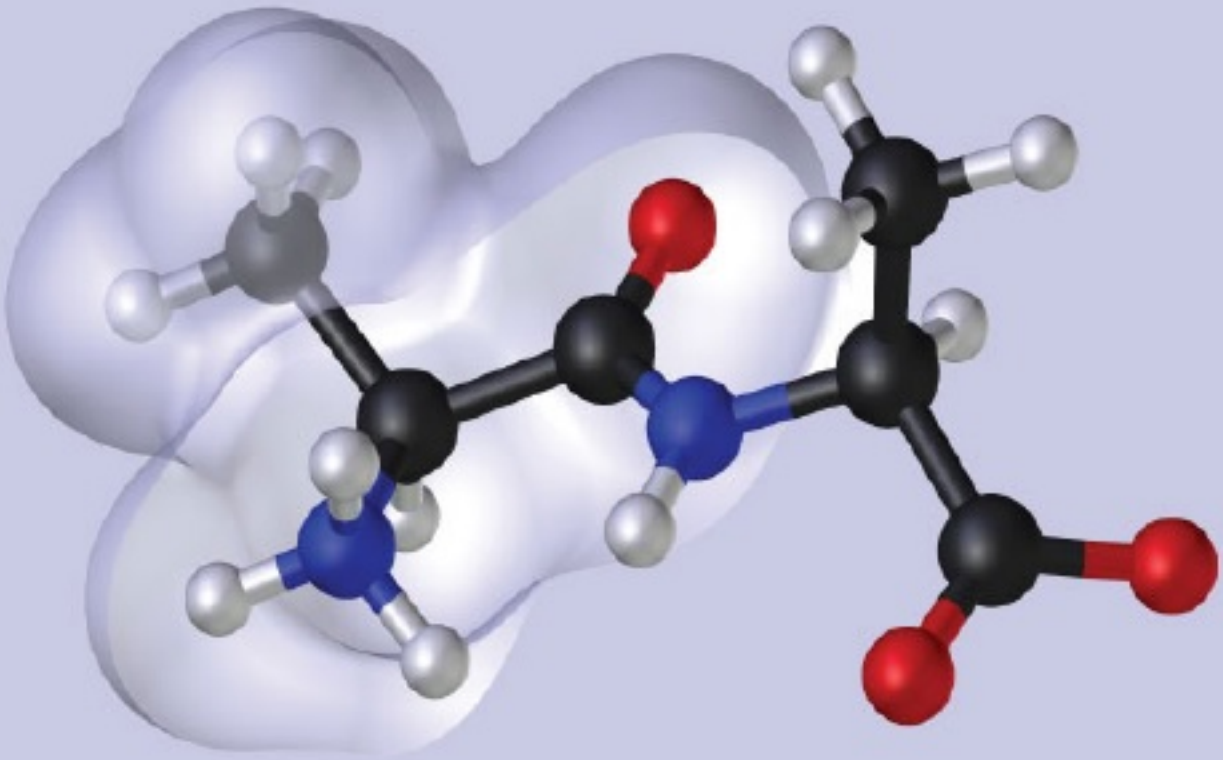
ISBN: 978-0-9971499-4-4

Austin Publishing Group

Software and Techniques for Bio-Molecular Modelling



Edited by

Azat Mukhametov

Austin Publishing Group

Software and Techniques for Bio-Molecular Modelling

Azat Mukhametov

Published by Austin Publications LLC

Published Date: December 01, 2016

Online Edition available at http://austinpublishinggroup.com/ebooks

For reprints, please contact us at ebooks@austinpublishinggroup.us

We consider to publish more books on the topics of drug design, molecular modelling, and structure-activity relationships. Those, wishing to contribute a chapter or article please contact this Book's Editor by e-mail:
azat@venture-pharmaceuticals.com

Austin Publishing Group

Table of Contents

Austin Publishing Group

Preface

The purpose of this book is to introduce new researchers into the field of Bio-Molecular Modelling. The field lays at intersection of such disciplines as Biology, Chemistry, Physics, and Computer Sciences. It is not surprising that researchers with respective backgrounds are attracted.

Historically, modelling techniques applied to discovery and development of new bio-active compounds, were ligand-based. These are methods of quantitative structure-activity (QSAR) and structure-property (QSPR) studies. To date, these methods include multiple techniques. Linear regression analysis, support vector machines, random forest, naive Bayes, and artificial neural networks are just some examples that utilize massive mathematical apparatus. With introduction of protein structure determination methods and exponential increase in computer power, structure-based methods of designing bio-active compounds became also available.

Modelling techniques got fast developing being applied to the challenges of life sciences (pharmaceutical, bio-technology, and agricultural sciences). These are search for biologically active molecules with targeted properties, development of special purpose proteins for industry, and fundamental studies of biological processes at molecular level.

Bio-Molecular Modelling includes various sub-fields. Bio-informatics lays at intersection of biology and computer sciences. Chemo-informatics utilizes computer sciences to operate with chemical structures. Theoretical bio-physics applies methods of computational physics to biology challenges. This way the fields of Bio-Molecular Modelling develop extensively and intensively in various directions.

Nova days Bio-Molecular Modelling includes multiple techniques. Methods of homology modelling let model three-dimensional structures of proteins based on the sequence of target protein and x-ray structural data of the template protein(s). Automated computational docking approaches make it possible to study the way small molecules may interact with bio-molecular targets. Being utilized to large library of ligands, it gets name of virtual screening. Pharmacophore modelling techniques let make pharmacophore models based on the structures of ligands or protein-ligand complexes for following pharmacophore-based virtual screening of small

molecular ligands. Techniques of Molecular Dynamics simulations let computationally model dynamics of bio-molecules in native environment. Atomic-level interactions between parts of bio-molecules or their fragments can be studied with Quantum Chemistry approaches. Bio-Informatics approaches let understand interactions between bio-molecular targets and select key ones for Bio-Molecular Modelling.

The listed are not all topics and fields of Bio-Molecular Modelling. However, these are the approaches which formed ground for future developments. Multiple methods and approaches are arising currently.

Structurally, this Book "Software and Techniques for Bio-Molecular Modelling" consists of three parts: "Software", "Techniques", and "Case Studies".

Modern and powerful software is a key component for Bio-Molecular Modelling. These are complex packages, small applications, as desktop as network based. Databases can also be considered with software. "Software" – includes small chapters on software packages, and algorithms, available as on open-source as on commercial basis. Each article gives information on software, algorithms involved, purposes, pros and cons, license terms, references.

Chapters on special-purpose applications are included. These are GROMPALA (a membrane-implicit modelling method to screen lipid-interacting molecules) by Dr. Laurence Lins; AtlasCBS (a graphic tool to map the content of structure-activity databases) by Dr. Cele Abad-Zapatero; and Modeller (an application for homology modelling) by Dr. Raghvendra Singh.

Many techniques for Bio-Molecular Modelling, are included into the "Techniques" part of the Book. Each chapter includes introduction, background, technique description, references. All of the techniques described have been successfully applied to bio-molecular challenges by scientists in the field. These are the technique for Homology Modelling: Protein structure prediction using molecular homology modelling, by Dr. Luiz Carlos Bertucci Barbosa. Technique for Virtual Screening: Structure, shape and electrostatic based virtual screening to discover small molecule therapeutics, by Dr. Raman Parkesh. Techniques for Molecular Dynamics Simulations: Introduction to the molecular dynamics of biomolecules, by Dr. D. A. Morton-Blake; GPU accelerated molecular dynamics simulations in predicting the protein-protein binding affinity from residues interactions within the binding surface, by Dr. Vannajan Sanghiran Lee. And the technique for Bio-informatics: Knowledge formalization and high-throughput data visualization using signaling network maps, by Dr. Andrei Zinovyev.

It is useful for as novice as more experienced researcher to look through a sample study before planning personal one. Multiple examples of successful studies in the field of Bio-Molecular Modelling are given in a "Case Studies" part of the Book. Experienced scientists have demonstrated the way modelling techniques can be applied to different research challenges. Hopefully this part of the Book will help a novice researcher to comprehend methods and techniques of Bio-Molecular

modelling. These are the Case Studies on various methods for neuro-ligands optimization using molecular modelling, by Dr. Anil Kumar Mishra; on virtual screening of DEN-2 non-competitive inhibitors, by Dr. Noorsaadah Abd Rahman; on comparative evaluation of docking programs - a case study with small peptidic ligands, by Dr. Punit Kaur; on protein Interaction study of novel mutants of human Hsp70 and Ad5 motif (PNLVP), by Dr. Salehhuddin Hamdan; on structure based drug design in identification of novel androgen receptor antagonist, by Dr. Muthiah Ramanathan; on Hepatitis C viral polymerase inhibition using directly acting antivirals - a computational approach, by Dr. Abdo A Elfiky; on application of structure and ligand-based drug design for finding lead compounds from natural product source: case of influenza targeted, by Dr. Muchtaridi; on molecular dynamics of E. coli undecaprenyl diphosphate synthase - asymmetry in a homodimer, by Dr. Irene E Newhouse; and on molecular dynamics simulations and molecular docking approaches in endoinulinase chemical modification, by Dr. Homa Torabizadeh.

I am very grateful to all of the Authors and their colleagues, who accepted invitation to contribute to this Book. Also I am very thankful to the external reviewers who were very helpful to assist in improving the chapters. I am extremely grateful to the company Venture Pharmaceuticals Ltd (Belize) which was a Technology Sponsor for all network communications with all of the perspective and the current Authors.

Dr. Azat Mukhametov,

Editor of the Book

Austin Publishing Group

GROMPALA: A Membrane-Implicit Modelling Method to Screen Lipid-Interacting Molecules

Steinhauer S, Crowet JM, Brasseur R and Lins L*

Laboratoire de Biophysique Moléculaire aux Interfaces, GBX-ABT, University of Liège, Belgium

***Corresponding author:** Lins L, Laboratoire de Biophysique Moléculaire aux Interfaces, GBX-ABT, University of Liège, Passage des Déportés, 2-5030 Gembloux, Belgium, Email: l.lins@ulg.ac.be

Published Date: December 01, 2016

ABSTRACT

In this chapter, we describe an improved version of our previously published Monte Carlo method IMPALA, based on an implicit description of the membrane. The implementation of the implicit water-membrane forcefield IMPALA into GROMACS molecular dynamics software suite is called GROMPALA. The method aims to decrease computational costs compared to explicit environment representation in MD simulation. We attend to gain a more accurate description as compared to IMPALA by taking advantage of the all-atom molecular dynamics algorithms. GROMPALA is designed to get insight into molecule-membrane interactions taking place on reasonable time scales, notably to screen large sets of peptides, than can serve as primary selection tool for further atomistic molecular dynamics simulations.

Keywords: Molecular Modelling; Amphipathic Peptides; Implicit Membrane; Hydrophobicity

Abbreviations: MD: Molecular Dynamics, MC: Monte Carlo, ASA: Accessible Surface Area, VdW: Van der Waals, MCe: Mass Centre; GB: generalized Born; MAG: Magainin; DDK: Dermadistinctin K; MLT: Mellitin; TM: Transmembrane

INTRODUCTION

Biological life sciences such as pharmacology, toxicology, bioindustry or cosmetology depend on knowledge about how membrane-related metabolism, transport and disruption processes take place. Proteins interacting with the membrane are essential in those phenomena, are present in all cells and represent more than one third of the genome. Understanding protein-membrane interactions is hence of fundamental importance.

Over the last decade, molecular dynamics (MD) has gained attractiveness in that domain. MD is a valuable tool to study interactions between proteins or peptides and membrane because it gives access to the atomistic details of the interaction as well as energetics and dynamics of the observed processes [1]. MD is based on the use of the motion equations of Newton and on a forcefield to simulate how an ensemble of atoms moves relative to each other. Forcefields include potential equations and parameters to reproduce stretching, bending and rotations of covalent bonds, to maintain planarity and chirality of several groups as well as to simulate Van der Waals and electrostatic interactions. The parameters which depend on defining atom types have been calibrated to reproduce a wide range of experimental values. MD studies have proven to be able to reproduce biophysical and biological processes, for solutes in uniform solvent, as well as for membrane environments. However, due to the high computational cost of molecular mechanics simulations using explicit membrane, there is a growing interest for implicit representation of the lipid bilayer.

A wide range of models for the interaction with implicit membranes have been developed and are the object of several reviews [2,3]. These models can be classified as knowledge-based [4–10] or physics-based [11–14]. The former are usually based on experimental data for the free energy of transfer of small molecules, typically amino acids, from water to apolar media, e.g. octanol [5,6,9]. Atomic solvation parameters are derived from these data and they are then often used with atomic solvent accessible surface area and tuned according to the atomic position in a membrane slab to compute the solvation energy of bigger molecules, such as peptides or proteins. The methods give optimal positions for the molecule inside the membrane which will be compared with experimentally known orientation for parametrization. Knowledge-based methods can also be combined with forcefields to include Van der Waals, electrostatic and torsion energy in the energy function and sample the protein conformational space [4,9]. The method presented in this paper is part of these methods, as described further below.

For the physics-based methods, the membrane protein interactions are decomposed into electrostatic and nonpolar contributions. For the electrostatics, a membrane can be approximated as a region with low dielectric constant, in contrast to water which has a high dielectric constant. This can be described with the Poisson-Boltzmann (PB) theory [15]. While several groups have used this model [10,12], it is computationally expensive and difficult to use for MD simulations [16]. Several faster methods have hence been developed [17,18] that mainly use the Generalized

Born (GB) approach, which has been first introduced by Still et al [19]. The GB equation is derived from the Born model [20], a solution of the PB equation for a charged spherical solute in a solvent with different dielectric constant. The GB approximation expresses the electrostatic solvation energy for a set of charged spheres, representing the biomolecule, and accounts for the effect of the dielectric medium on the pairwise interactions of charged particles [19]. The key point is the calculation of the sphere radius, named Born radius, because it depends on the position and volume of all the other atoms in the solute [21]. The Coulomb field approximation is used to compute Born radii [3]. However, as the membrane is modelled as a layer with different dielectric constants, the GB approach needs to be adapted. In the first membrane model based on GB, proposed by Spassov et al [22], the dielectric constant is the same within the membrane and for the protein. A procedure to handle multiple dielectric environments with GB was then proposed by Feig et al.[16]. For the nonpolar contribution, it mainly corresponds to the cost of cavity formation and is usually approximated by a term proportional to the solvent accessible area [2].

In our lab, we have developed a Monte Carlo method using an implicit description of the membrane, called IMPALA [6]. The forcefield was parameterized for mimicking a membrane in aqueous environment by considering 1) the hydrophobic effect between the membrane and a solute and 2) the perturbation effect of the solute on the lipid acyl chain organization. Both energy restraint terms depend on the solvent accessible surface area and a membrane potential which mimics the hydrophilic profile of the lipid bilayer. While being very simplistic, this method notably allowed to accurately study and classify different lipid-interacting peptides [23] and to predict entire membrane protein orientation into lipid bilayers [24,25]. The main limitation of the method resides in the fact that the conformation of the peptide is not modified following its insertion into the implicit membrane.

In this chapter, we describe the implementation of the implicit water-membrane forcefield IMPALA into GROMACS molecular dynamics software suite. We call the resulting hybrid method GROMPALA. The method aims to decrease computational costs compared to explicit environment representation in MD simulation. We attend to gain a more accurate description as compared to IMPALA by taking advantage of the all-atom molecular dynamics algorithms. GROMPALA is designed to get insight into molecule-membrane interactions taking place on reasonable time scales, notably to screen large sets of peptides, than can serve as primary selection tool for further atomistic molecular dynamics simulations.

DESCRIPTION OF THE GROMPALA METHODOLOGY

We have implemented the algorithm from Ducarme et al [6] into Gromacs software [26,27]. In the Impala membrane model, two pseudo energy terms are considered: a) the hydrophobic energy restraint (eq.2) and b) the lipid perturbation energy restraint (eq.3). Both terms depend on atomic transfer energy and accessible surface area (ASA). A Monte Carlo (MC) procedure is used

to explore optimal insertion configurations. The conformation of the lipid-interacting molecule is considered to be rigid and the membrane hydrophobicity is modelled by an empirically parameterized symmetric sigmoidal curve, C(z) (eq.1):

$$C_{(z)} = 1 - \frac{1}{1 + e^{\acute{a}(z - z_0)}} \qquad \text{(equation 1)}$$

where α is a constant equal to 1.99; z_0 corresponds to the middle of polar heads and z is the position in the membrane.

The hydrophobicity of the membrane for the interaction is simulated by E_{pho}:

$$E_{pho} = -\sum_{i=l}^{N} S_{(i)} E_{tr(i)} C_{(zi)} \qquad \text{(equation 2)}$$

Where N is the total number of atoms, $S_{(i)}$ the accessible surface to solvent of the *i* atom, $E_{tr(i)}$ its transfer energy per unit of accessible surface area and $C_{(zi)}$ the $_{zi}$ position of atom *i*.

The perturbation of the bilayer by insertion of the molecule was simulated by the lipid perturbation restraint (E_{lip}):

$$E_{lip} = Klip * a_{lip} \sum_{i=1}^{N} S_{(i)} \left(1 - C_{(zi)}\right) \qquad \text{(equation 3)}$$

where a_{lip} is an empirical factor fixed at 7.53624 Kj.mol^{-1}Å^{-2} and Klip, a weight factor comprised between 0.1 and 1.

In our Impala implementation to Gromacs (Grompala), the MD routine replaces the original MC approach. We have adapted the values of accessible surface area and VdW radii for each atom type of Grompala. A performance optimized program routine for the calculation of solvent accessible surface areas has been implemented to the Gromacs core program by modifying the Gromacs tool g_sas, based on the DCLM method [27]. After calibration, ASA probe radius was set to 0.115. Both parameters (ASA and VdW radius) were associated to the corresponding atom type definition of the OPLS-AA forcefield [28]. The transfer energy values were taken from [29].

Preliminary energy minimization has been carried out in absence of Impala forcefield. The MD part of the simulations was done without periodic boundary conditions at 323 K for 10 ns by steps of 2 fs. A dielectric constant of 1 was used and Coulomb and Van der Waals have been computed without cutoff, since the systems studied are small. Each calculation is repeated 7 times.

To test and calibrate Grompala, we investigated four amphipathic peptides and one transmembrane domain that have been described in the literature for their interaction with lipids, notably by NMR. Concerning the amphipathic helical peptides, Magainin2 (MAG) (PDB ID: 2MAG) and Dermadistinctin K (DDK) (PDB ID: 2JX6) are antimicrobial peptides that have been shown to form amphipathic α-helices oriented parallel to the membrane surface [30,31]; Magainin has been previously used as test peptide for IMPALA [6]. The Influenza hemagglutinin HA2 fusion peptide (PDB ID: 2XKA) has been reported to be a helical hairpin at the lipid/water interface [32]. Mellitin (PDB ID: 2MLT) is a highly hemolytic helical peptide from Bee venom. It has been shown to be able to adopt a wide range of orientations in the membrane, from parallel to the membrane surface to a transmembrane configuration [33–35]. This peculiar behavior has also been observed with IMPALA [6] and is hence a good check for Grompala. To compare those amphipathic helices to transmembrane domains, the M3 segment of the alpha subunit of the nicotic acetylcholine receptor from Torpedo californica has been chosen (PDB ID: 3MRA) [36].

CALIBRATION OF K_{LIP}

Previously, we observed that the lipid perturbation restraint had to be weighted for some molecules, since in some cases, molecules interacting experimentally with the membrane were unable to insert into the IMPALA bilayer. The weighting factor Klip allows restoring lipid insertion with values varying between 0.1 and 0.8 (unpublished data). For Grompala, we tested Klip values between 0.1 and 0.8 for the different peptides. The best results were obtained with values between 0.45 and 0.65 for all the molecules (data not shown) and we show the results obtained for Klip=0.5 for all the peptides.

PEPTIDE BEHAVIOR IN THE GROMPALA MEMBRANE

Different parameters have been analyzed for the five peptides: the conservation of the helical conformation along the simulations, the position of the mass centre (MCe) and the orientation (tilt) of the peptide into the implicit membrane. For the helical conformation calculation, the DSSP method is used [37]; the orientation of the peptide is defined by the axis between the MCe of the 3 first and 3 last Cα of the helical part of the peptide only. For DDK peptide, the tilt was calculated using the 7 first and 7 last atoms, whatever the conformation is (the peptide is mainly destructured-see below).

For the antimicrobial peptides DDK and MAG that have been found experimentally to orient on the surface of the membrane, they are both oriented mostly parallel to the lipid plane with Grompala (Figure 1A and Figure 2A respectively), with insertion angles around 90° and around 60°-70° towards the membrane normal, for DDK (Figure 1D) and MAG respectively (Figure 2D). Both peptides have their mass centre located in the hydrophilic interface, at 11-12 Å from the bilayer centre (Figure 1C and Figure 2C). This is in very good agreement with the experimental data [30] and with IMPALA calculations (not shown).

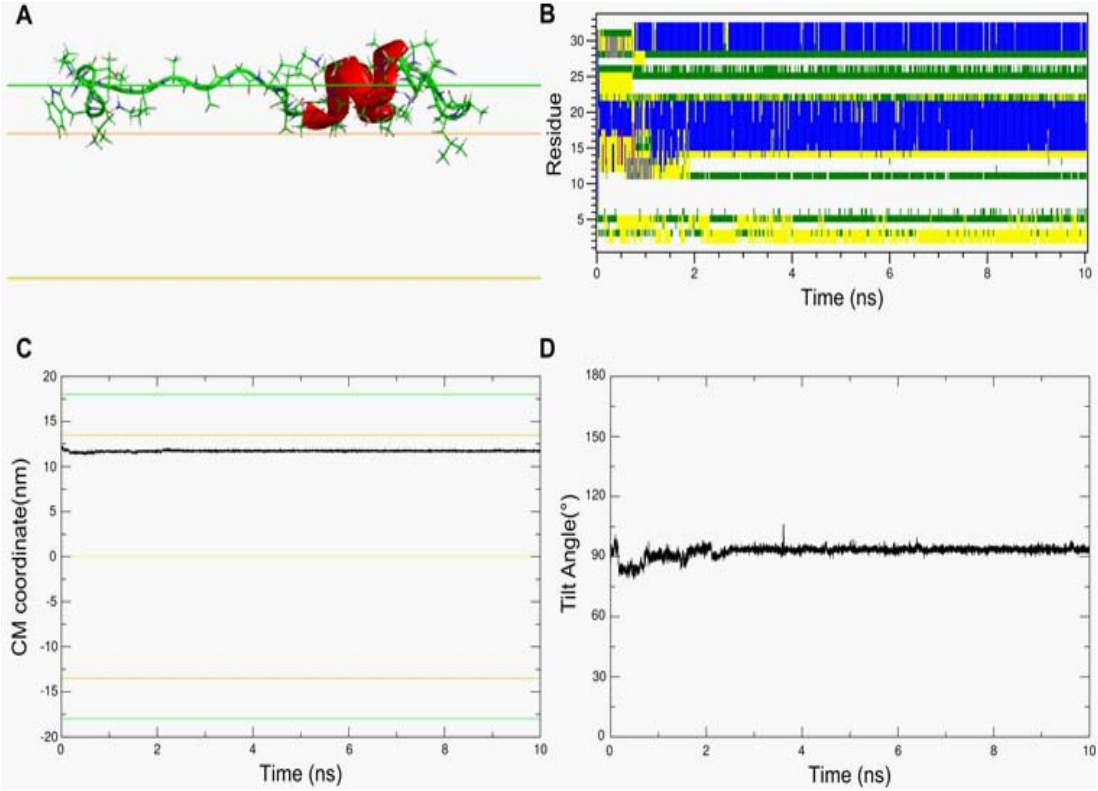


Figure 1: Grompala simulation of DDK peptide.

A. Final position of the peptide into the implicit membrane. The yellow plane represents the centre of the bilayer, the orange plane, the interface between the lipid polar headgroups and the hydrophobic tails and the green plane, the interface between water and lipid polar heads.

B. Evolution of the secondary structure of each residue of the peptide along the simulation (10 ns). Blue: α helix; green: bend, yellow: β structure; white: coil; mauve: 5-helix; grey: 3-helix.

C. Evolution of the mass centre position of the peptide along the simulation. The planes are the same as in A. The bilayer is symmetric and the thickness is ±18 Å.

D. Evolution of the angle of insertion (°) of the peptide (as defined in the text) along the simulation.

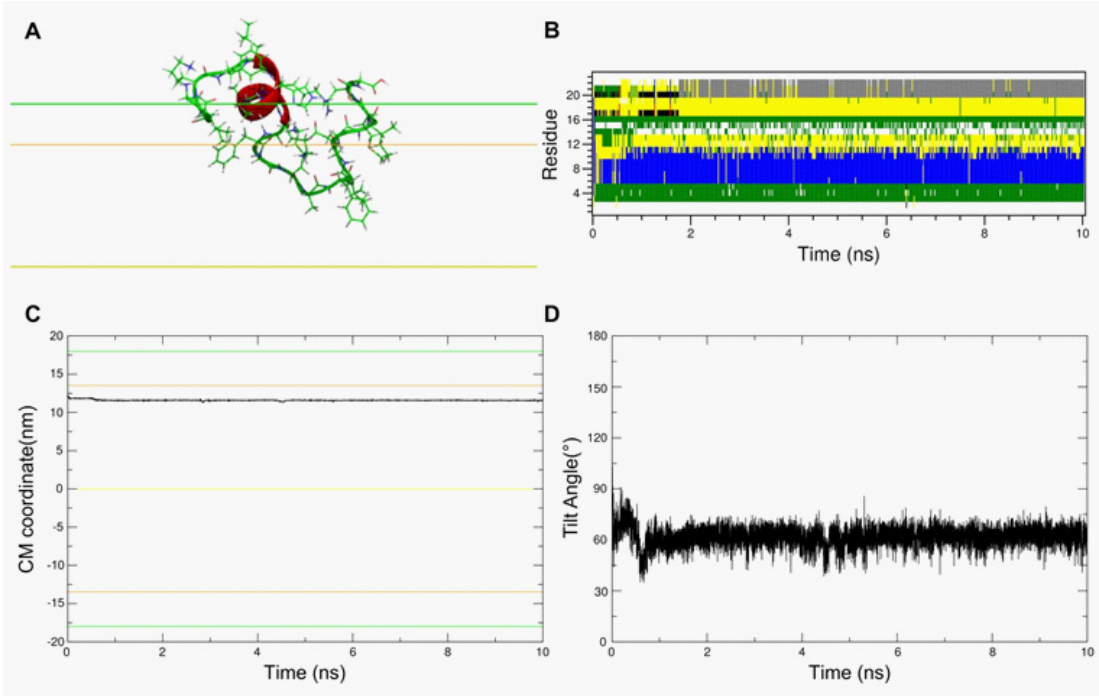


Figure 2: Grompala simulation of MAG peptide-same representation as for Figure 1.

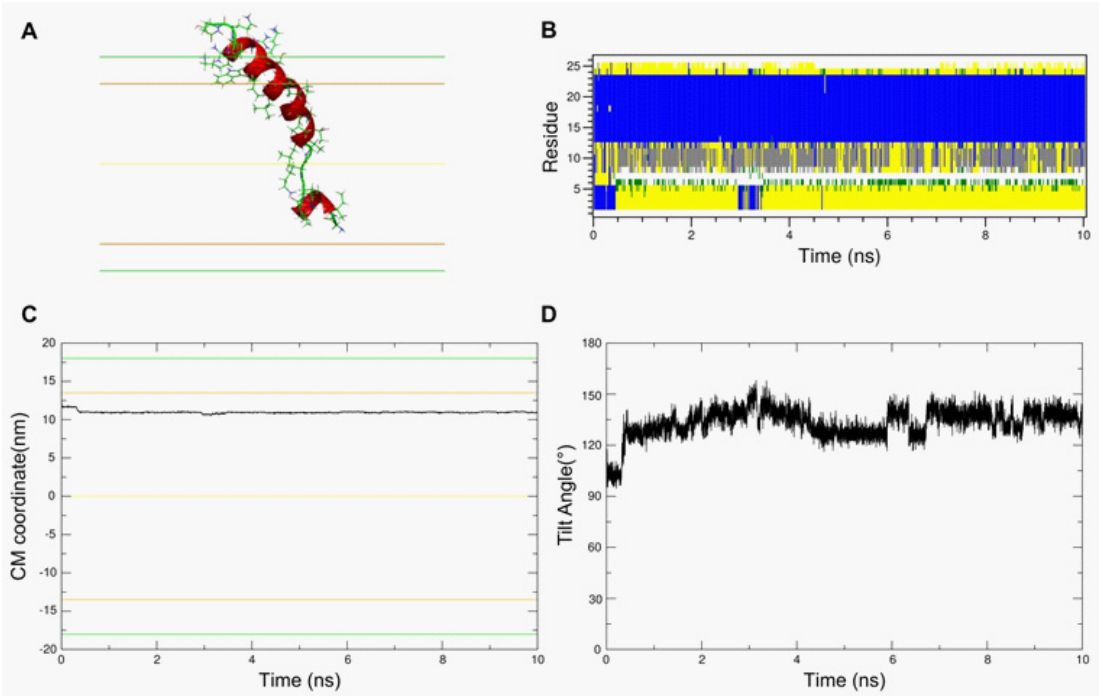


Figure 3: Grompala simulation of MLT peptide-same representation as for Figure 1.

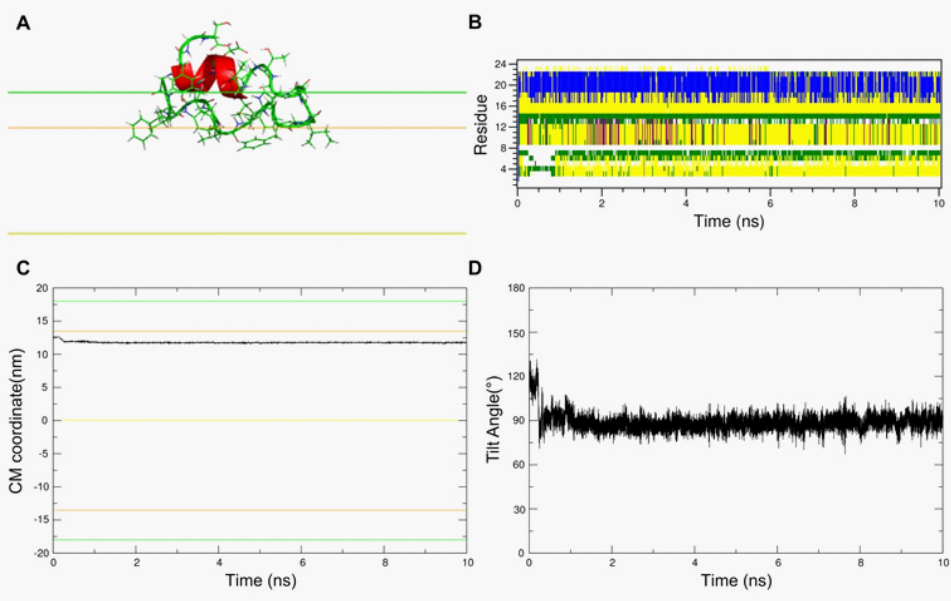


Figure 4: Grompala simulation of HA2 fusion peptide-same representation as for Figure 1.

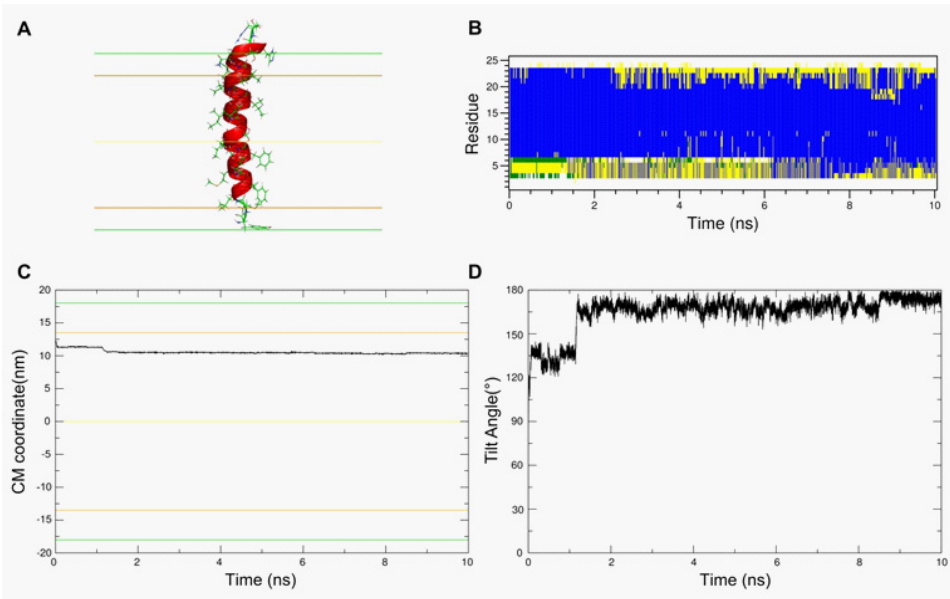


Figure 5: Grompala simulation of M3 TM segment-same representation as for Figure 1.

The peptide conformation was also followed during the simulation. Figures 1A and C shows that DDK is partly helical (residues 15-21 and 29-32), but the N-terminal part is destructured. For MAG (Figure 2A and Figure 2B), a helical structure is observed at residues 5-11 in few simulations; in the other calculations, the MAG peptide is even more destructured (data not shown).

For mellitin, the results are presented on Figure 3. We can clearly see that the helical structure is pretty well conserved, with the N-terminal part being in turn (Figure 3A,B). As previously observed by Impala [6] and also experimentally [33–35], the insertion angle can adopt a larger array of values, from 100 to 150° (Figure 3D), the mass centre being located at the interface between the hydrophobic tails and the hydrophilic lipid headgroups (around 10-11 Å) (Figure 3C).

The HA2 peptide that shows a helical hairpin structure experimentally [32], is oriented parallel to the membrane interface (Figure 4A) with its mass centre around 12 Å (Figure 4C), the tilt staying at 90° during the whole simulation (Figure 4D). The helical structure is preserved for the second helix (residues 16 to 22) (Figure 4B), and the overall hairpin configuration is observed (Figure 4A).

For 3MRA (Figure 5), the results are in very good agreement with the experimental behavior, i.e. a transmembrane insertion (Figure 5A) that appears in the few first nanoseconds of the simulation (MCe position around 10 Å- Figure 5C and an angle around 180°- Figure 5D). The helical conformation is preserved along the whole peptide for the whole simulation (Figure 5B).

CONCLUSIONS AND PERSPECTIVES

The results of Grompala for the four amphipathic peptides are encouraging, since an interfacial position is predicted for all of them. For mellitin, our results are in agreement with the fact that this peculiar peptide can adopt a wider variety of positions into the membrane, suggesting that our method is adapted to distinguish between strictly interfacial peptides and those having more specific features. It also clearly distinguishes between amphipathic and transmembrane peptides, as shown for 3MRA.

Concerning the structure of the peptide along the simulations, the helical conformation is not as well conserved as it would by with atomistic molecular dynamics approach, especially in the case of DDK and MAG. The former is defined by NMR as a 33-residues helical structure in the presence of DPC micelles, with the 5 N-terminal amino acids being coiled, and with a distorsion around residues 10-16 [30]. In water, the peptide is destructured, as shown by CD measurements [30]. In the Grompala simulations, the 1-14 N-terminal domain is not structured as an helix, and the 22-28 residues are configured as a kink. In the same way, Magainin is relatively destructured as compared to NMR results [31]. This should be due to the fact that the peptide lays in the vacuum when not inserted into the implicit membrane. For some peptides, this could induce a destabilization of the hydrogen bonding maintaining the helical conformation, since both peptides are indeed destructured in water (i.e. not in a hydrophobic medium). When inserted at the interface, the duration of the simulation might not be enough to allow reappearance of helical structure. Future investigation in that direction should help to improve the structural stability of the peptides in the Grompala methodology.

In conclusion, by using the gold standard molecular dynamics approach combined to the implicit membrane representation of our home-designed IMPALA methodology, we built up an original and sufficiently fast method that should help to predict and screen peptide membrane behavior. This could be the starting point to subsequent MD simulations that are more time-consuming.

ACKNOWLEDGMENTS

We thank the FNRS (PDR grant T.1003.14), the Belgian Program on Interuniversity Attraction Poles initiated by the Federal Office for Scientific, Technical and Cultural Affairs (IAP P7/44 iPros), The University of Liège (Fonds Spéciaux de la Recherche, Action de Recherche Concertée-Project FIELD) for financial support.

LL and RB (currently retired) are Senior Research Associate and Research Director for the Fonds National de la Recherche Scientifique (FRS-FNRS), JMC is supported by the IAP iPros project and SS was supported by the Walloon Region (RaidGBS project).

The software GROMPALA is available on request.

References

1. Deleu M, Crowet J-M, Nasir MN, Lins L. Complementary biophysical tools to investigate lipid specificity in the interaction between bioactive molecules and the plasma membrane: A review. Biochim Biophys Acta. 2014; 1838: 3171–3190.
2. Feig M1. Implicit membrane models for membrane protein simulation. Methods Mol Biol. 2008; 443: 181-196.
3. Grossfield A. Computational Modeling of Membrane Bilayers. Philadelphia: Elsevier. 2008.
4. Lazaridis T. Effective energy function for proteins in lipid membranes. Proteins. 2003; 52: 176-192.
5. Lomize AL, Pogozheva ID, Lomize MA, Mosberg HI. Positioning of proteins in membranes: a computational approach. Protein Sci. 2006; 15: 1318-1333.
6. Ducarme P, Rahman M, Brasseur R. IMPALA: a simple restraint field to simulate the biological membrane in molecular structure studies. Proteins. 1998; 30: 357–371.
7. Jähnig F, Edholm O. Modeling of the structure of bacteriorhodopsin. A molecular dynamics study. J Mol Biol. 1992; 226: 837-850.
8. Sanders CR 2nd, Schwonek JP. An approximate model and empirical energy function for solute interactions with a water-phosphatidylcholine interface. Biophys J. 1993; 65: 1207-1218.
9. Efremov RG, Nolde DE, Vergoten G, Arseniev AS. A solvent model for simulations of peptides in bilayers. I. Membrane-promoting alpha-helix formation. Biophys J. 1999; 76: 2448-2459.
10. Ben-Tal N, Ben-Shaul A, Nicholls A, Honig B. Free-energy determinants of alpha-helix insertion into lipid bilayers. Biophys J. 1996; 70: 1803-1812.
11. Kessel A, Cafiso DS, Ben-Tal N. Continuum solvent model calculations of alamethicin-membrane interactions: thermodynamic aspects. Biophys J. 2000; 78: 571-583.
12. Murray D, Ben-Tal N, Honig B, McLaughlin S. Electrostatic interaction of myristoylated proteins with membranes: simple physics, complicated biology. Structure. 1997; 5: 985-989.
13. Roux B, MacKinnon R. The cavity and pore helices in the KcsA K+ channel: electrostatic stabilization of monovalent cations. Science. 1999; 285: 100-102.
14. Im W, Roux B. Ion permeation and selectivity of OmpF porin: a theoretical study based on molecular dynamics, Brownian dynamics, and continuum electrodiffusion theory. J Mol Biol. 2002; 322: 851-869.
15. Sharp KA, Honig B. Electrostatic interactions in macromolecules: theory and applications. Annu Rev Biophys Biophys Chem. 1990; 19: 301-332.

16. Feig M, Onufriev A, Lee MS, Im W, Case DA. Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. J Comput Chem. 2004; 25: 265-284.

17. Im W, Feig M, Brooks CL 3rd. An implicit membrane generalized born theory for the study of structure, stability, and interactions of membrane proteins. Biophys J. 2003; 85: 2900-2918.

18. Feig M, Im W, Brooks CL 3rd. Implicit solvation based on generalized Born theory in different dielectric environments. J Chem Phys. 2004; 120: 903-911.

19. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc. 1990; 112: 6127-6129.

20. Born M. Volumen und hydratationswarme der ionen. Z Phys. 1920; 1: 45.

21. Qiu D, Shenkin PS, Hollinger FP, Still WC. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. J Phys Chem A. 1997; 101: 3005–3014.

22. Spassov VZ, Yan L, Szalma S. Introducing an Implicit Membrane in Generalized Born/Solvent Accessibility Continuum Solvent Models. J Phys Chem B. 2002; 106: 8726-8738.

23. Lins L, Charloteaux B, Thomas A, Brasseur R. Computational study of lipid-destabilizing protein fragments: towards a comprehensive view of tilted peptides. Proteins. 2001; 44: 435-447.

24. Basyn F, Charloteaux B, Thomas A, Brasseur R. Prediction of membrane protein orientation in lipid bilayers: a theoretical approach. J Mol Graph Model. 2001; 20: 235-244.

25. Basyn F, Spies B, Bouffioux O, Thomas A, Brasseur R. Insertion of X-ray structures of proteins in membranes. J Mol Graph Model. 2003; 22: 11-21.

26. Lindahl E, Hess B, Spoel D Van Der. GROMACS 3.0: a package for molecular simulation and trajectory analysis. Mol Model Annu. 2001; 7: 306–317.

27. Hess B, Kutzner C. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J Chem Theory Comput. 2008; 4: 435-447.

28. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. J Phys Chem B. 2001; 105: 6474–6487.

29. Brasseur R. Differentiation of lipid-associating helices by use of three-dimensional molecular hydrophobicity potential calculations. J Biol Chem. 1991; 266: 16120-16127.

30. Verly RM, de Moraes CM, Resende JM, Aisenbrey C, Bemquerer MP. Structure and membrane interactions of the antibiotic peptide dermadistinctin K by multidimensional solution and oriented 15N and 31P solid-state NMR spectroscopy. Biophys J. 2009; 96: 2194-2203.

31. Bechinger B, Zasloff M, Opella SJ. Structure and interactions of magainin antibiotic peptides in lipid bilayers: a solid-state nuclear magnetic resonance investigation. Biophys J. 1992; 62: 12-14.

32. Lorieau JL, Louis JM, Bax A. The complete influenza hemagglutinin fusion domain adopts a tight helical hairpin arrangement at the lipid:water interface. Proc Natl Acad Sci U S A. 2010; 107: 11341-11346.

33. Irudayam SJ, Pobandt T, Berkowitz ML. Free energy barrier for melittin reorientation from a membrane-bound state to a transmembrane state. J Phys Chem B. 2013; 117: 13457-13463.

34. Niu W, Wu Y, Sui SF. Orientation of membrane-bound melittin studied by a combination of HPLC and liquid secondary ion mass spectrometry (LSIMS). IUBMB Life. 2000; 50: 215-219.

35. Raghuraman H, Chattopadhyay A. Orientation and dynamics of melittin in membranes of varying composition utilizing NBD fluorescence. Biophys J. 2007; 92: 1271-1283.

36. Lugovskoy AA, Maslennikov IV, Utkin YN, Tsetlin VI, Cohen JB. Spatial structure of the M3 transmembrane segment of the nicotinic acetylcholine receptor alpha subunit. Eur J Biochem. 1998; 255: 455-461.

37. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22: 2577-2637.

Austin Publishing Group

AtlasCBS: A Graphic Tool to Map the Content of Structure-Activity Databases

Abad-Zapatero C*

Department of Medicinal Chemistry and Pharmacognosy, Center for Pharmaceutical Biotechnology, University of Illinois at Chicago, USA

***Corresponding author:** Abad-Zapatero C, Department of Medicinal Chemistry and Pharmacognosy, Center for Pharmaceutical Biotechnology, University of Illinois at Chicago, 900 South Ashland Av. Chicago, IL. 60607, USA, Tel: 312-355-4105; Fax: 312-413-9303; Email: caz@uic.edu

Published Date: December 01, 2016

ABSTRACT

The number of databases containing biologically active compounds and the bioactivities towards the corresponding targets (SAR-Databases) has grown substantially in the last decade. Concurrently, the number of bioactivities and targets has also grown dramatically following the extensive and growing data available in the medicinal chemistry and biological literature. The web resources and tools have expanded the range of possibilities on how best to present this myriad of data entries to the drug discovery community. Naturally, the connection between the chemical space of ligands and the biological space of targets is the experimentally determined affinities estimated by a variety of assays (K_i, IC_{50}, K_d, MIC and others).

In the past few years, the medicinal chemistry community has begun exploring the use of alternative variables (also referred to as 'combined variables') to better understand the interrelation between the chemical and biological spaces. The concepts of 'Ligand Efficiency' and 'Ligand Efficiency Indices' are now commonly accepted as variables and defined as a combination of affinities (typically in the numerator) with certain physico-chemical properties of the ligands (normally in the denominator). Efficiency indices based on a subtraction definition (e.g. affinity minus a term related to certain properties of the ligand) are also widely used.

We present the use of variables related to the concept of Ligand Efficiency Indices to map the content of SAR-Databases in a series of Cartesian planes that combine three critical variables in drug discovery (affinity, size and polarity) into easily interpretable 'efficiency planes' to guide the drug discovery process. The series of planes, using different physico-chemical variables at different scales resembles an atlas of chemico-biological-space (AtlasCBS).

The existing web tool available at the European Bioinformatics Institute is briefly discussed and illustrated. The limitations of the current application are also examined. A concrete proposal is made to convert the AtlasCBS concept into an effective personalized application using the virtual machine environment MyCHEMBL combined with open access Chemoinformatics tools such as KNIME.

Keywords: SAR-Databases; Ligand Efficiency Indices; Chemoinformatics; Chemico-Biological Space; Drug Discovery Software; AtlasCBS

Abbreviations: SAR-Databases: Structure Activity Relationship Databases; AtlasCBS: Atlas of Chemico-Biological Space; LE: Ligand Efficiency; LEI(s): Ligand Efficiency Indice(s); BEI: Binding Efficiency Index; SEI: Surface Efficiency Index. LEM: Ligand Efficiency Metrics; PDB: Protein Data Bank; BindingDB: Binding Data Base; PDBBind: Protein Data Bank Binding Data Base, a subset of protein-ligand complexes extracted from PDB; ChEMBL: Chemistry (small molecule database) of the European Molecular Biology Organization. KNIME: Konstanz Information Miner, open source data analytics, reporting and integration platform. WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. SQL: Structured Query Language, an international standard for database manipulation

INTRODUCTION

The purpose of this brief note is to introduce the concept of a web tool named AtlasCBS developed to facilitate the extraction, mapping, representation and analysis of the content of SAR-databases, which contain extensive collections of chemical compounds (ligands), and the associated data relating those compounds to the biological entities (targets) against which the chemical matter has some activity. These affinity or activity data are predominantly (if not exclusively) experimental in nature, as measured by some experimental assays. Because these databases relate the structure of the compounds to their biological activity, they will be referred to as Structure-Activity Relationship Databases (SAR-DB, or SAR-Databases).

This contribution will present the numerical and algebraic relationships among the different variables used in the current implementation, based on Ligand Efficiency Indices (LEIs) and Ligand Efficiency Metrics (LEM). The usage of the current web server will be discussed as well as its limitations. Within that framework a simple, effective and practical way to use these ideas and concepts individually will be presented and illustrated, using the content of the public databases in combination with open source modules to create workflows that can be independently tailored for specific needs within the drug-discovery community.

SAR-DATABASES

The availability of extensive Structure-Activity-Relationship data for a wide variety of biological targets in the literature and in experimental high-throughput centers and installations in the drug discovery community, made it imperative the development of databases that could store, annotate and easily retrieve the information. Within a period of a few years several vigorous databases originated that are now well established in biomedical community. Among them, BindingDB [1-3], PDBBind [4,5], ChEMBL [6], PubChem [7] and WOMBAT [8] are quite reputable, and with the unique 3-D flavor of the Protein Data Bank (PDB) [9,10] form the core of critical resources for the future of drug discovery. The database DrugBank [11,12] has a unique position in the community but cannot be considered to be exclusively or predominantly a SAR-Database. DrugBank focuses more on the clinical/pharmacological aspects of the compounds and therapies and does not contain any significant affinity data.

Naturally, all these databases are now well entrenched and do have a critical presence as Internet sources from where they derive their tremendous power and versatility. Concurrently, the absolute growth as revealed by the total number of entries, and the development rate of these resources is staggering. For example, ChEMBL20: 10,774 targets, 1,715,667 compound records, 1,463,270 distinct compounds and 13,520,737 activities extracted from nearly sixty thousand publications (https://www.ebi.ac.uk/chembl/). Some of the details have been presented in more extensive publications as well as some of the issues now facing due to the expanded rate of data deposition [13]. Among them are: redundancy, complementarity vs. uniqueness, private vs. public databases and access, quality of the data, inclusion of errors as well as others. From the viewpoint of this brief communication a critical issue is to find the 'best' representation of the available data.

In a simplistic way, databases are basically collections of interconnected numbers in groups of internal tables. In a way, it is as if we had a list of latitudes and longitudes of an immensely vast chemico-biological space (CBS). Using a historical perspective, this collection could be considered to be similar to the data collected by Thyco Brahe (circa 1600) relating the planets to the orbits that they circumnavigate in relation to the point of observation, Earth. By themselves, the list of numbers does not provide any clue as to nature of the orbits. In the same way that a list of x,y coordinates does not reveal directly the shape of a circumference or geometrical figure. It is only when we 'map' these numbers in a certain frame of reference in Cartesian (or polar) coordinates that we begin to understand the shapes, trajectories, paths and orbits of the planets.

This is what the AtlasCBS concept and application does in the context of chemico-biological-space. It highlights the relationship between targets and ligands in a series of efficiency planes or charts. The collection of maps and charts (in different scales and using different variables) is what suggests the idea of an atlas-like representation or an atlas of chemico-biological-space (AtlasCBS).

ALTERNATIVE VARIABLES IN MAPPING THE CONTENT OF SAR-DATABASES

In spite of the enormous amount of data contained in the existing SAR-Databases the key element is a measure of the affinity of the ligand (chemical entity) towards the biological target (typically a protein). The variety of assays from which those data are obtained are also included (i.e. in vitro, in cell, etc.) and in most cases a cross-referenced to the original publications. Most SAR-Databases (for example ChEMBL) now contain (or are easily calculated via the SMILES representation) a large set of physico-chemical properties: Molecular Weight (MW), Polar Surface Area (PSA), logP, logD among others.

Most of them are constructed based on relational databases (SQL, MySQL or related) with a large number of interconnected tables via primary keys. Yet, the key variable to analyze the data is essentially one of several possible affinity variables (K_i, IC_{50}, K_d and related).

Undoubtedly, this is the key variable that links the chemical and biological spaces. A ligand, by the fact that has some activity towards a concrete biological molecule (typically a macromolecule: protein, nucleic acid, lipid(s), etc.) connects the two domains: chemical and biological.

The underlying theme of this article is that using certain 'alternative variables', and particularly variables that combine the affinity with other physico-chemically relevant properties of the ligands, it is much easier to represent, interpret and analyze the content of those SAR-Databases. In turn, this ease of interpretation will simplify the mapping of CBS and the navigation in the tumultuous waters of drug-discovery facilitating productive, efficient and successful expeditions in drug-discovery.

This concept has been highlighted and discussed recently using certain proprietary software tools currently available to the community [14].

ATLASCBS: THE ESSENTIAL CONCEPT

The initial concepts and publications related to Ligand Efficiency (LE) were prompted by the notion of quantifying the quality of fragments, as seen and evaluated in fragment-based drug discovery (FBDD). This was the milieu within which Hopkins and colleagues introduced in 2004 the initial definition of LE (see Table 1) as the quotient between the free energy of binding (ΔG) and the number of non-hydrogen atoms (NHA) of the ligand. It was introduced basically as a size-related efficiency.

Table 1: List of Ligand Efficiency (LE) and Ligand Efficiency Indices (LEIs) definitions relevant to the AtlasCBS application.

Variable Name	Definition	Example Value[a]	Equation
LE	ΔG/NHAC	0.50	[1]
BEI	$p(K_i)$, $p(K_d)$, or $p(IC_{50})$/ MW(kiloDaltons)	27	[2]
SEI	$p(K_i)$, $p(K_d)$, or $p(IC_{50})$/(PSA/100 Å^2)	18	[3]
Slope of lines: 10(PSA/MW). Algebraic description: BEI = 10(PSA/MW) SEI; Description: Efficiency plane based on macroscopic physico-chemical properties of the ligand: PSA, MW.			
NSEI	$-\log_{10} K_i$/(NPOL) = pK_i/NPOL(N,O)	1.5	[4]
NBEI	$-\log_{10} K_i$/(NHA)= pK_i/(NHA)	0.36	[5]
Slope of lines: NPOL/NHA. Algebraic description: NBEI = (NPOL/NHA) NSEI. Description: Efficiency plane based on atomic properties of the ligand. Slope of the lines is always a rational number given as NPOL/NHA.			
nBEI	$-\log_{10}[(K_i/NHA)]$	10.25	[6]
mBEI	$-\log_{10}[(K_i/MW(KiloDaltons)]$	11.5	[7]
NSEI	$-\log_{10} K_i$/(NPOL) = pK_i/NPOL(N,O)	1.5	[4]
Slope of lines: NPOL. Algebraic description:			
nBEI = NPOL*NSEI + $\log_{10}$(NHA);			[8]
mBEI= NPOL*NSEI + $\log_{10}$(MW);			[9]
Intercept: $\log_{10}$(NHA) or $\log_{10}$(MW) respectively.			
mBEI	$-\log_{10}[(K_i/MW]$	11.5	[10]
SEI	$p(K_i)$, $p(K_d)$, or $p(IC_{50})$/(PSA/100 Å^2)	18	[3]
Slope of lines: PSA/100. Algebraic description: mBEI = (PSA/100) SEI + $\log_{10}$(MW) Intercept: $\log_{10}$(MW).			

[a]The examples refer to the extensive table presented in Abad-Zapatero (2013) [13] using a hypothetical compound with 1 nM K_i (K_d), affinity; MW = 333 Da.; PSA= 50 Å^2; NPOL =6, NHA=25. If MW = 333, and NHA=25, the approximate mean molecular weight of a typical non-hydrogen, medicinal chemistry, atom is ~13.3 [18]. Constants: RT= 0.594.

Seeing the limitations of atom counting vs. MW to assess the size of a ligand, Abad-Zapatero and Metz introduced in 2005 the concept of Binding Efficiency Index (BEI) [15], using only the pK_i (or equivalent affinity quantity) in the numerator and the MW scaled to a thousand (for convenience) in the denominator. In addition, they suggested an analogous polarity-efficiency combining the same numerator (pK_i or equivalent to make it consistent) divided by the polar surface area (PSA) of the ligand scaled down to one hundred. These pair of combined variables provided two complementary variables (in approximately the same scale) to judge the quality of compounds and fragments and also to represent three critical variables in drug discovery (affinity, size, polarity) in a Cartesian plane (Table 1), preferably SEI (x) and BEI (y).

The concept of an atlas-like representation begins to emerge if one realizes that the ratio of the LEIs defined above (BEI/SEI; y/x) depends only on the properties of the ligand, i.e. BEI/SEI = 10 • PSA/MW (Table 1). Thus, one can think of the points in the SEI, BEI planes as having an angular coordinate (slope of the lines) given only by the physico-chemical properties of the ligand (PSA, MW) and a radial coordinate corresponding to the affinity of the ligands towards specific

targets. Highly polar compounds will map along lines of steep slopes (large PSA/MW) and very size-efficient compounds will map far away from the origin. Polarity increases counterclockwise for the different compounds, from nearly horizontal slopes (hydrophobic) to nearly vertical lines (very polar). This was initially described in the analysis of compounds for PTP1B as related to compounds for other less polar targets [16].

Using this basic concept, several other combinations of variables were explored and the notion of an atlas-like representation was published in 2010, presenting a very intuitive and easy-to-understand combination of variables, referred to as (NSEI, nBEI; see Table 1 for definitions). Details can be found in the initial publication [17] and also in the more recently extended description of these ideas [13].

Briefly, NSEI is an efficiency index related to the number of polar atoms (understood as the sum of Nitrogens and Oxygens: NPOL=N+O) and nBEI is a variation of the BEI index, where the denominator is NHA (as in the original Hopkins publication [18]) but the logarithm operation is taken after the ratio of affinity to the number of non-hydrogen atoms (see Table 1 for definitions):

$$NSEI = -\log Ki / NPOL\ (N+O) = pKi / NPOL \quad \text{(equation 4)}$$

$$nBEI = -\log [(Ki/NHA)] \quad \text{(equation 6)}$$

From these two equations, the appearance of Cartesian planes (NSEI, nBEI; x,y) can be inferred by eliminating the affinity variable from the two equations (Abad-Zapatero, 2013 [13]. Appendix A). Substitution of the value of log Ki from equation 4 in equation 6, yields:

$$nBEI = NPOL \bullet NSEI + \log (NHA) \quad \text{(equation 11)}$$

This linear equation in y (nBEI) and x (NSEI) shows that the ligands present in the dataset will be arranged in the planes in lines of slope NPOL (=N+O) and intersects log (NHA). More polar compounds will be along lines of increasing slope, the most efficient being further away from the origin. This appearance has been illustrated in Figure 1.

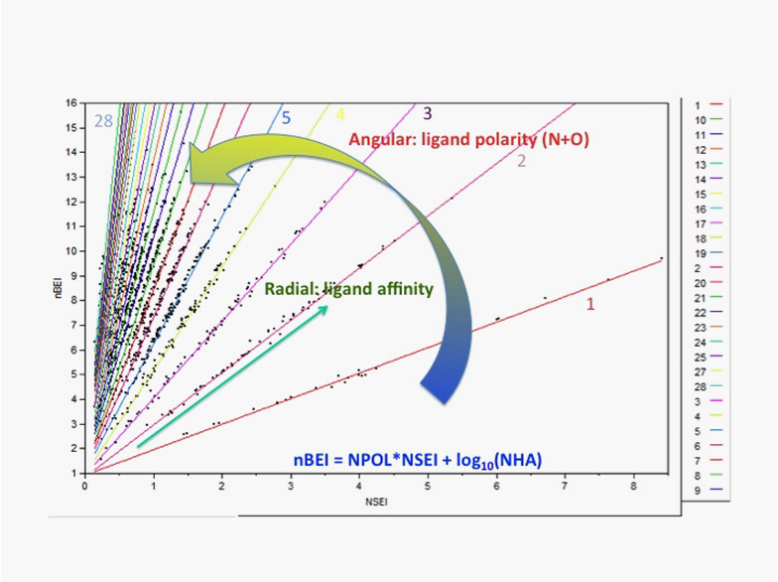


Figure 1: PDBBind in AtlasCBS. The content of a limited SAR-Database (PDBBind 2007) in the NSEI, nBEI (x,y) plane to highlight the distinct pattern of lines of slope NPOL and intersect $\log_{10}$(NHA), as indicated in equation 11. The polarity increases counterclockwise as NPOL increases. The different lines were modeled statistically using the statistical software JMP [17] and the corresponding colors for each value of NPOL are shown on the right panel.

THE ATLASCBS WEBTOOL

The web tool was designed and implemented as a proof of concept having the most essential capabilities to make it practical. Details have been published and will be reviewed here briefly only to provide the basis for the introduction of the new tools using the KNIME workflows in the next section.

The design architecture is based on three layers: *database, application server* and *client*. The application is between the client and the database preventing direct access to the database by the client. The database server is MySQL(v.5). The data were extracted from the three SAR-databases (BindingDB, PDBBind, and ChEMBL) using an off-line custom Java application that extracts the target, molecule activity (K_i, IC_{50} or K_d) and calculates the basic properties needed (MW, SMILES, PSA, NHA, and NPOL) to compute the ligand efficiency indices. This *database* is interrogated by the *client* via the Java Virtual Machine (JVM) *server*, resulting in different plots and maps as well as other options described in the publications. Of notice is the capability of accessing the affinity data from the compounds associated with a specific target by entering the PDB access code of the protein-ligand complex.

The entry portal and a few snapshots of different screens introducing the webserver application are shown in Figures 2-4. The entry portal contains five tabs shown in Figure 4 (upper left). Briefly, their functionalities can be described as follows.

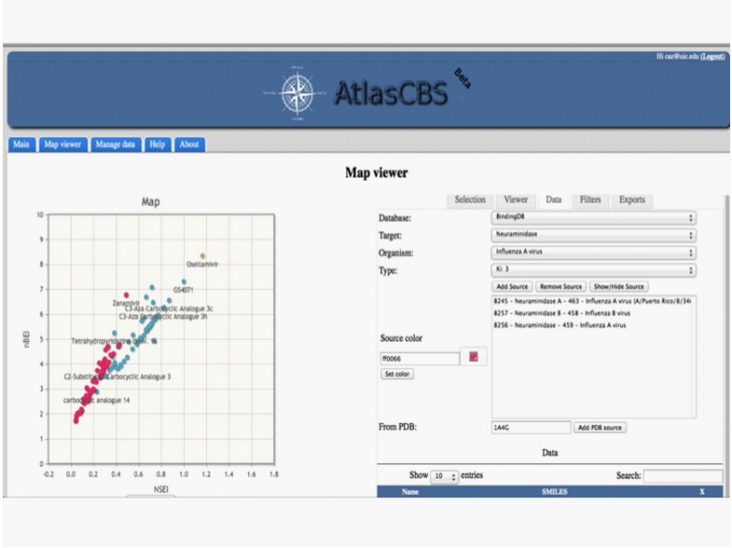


Figure 2: Mapping of Neuraminidase in efficiency space. Example of the AtlasCBS mapping of the data available for Neuraminidase within the **Map Viewer** window. Three different datasets are represented (see working window in the center right) from BindingDB. One of the sets was extracted entering the PDB code 1A4G (see the **Add PDB source** window) that corresponds to the complex of Neuraminidase and Zanamivir.

Main. Contains the basic information about the application and its purpose with references, contact information and entry to the **Help** tab. The application opens on a register (login) window (Figure 4) but it is not necessary to register unless the user wishes to upload personal data. Clicking on **Main** reverts to the main entry portal with basic information.

Map Viewer. This is the central element of the AtlasCBS server (Figs. 2,4). The page contains a plotting window (upper left) with the various definitions of LEIs below (BEI, SEI, etc.) and five major tabs: **Selection, Viewer, Data, Filters** and **Exports** at the top. There are three minor grey tabs below that are used to add/remove targets and data to/from the map window (see below). The **Data** tab is auto selected (whitened) upon opening and shows the various databases available at the time the application was installed: BindingDB, PDBBind and ChEMBL. Upon registering a '**Userset**' database will also be available. The data in those databases have been extracted from the databases as indicated above and organized in the internal MySQL database that is part of the application.

For each database, the target, organism and type of affinity data available (K_i, IC_{50}, K_d are considered only) are visible. After selection, one needs to press the Add Source tab to include the selection in the working window and a plot of nBEI vs. NSEI is presented by default. This is considered to be the simplest efficiency plane to interpret, as explained above. Obviously, other planes can be also seen by changing the x,y axes opening the **Viewer** tab. The definition of the various variables is listed on the lower part of the map panel.

The response will be slower for targets and organism that are heavily populated in the databases, such as HIV-1 Protease with over a thousand (1346 in the current version) entries for human immunodeficiency virus type 1, Ki values (Figure 3). The other tabs (**Remove, Show/ Hide**) permit deleting a set from the working window and display or blanking off the plot of that particular set after selection, respectively. The colors of the different target-ligand pairs can also clicking on the '**Set color**' square button after selecting the appropriate data set.

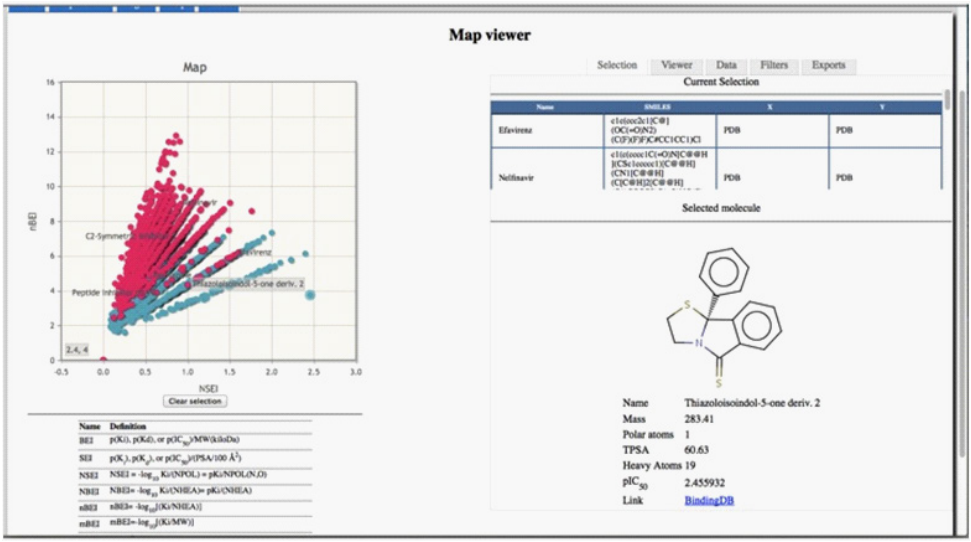


Figure 3: Mapping of HIV Protease and Reverse Transcriptase. Map viewer page of the AtlasCBS server after using the **Add PDB source** option for PDB codes 1OHR (Nelfinavir, HIV Protease) and 1FK9 (Efavirenz, HIV Reverse Transcriptase) within the Selection option that displays the structure of the compound last hit. Colors were reset with the set color option. The wide range of overlapping chemical space encompassed by both targets can be appreciated.

A particularly useful addition to the server application that was added in collaboration with the PDB is the possibility of accessing (plotting) data related to a specific target and compound directly via the PDB entry code. Below the working window there is a '**From PDB**' space that can be used to enter a PDB access code to extract data for the available set of compounds for a certain target connected to the 3-D structure available in the PDB. This feature can be tested by using the PDB access code '1A4G' and pressing the '**Add PDB source**'. Almost immediately, the target Neuraminidase B-458-Influenza virus will appear in the working window and the compounds will be displayed on the corresponding NSEI-nBEI plot with the annotation corresponding to the location of 'Zanamivir' in the plane (Figure 2).

Unfortunately, due to the limitations of the content of the existing tables in the server database, not all PDB entries codes can be accessed so effectively. A suitable message will appear to alert the user if the search failed to extract any affinity data from BindingDB. Examples are presented in Figures 2 and Figure 3.

Login. Is required if the user wishes to upload proprietary data to view it in the context of what is available in the publicly available databases for the target of interests. Upon login in, the new database '**UserSet**' will be available in the data window after the ones currently available. The selection process and mapping in the different planes will be same as the required for the pre-existing public databases. After **Login** this tab is replaced by **Manage Data** (see Figure 2) and allows the uploading of user data into the **UserSet** database that appears after the core databases contained in the AtlasCBS. Details of the input format to upload data are given in the Help tab and in the publications [13].

Help. Basic information on how to use the AtlasCBS server is available here.

About. Contains information related to the institutions, people and supporting institutions that were involved in the AtlasCBS project.

It is impossible in this brief overview to illustrate or even introduce all the functionalities of the current AtlasCBS server. The few images (Figures 2-4) discussed have been added to illustrate the discussion. The interested reader is referred to the original publication describing the web tool application [19] and is encouraged to use the application server https://www.ebi.ac.uk/chembl/atlascbs/intro.jsp.

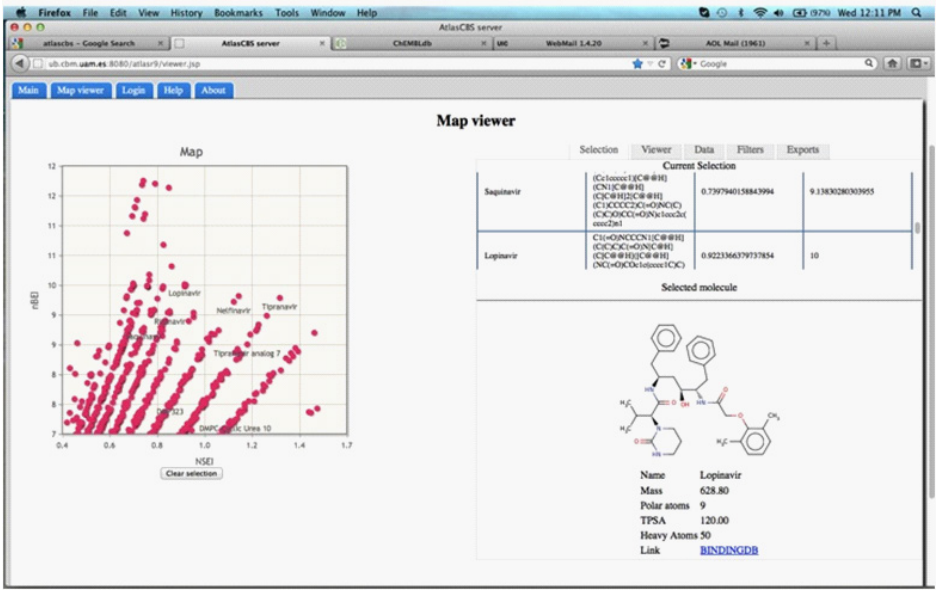


Figure 4: Close-up of the region of chemico-biological-space containing drugs against the HIV Protease. This window was selected and scaled-up with the mouse from the previous figure, after blanking the Reverse Transcriptase compounds. The migration in the efficiency from the early 'peptide mimetic' (Saquinavir, Ritonavir and Lopinavir) compounds towards the more 'drug-like' (Nelfinavir or Tipranavir), while maintaining or improving the size-efficiency, can be seen. Figures 3-4, adapted from (Abad-Zapatero, 2013) [13].

USAGE

The universes of chemistry and biology are overwhelming. Understanding, absorbing or even digesting the myriad of biologically active chemical entities is extremely difficult. However, mapping them in relation to the targets upon which they act and ranking them according to their individual physico-chemical properties creates images that are easier to interpret. This is the power of the AtlasCBS notion and even though the current implementation has limitations, the concept has merit. The author has found the webserver particularly useful at two levels. First, it is convenient to have an overview of the content SAR-data available in the most prominent databases. This has been illustrated in the publications, particularly in the monograph focused on the use of efficiency indices to map chemico-biological space [13]. The plots produced by the server give a rapid birds-eye view of the chemical matter available for each particular target and an easy way to assess the quality of the available compounds in terms of three key variables: affinity, size and polarity. Consistently, it is found that compounds that correspond to marketed drugs typically occupy the upper right quadrant of the plots, where both the efficiency-per-size and efficiency-per-polarity are optimized. Since the databases contain affinity value for other related targets (often mutants, or homologous targets in different organisms), frequently is also found that the same compound has been tested for efficacy or specificity in other targets. Given the definition of the NSEI, nBEI (x,y) variables, the same compound will appear along the same line (same slope NPOL, $\log_{10}$[NHA]) for different targets, with the intended target occupying the position furthest from the origin (highest affinity) and the 'anti-target' the lowest, closest to the origin. The Cartesian distance between the two can be used as a quantitative measure of specificity.

In addition, the analysis of limited sets within those databases and other publications, for instance the 'lead-drug' set analyzed by Perola (2010) [20], has suggested that within the efficiency planes produced by the AtlasCBS server it is possible to 'extract' a general direction, like a 'compass bearing', to guide the drug-discovery process for various targets. I am referring to a general 'North-East' direction in the efficiency planes defined by polarity-efficiency (x-axis) and size-efficiency (y-axis). Moving towards the NE, the successive compounds move towards regions of chemico-biological space where the three-variables (affinity, size and polarity) are optimized and result in compounds with high probability of being successful upon further development. Specific examples of the use of the AtlasCBS server for these applications can be found in Chapters 7,8 of Abad-Zapatero, 2013 [13].

LIMITATIONS

Inevitably, the current implementation has limitations that need to be addressed in order to take the AtlasCBS to the next level of design, implementation and effectiveness to make it a powerful global resource to expedite drug discovery. This overarching goal can be reframed in three specific improvements listed in order of importance and priority.

1. Expedite the updating of the MySQL database that is a critical part of the server and that is now updated 'off-line' by a time consuming and inefficient procedure.

2. Improve the linking between the AtlasCBS and the Protein Databank (PDB) and possibly other database to facilitate client queries and access to the combined data.

3. Expand the number of ligand properties extracted and computed from each compound in the SAR databases, namely BindingDB, ChEMBL and PDBBind to include pharmacokinetic data (Log P, Log D, among others) to relate efficiency indices to pharmacokinetic properties. This is a critical step in order to use LEIs as optimization variables in the future and also to critically assess the value of LEM [14] [21].

Alternatively, given the developments in the computing and programming world, it is possible that a different approach might be more effective and practical in the long term. There is a critical issue that needs to be considered first. Is it necessary to keep the AtlasCBS server as a central entity, with a proper identity, installed and maintained within a central public facility as it is currently at the EBI server? Or is it better to make it a more 'democratic' tool where the mapping of chemico-biological space is available within any rendition of SAR-databases? Are the two approaches necessary and do they complement each other? Or are they exclusive of each other? The second option will permit the community to explore fully these concepts and ideas by incorporating them into their workflows and decision-making. Time and the ingenuity of the future generations of drug discoverers will provide an answer to these questions.

FUTURE DEVELOPMENTS AND APPLICATIONS: MyCHEMBL

The initial AtlasCBS web tool[19] was conceived as a proof of concept and illustration of the possibility of having access to the content of three SAR-Databases (BindingDB, ChEMBL and PDBBind) with a series of tools to represent, map and analyze the content of those databases within a LE framework. The web tool also allowed the combination of publicly available SAR data with separate, user uploaded, and project related databases. Some examples have been illustrated in the previous sections.

Given the rate at which the public and private affinity-containing databases is constantly growing, updating the MySQL database associated independently with the AtlasCBS application presented serious problems given the limited resources of the team that initially developed the application. An expanded and updated combined database (ChEMBL, BindingDB, PDBBind and User Set), within the same web server environment, including additional efficiency indices and expanded capabilities for data uploading by the user is still being contemplated.

Simultaneously, the various SAR-Databases are also developing their own strategies to handle issues such as the increase number of entries, frequency of updates, user support and the efficient use of the new computational and internet-based resources, including availability of powerful portable devices.

Within this context, the ChEMBL team developed the independent, virtual machine and user controlled, application named MyChEMBL that proved to be an effective and practical way of making the entire content of the database available to the user and including an initial set of tools to interact and exploit its usage.

MyCHEMBL is an open virtual machine implementation of open data that includes chemoinformatics tools. It has been described in two recent publications [22,23]. Basically, it consists of the following elements: i) a linux (Ubuntu) Virtual Machine (VM); ii) a PostgreSQL version of the corresponding version of the ChEMBL (currently ChEMBL19); iii) the latest version of the RDKit chemoinformatics toolkit and chemistry cartridge. In addition, it provides local and secure access to the latest ChEMBL web services, interactive IPython notebook tutorials, a PostgreSQL schema browser and KNIME example workflows.

Given these components, MyCHEMBL presents several advantages that are discussed by the authors: no cost and runs locally behind a firewall and therefore security concerns over uploaded data are nonexistent. In addition, the authors make available the source for all the applications and consequently further development is not only easy but also encouraged. In our limited experience, it has proven to be easy to use for novices and experienced researchers and there are plenty of learning resources available within the ChEMBL/EBI environment.

This concept provided the opportunity of a very simple, inexpensive and effective way to develop an 'atlas-like' representation of the updated content of ChEMBL (CHEBML19) using workflows prepared with the public domain application KNIME [24]. This is what we introduce in this brief update on available software tools related to analyze and map the content of SAR-databases as well as how to use them in your own drug-design applications.

The workflows described and illustrated were developed using MyCHEMBL version 19 (MyCHEMBL19) and KNIME 2.12.2. Several examples of the three basic workflows have been deposited in the KNIME 3.0 Example server and are accessible from /050_Applications/05022_ AtlasCBS. They are ranked from simpler to more complex.

Workflow 1. Read a user/confidential data set in CSV format, calculate LE variables and map the resulting chemical matter in efficiency space. The user provides a text file (ex. 11HSD_text.csv) in CSV format containing the following information: compoundID, SMILES, affinity variable type (K_i, IC_{50}, K_d in consistent units, typically nM), value, TargetID. In this example only BEI and SEI are calculated as well as PSA/MW that is useful to show the polarity gradient counterclockwise in the corresponding planes. The workflow reads the information, calculates desired efficiency indices based on simple 'calculating algebraic modules' adding additional colums to the data, and produces a 3D-scatter plot that allows flexibility as to which variables to choose, label and annotate. An example is provided in Figure 5 (top).

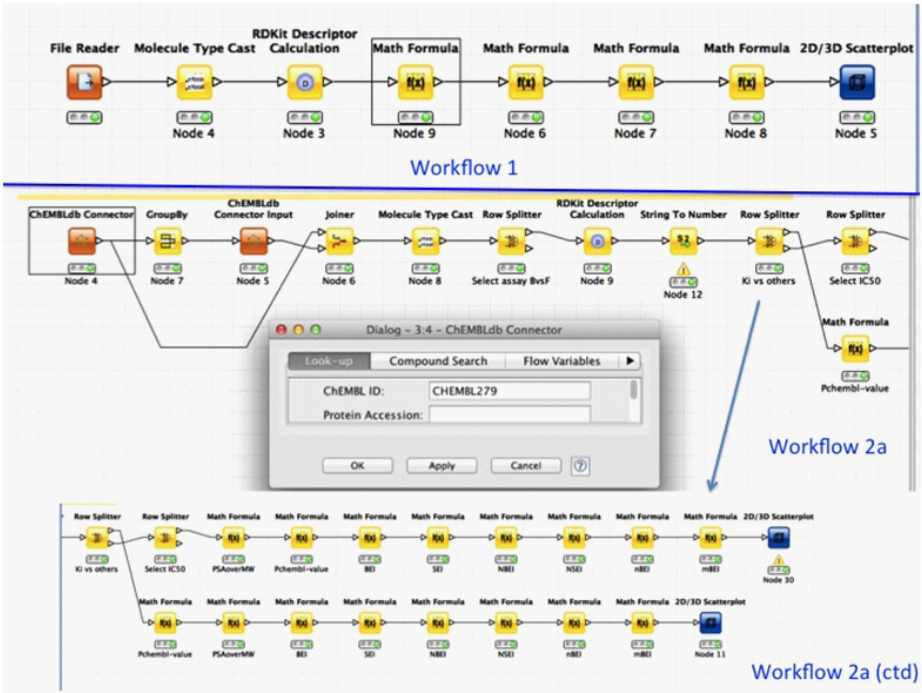


Figure 5: View of KNIME for workflows 1, 2. Workflow 1 (top) consists simply in reading a user prepared input file in CSV format with basic information. It can also be a downloaded file from ChEMBL or any other SAR-Database for any specific target. For this example only three new variables are calculated: BEI, SEI and 10(PSA/MW) for convenience in plotting the gradient of PSA/MW counterclockwise. **Workflow 2** (lower part) is divided in two parts: **2a** and **2a (ctd)**. Part **2a**. Data extraction section showing the configuration window for CHEMBLID target 279 corresponding to humanVEGRF. Over 7,000 compounds were extracted. Notice the row splitter in the top middle to separate two different types of assays: B (Binding) and F (Functional) part **2a (ctd)**. The LEIs calculations for all the indices listed in Table 1. LE is equivalent by an approximate conversion rational factor to BEI/LEI ~54 (Abad-Zapatero, 2013) [13]. Using the K_i values from ChEMBL LE can be easily calculated from equation 1 (Table 1) as well as other efficiency indices (e.g. LLE).

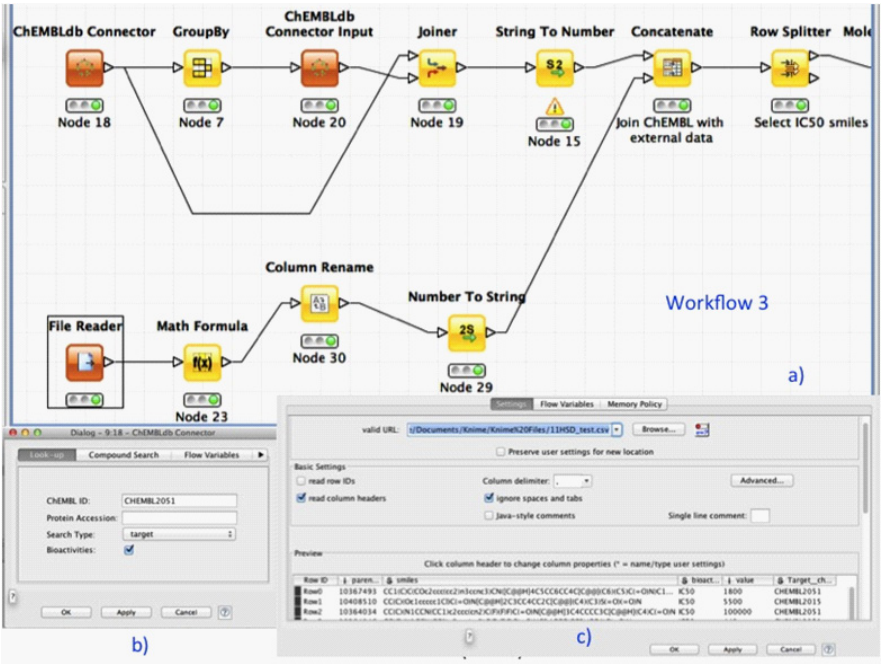


Figure 6: KNIME workflow 3. The workflow combines data extraction from ChEMBL19 as in workflow **2a (top)** with data read from an external file (below) **(6a)**. The configuration file for CHEMBL target ID 2051 (Neuraminidase) is shown **(6b)** as well as the input variables/ format for the external data (example: 11HSD_test.csv) in CSV format **(6c)**. The image represents only the workflow up to the 'row splitter' selecting the IC_{50} values. Calculations of the LEIs are as in the previous workflows (consult the comments on the deposited workflow). The key node for this operation is the one named 'Concatenate' on the upper right. Target ID numbers can be easily consulted and identified using the search options of the ChEMBL website.

Workflow 2. Use the available MyCHEMBL data extractor to extract all the available data from a certain target(s), calculate desired efficiency variables and view and explore in the wider AtlasCBS context (Figure 5, lower portion). Configuration box for Target humanVEGFR2 (CHEMBL279) is shown.

Workflow 3. Read an external data set, calculate LE variables and merge with existing data for an evaluation and comparison of different series. A simple example is presented merging data from two different targets. (Details are documented in the KNIME 3.0 workflow example server). The workflow is illustrated in Fig. 6. Resulting plots for Workflows 2 and 3 are presented in Figure 7 and Figure 8. Data extracted for JAK1 kinase (ChEBML2835) in Fig. 7 and a combination from two targets human VEGFR2 (CHEMBL279 more than 7,000 entries) and a previously prepared external set (11HSD_text.csv, see workflow 1) is presented in Figure 8.

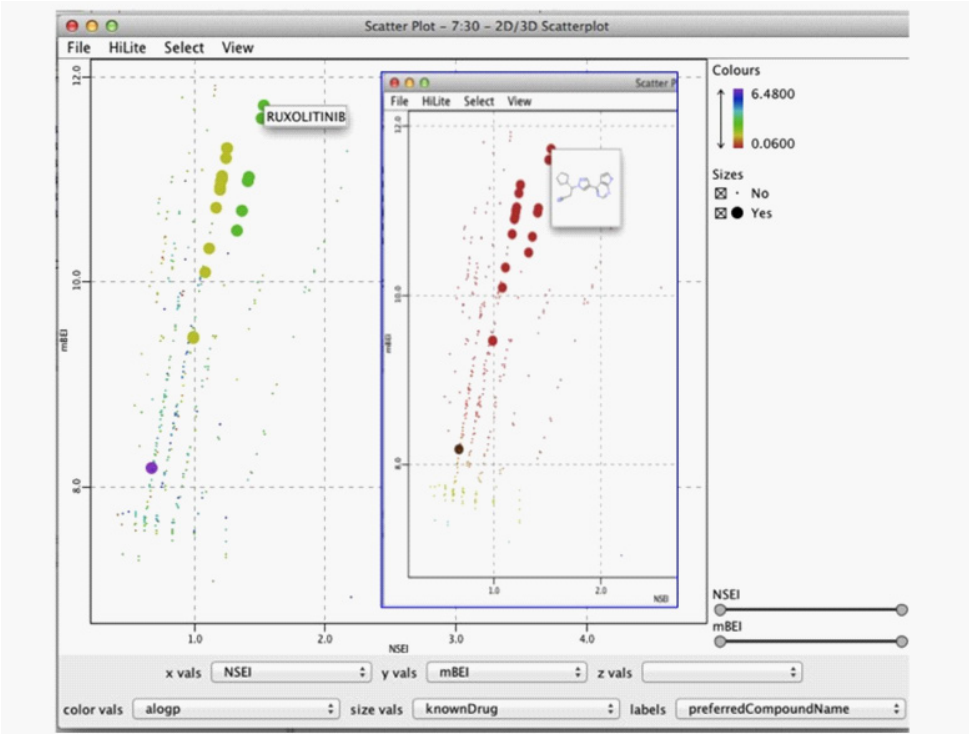


Figure 7: Extracting and mapping data in AtlasCBS. Plotting of data extracted from ChEMBL for JAK1 Tyrosine kinase (ChEMBL2835) annotated by the compound preferred name included in the dataset. Color gradient is based on the available value of 'alogp' (see upper right panel). Insert shows the same data but the labeling is now done by the chemical structure as stored in the SMILES column. These examples illustrate the versatility of the Scatter 2D-3D module of KNIME.

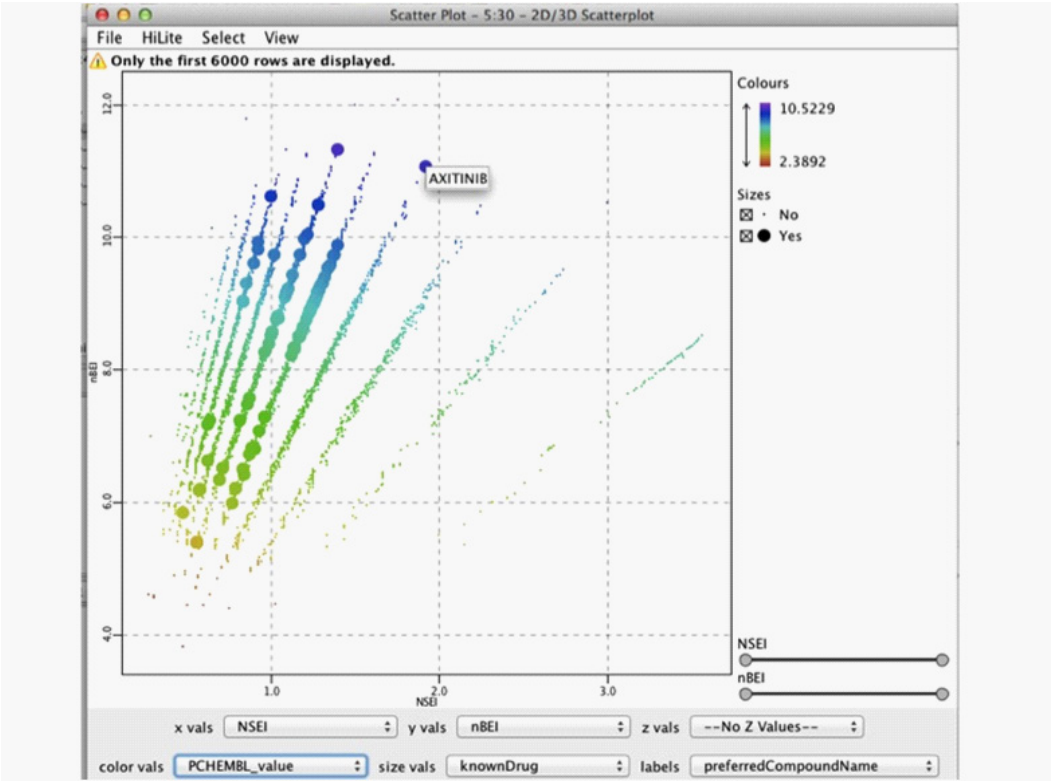


Figure 8: Extraction of large data set from MyChEMBL19. Mapping the chemistry available for human VEGRF2 with over 7,000 entries. The color gradient is based on the pKi affinity value (PCHEMB_value) read from the dataset. Notice the change from low to high values along the vertical axis. Large circles correspond to existing drugs as illustrated in the right lower panel ('Size') in Figure 7 and Figure 8. The preferred compound name for 'AXITINIB' is shown.

Simple options within the 2D-3D scatter plot in KNIME make it possible to change the corresponding axes in the Cartesian 'efficiency planes' that are equivalent to the ones produced by the AtlasCBS application; 3D plotting is also possible and is left to the discretion of the user; it is not illustrated here. The different points in the plots (each corresponding to a target-ligand pair, like in the AtlasCBS) can be annotated and colored based on any of the available variables in the original ChEMBL data set or in the combined set. The workflows with appropriate documentation have been deposited in the KNIME example server http://www.knime.org/example-workflows (for KNIME 3.0 and above).

These workflows are presented here only as an introduction and to impress upon the reader how simple it is to graphically organize and represent the content of SAR-databases using open access tools. The substantially large collection of modules currently available on KNIME [24] will permit the individual tailoring of drug-discovery approaches and strategies to expedite and

succeed in this multi-parameter optimization problem [24-27]. In particular, the author expresses his interest in critically assessing the value and use of LEIs and LEM as variables and parameters to drive and optimize drug design. Quite possibly, the broader dissemination and usage of these concepts and ideas will permit the testing of approaches, algorithms and workflows in a wider range of targets and biomedical problems [28]. Perhaps, workable and effective solutions might be around the corner for the future generations to discover and implement.

ACKNOWLEDGMENTS

The hospitality of the ChEMBL group at EBI directed by John Overington is greatly appreciated. The guidance and help of Drs. George Papadatos and Mark Davies within the group in the installation and use of the MyCHEMBL application and KNIME modules is fully acknowledged. The comments and suggestions of Dr. A. Morreale, one of the early supporters of the project, as well as Dr. A. Cortés-Cabrera who initially programmed the application are greatly appreciated.

References

1. Chen X, Liu M, Gilson MK. BindingDB: a web-accessible molecular recognition database. Combinatorial chemistry & high throughput screening. 2001; 4: 719-725.
2. Liu T, Lin Y, Wen X, Jorissen RN, Gilson M. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic acids research 00(Database issue). 2006; D1-D4.
3. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res. 2007; 35: D198-201.
4. Wang R, Fang X, Lu Y, Wang S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J Med Chem. 2004; 47: 2977-2980.
5. Wang R, Fang X, Lu Y, Yang CY, Wang S. The PDBbind database: methodologies and updates. J Med Chem. 2005; 48: 4111-4119.
6. Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic acids research 40(Database issue). 2012; D1100-1107.
7. Li Q, Cheng T, Wang Y, Bryant SH. PubChem as a public resource for drug discovery. Drug Discov Today. 2010; 15: 1052-1057.
8. Olah M, Rad R, Ostopovici L et al. WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. In: Schreiber SL, Kapoor T, Wess G, editors. Chemical Biology: From Small Molecules to Systems Biology and Drug Design. Weinheim: Wiley-VCH Verlag CmbH & Co. 2007.
9. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Biol. 2003; 10: 980.
10. Berman HM. The Protein Data Bank: a historical perspective. Acta Crystallogr A. 2008; 64: 88-95.
11. Knox C, Law V, Jewison T, Liu P, Ly S. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res. 2011; 39: D1035-1041.
12. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006; 34: D668-672.
13. Abad-Zapatero C. Ligand Efficiency Indices for Drug Discovery. Towards an Atlas-Guided Paradigm. Philadelphia: Elsevier. 2013.
14. Abad-Zapatero C, Champness EJ, Segall MD. Alternative variables in drug discovery: promises and challenges. Future Med Chem. 2014; 6: 577-593.
15. Abad-Zapatero C, Metz JT. Ligand efficiency indices as guideposts for drug discovery. Drug Discov Today. 2005; 10: 464-469.
16. Abad-Zapatero C. Ligand efficiency indices for effective drug discovery. Expert Opin Drug Discov. 2007; 2: 469-488.
17. Abad-Zapatero C, O Perisic, J Wass, PA Bento, J Overington, et al. Ligand Efficiency Indices for an Effective Mapping of Chemico-Biological Space: The concept of an Atlas-like representation. Drug discovery today. 2010; 15: 804-811.
18. Hopkins AL, Groom CR, Alex A. Ligand efficiency: a useful metric for lead selection. Drug Discov Today. 2004; 9: 430-431.

19. Cortés-Cabrera A, Morreale A, Gago F, Abad-Zapatero C. AtlasCBS: a web server to map and explore chemico-biological space. J Comput Aided Mol Des. 2012; 26: 995-1003.

20. Perola E1. An analysis of the binding efficiencies of drugs and their leads in successful drug discovery programs. J Med Chem. 2010; 53: 2986-2997.

21. Shultz MD1. Improving the plausibility of success with inefficient metrics. ACS Med Chem Lett. 2013; 5: 2-5.

22. Ochoa R, Davies M, Papadatos G, Atkinson F, Overington JP. myChEMBL: a virtual machine implementation of open data and cheminformatics tools. Bioinformatics. 2014; 30: 298-300.

23. Mark Davies MN, George Papadatos, Francis Atkinson, Gerard JP Van Westen, Nathan Dedman, et al. Overington: MyCHEMBL: A Virtual Platform for Distributing Cheminformatics Tools and Open Data. Challenges. 2014; 5: 334-337.

24. Beisken S, Meinl T, Wiswedel B, de Figueiredo LF, Berthold M. KNIME-CDK: Workflow-driven cheminformatics. BMC Bioinformatics. 2013; 14: 257.

25. Jagla B, Wiswedel B, Coppée JY. Extending KNIME for next-generation sequencing data analysis. Bioinformatics. 2011; 27: 2907-2909.

26. Lindenbaum P, Le Scouarnec S, Portero V, Redon R. Knime4Bio: a set of custom nodes for the interpretation of next-generation sequencing data with KNIME. Bioinformatics. 2011; 27: 3200-3201.

27. Mazanetz MP, Marmon RJ, Reisser CB, Morao I. Drug discovery applications for KNIME: an open source data mining platform. Curr Top Med Chem. 2012; 12: 1965-1979.

28. Reynolds CH. Ligand efficiency metrics: why all the fuss? Future Med Chem. 2015; 7: 1363-1365.

Austin Publishing Group

Modeller: An Application for Homology Modeling

Singh R* and Gaur P

Center of Bioinformatics, IIDS, University of Allahabad, India

***Corresponding author:** Singh R, Center of Bioinformatics, IIDS, University of Allahabad, Allahabad 211002, India, Email: raghvendra1986singh@gmail.com

Published Date: December 01, 2016

ABSTRACT

One of the major goals of Bioinformatics is to understand the relationship between amino acid sequence and three –dimensional structure. The biological role of protein can be determined by its function, which in turn, is largely determined by its structure.The role of structure for biological sciences and research has grown considerably since the advent of systems biology. The task of functional characterization of a protein sequence is one of the most frequent problems in biology which is usually facilitated by accurate three-dimensional (3-D) structure of the studied protein. In the absence of an experimentally determined structure of protein, comparative or homology modeling can provide a useful 3-D model, related to at least one known protein structure. Comparative modeling predicts the 3-D structure of a given protein sequence (target), based primarily on its alignment to one or more proteins of known structure (templates).The goal of protein structure prediction is to estimate the spatial position of each atom of protein molecules from the amino acid sequence by computational methods. Before 1993, protein modeling was done through a semiautomatic and multi-step fashion, including distinct modeling procedure for SCRs (Structurally conserved regions), SVRs (Structurally variant regions), and side chains. In this chapter, MODELLER, the first automatic protein structure prediction tool, developed by Sali and Blundell [1], is elaborated along with the description of download and installation.

Keywords: Modeller; Comparative Modeling; Structure Prediction

Abbreviations: SCRs - Structurally conserved regions; SVRs - Structurally variant regions; HM- Homology Modeling; PDB- Protein Data Bank; PDF- Probability Density function

INTRODUCTION

The basic assumption for protein structure prediction is associated with similarities shared by sequences, but known sequences and structures have shown us the divergence from this basic assumption. Globin family is the best known example for this. Globin family includes haemoglobin, myoglobin and plant leg haemoglobin which adopt the same overall structure and carry out same function of oxygen transporting by same mechanism, but have a very low sequence identity (below 20%). This shows the importance of protein structure, as structure is much more conserved than the sequences ever are. With the increasing scope and utility of Bioinformatics, finding the solution to the problem of protein structure prediction started to become easier and pursuable. Solved structure of proteins with help of techniques like NMR and X-ray Crystallography made a base for Bioinformaticians to predict the structure of unknown proteins. PDB [2], which is the repository for solved structures of proteins, was used as a source and with the help of key procedures of Bioinformatics like sequence alignment, fold recognition, fragment based structural assembly and multiple structural refinement, the prediction technique of protein structure enjoyed a considerable and perceivable success. Protein structure prediction broadly can be categorized as: ab initio folding, comparative/ Homology modelling and threading.

Homology modeling (or comparative modeling) is considered as the most successful category of protein structure prediction so far. It is based on our understanding of protein evolution with which we can draw two inferences; (1) proteins that have similar sequences usually have similar structures and (2) protein structures are more conserved than their sequences. HM is based on the notion that new proteins evolve gradually from existing ones by amino acid substitutions, additions, and/or deletions and that the 3D structures and functions are often strongly conserved during this process. Many proteins thus share similar functions and structures and there are usually strong sequence similarities among the structurally similar proteins. Strong sequence similarity often indicates strong structure similarity, although the opposite is not necessarily true as we have discussed with the case of Globin protein family.

A QUICK REVIEW OF HOMOLOGY MODELING

In HM, the sequence of the protein of interest (target) is matched to an evolutionarily related protein with a known structure (template) in the PDB to construct protein structure. Thus, only those proteins having appropriate templates can be modeled by homology modeling. For the protein targets where templates with a sequence identity > 50% are available in the PDB, the homologous templates can be easily identified with the sequence-template alignments conducted precisely.

Below are the brief standard processes involved in HM:

- Using the unknown sequence as a query to search for known protein structures.
- Producing the best possible global alignment of the template sequence(s) and unknown sequence.
- Building a model of the protein backbone and taking the backbone as a model.
- Using a loop- modeling procedure in gap region of target or template.
- Adding side-chains to the model backbone.
- Optimizing positions of side-chains.
- Optimizing the structure with energy minimization or knowledge-based optimization.

AUTOMATIC PROTEIN MODELING USING MODELLER

As stated earlier, protein modeling was not fully automatic until1993, and done through semi-automatic fashion which included different procedures for SCRs, SVRs, and side chains. Modeller is the first automatic, full-atom protein modeling computer program.

What is Modeller?

The first automatic protein modeling program- Modeller, developed by Sali and Blundell (http://salilab.org/modeller), is a very popular and widely used modeling package for homology or comparative modeling of protein three-dimensional structures. To compute the structure of the target protein, MODELLER optimally satisfies spatial restraints derived from the alignment of the target protein sequence and multiple related structures [3]. As input, it takes alignment of a sequence to be modeled and automatically generates 3-D model that contains all non-hydrogen bonds. It also performs the additional tasks like de novo loop modeling, optimization of different models of protein structure, clustering, protein structure comparison, sequence database searching and protein sequence and/or structure alignment, etc. It does so, either by using distance geometry or optimization techniques, obtained from the alignment of the target sequence with the template structures. Modeller has no graphical interface of its own, but the command-line environment is comfortable to work with.

Obtaining and Installing Modeller

Modeller is written in Fortran 90 and Python is its control language. Hence, Python scripts are input scripts to Modeller. Although knowledge of Python is not mandatory to run Modeller, it is useful in performing more advanced tasks. Precompiled executable for Modeller can be downloaded from aforementioned website.

Required Operating System

Modeller can run on various operating Systems like Unix/ Linux, Apple Mac OS X andMicrosoft Windows.

Required Software

An up-to-date web-browser is required, such as Internet Explorer, Firefox and Chrome etc.

Installation

Installation of Modeller depends on its operating system .The procedures for different operating systems differ slightly. Detailed instructions for installing Modeller on machines running on different operating systems can be found at http://salilab.org/modeller/ release. html. Basically, the following steps are achieved for installing Modeller on Unix/Linux operating system.

- Go to http://salilab.org/modeller/download_installation.html.
- Download the distribution by clicking on the link indicating for Unix/ Linux.
- Obtain a key from URL http://salilab.org/modeller/registration.html. A valid license key, distributed free of cost to academic users, is required to use Modeller.
- Open a terminal or console and change to the directory according to the containing the downloaded zip file of Modeller latest version.
- Unpack the downloaded file with the gunzip / tar .command
- A new directory is created that contains files needed for installation. Move into that directory to install the Modeller with the help of command- ./Install
- The installation script will prompt the user with several questions and suggest default answers. Follow and answer accordingly and begin the installation.

HOW A MODEL IS BUILT BY MODELLER

Modeller initiates itself with multiple sequence alignment between the target sequence and the template protein sequence(s). Basically this alignment is the input to the program. A set of spatial restraints is generated by Modeller using the template structures. These restraints are generally formed on the basis of statistical analysis of the relationships between many pairs of homologous structures. This statistics contributes quantitative description of how much various properties are likely to vary among homologous structures. Modeller effectively limits the number of conformations the model can assume by applying these spatial restraints. In fact, the correlation between two equivalents, for instance - Cα – Cα distances, or between equivalent main chain dihedral angles from two related proteins is expressed as a probability density function

(PDF) which can be used directly as spatial restraint. This PDF-based restraint allows you to build a structure that isn't exactly like the template structure. Instead, in this manner the structure of the model would be allowed to deviate from the template but only in a way consistent with differences found between homologous proteins of known structure. For instance, if a particular dihedral angle in the template structure has a value of - x degree, the applied PDF-based restraint should allow the dihedral angle to assume a value of x plus or minus some value. This value is determined probabilistically by what is observed in known pairs of homologous structures, and also according to the form of the probability density function. Along with the Homology-based spatial restraints, a chemical restraint in form of force field (for example Charmm) is also applied. The use of force field enforces the control over proper stereochemistry [4], so that the model structure could not violate the rules of chemistry to satisfy the spatial restraints derived from the template structures. Next, these chemical and spatial restraints applied to the model are combined in a function which is called an objective function. Finally, the model is obtained by optimizing the objective function in Cartesian space. Also by varying the initial structure, several slightly different models can be calculated. The variability among these models can be used to estimate the errors in the corresponding regions of the fold. It could be said that the model building procedure of Modeller is similar to that of the structure determination by NMR spectroscopy.

HOW TO USE MODELLER

Information about downloading Modeller and installing it on Unix/Linux has already been mentioned. For more information, 'Readme' file available with the distribution can be looked upon. A simple demonstration and information on additional tools can be found at same aforementioned website of Modeller. The most basic use of Modeller in HM includes PDB atom files of known protein structures and their alignment with the target sequence to be modeled. The alignment may also contain very short segments such as loops, secondary structure motifs, etc.The output is a model for the target that includes all non-hydrogen atoms. Modeller is multifunctional and has built-in commands like SEQUENCE_SEARCH which searches for similar sequences in a database of fold class representative structures, MALIGN3D which aligns two or more structures, ALIGN that aligns two blocks of sequencesand CHECK_ ALIGNMENT which evaluates an alignment to be used for modeling. These built in commands help you prepare your input. There are other commands too like SUPERPOSE, ENERGYand COMPARE_SEQUENCES etc that need to be submitted to Modeller via a script that calls that command. Full details of writing scripts are described in the Modeller manual.

What the Inputs are

As input, Modeller takes three kinds of files; Protein Data Bank atom files with coordinates for the template structures, the alignment file with the alignment of the template and target sequence, and a script file that instructs Modeller what to do.

Each PDB atom file is named code.atm where code is a short protein code, preferably the PDB code. The file extensions could also be in pdb and ent format instead of atm. The code must be used as that protein's identifier throughout the modeling. The preferred format for alignment file is like the PIR database format. A sample alignment in the PIR format is shown here.

>P1;sirtuin_2

sirtuin2:sm:wt:::::::0.00

MSFDFLGIKKALFGDNTPRPELKSLNIEGVAQLIQDGQVNKIITMVGAGVSTAAGIPDFR

SPSSGIYDNLEDFNLPTPNAIFTIDYFRRDPRPFFEIARRLYRPEAKPTLAHCFIRLLHD

KGLLLRHYTQNVDSLERLSGLPEEKLVEAHGTFHTGHCIKCNKQHDFEFMLNEILAKRVP

QCLKCRNVVKPDVVLFGESMPKKFFKNLSSDLNNCDLLIIMGTSLTVLPFCAMIHRVGND

VPRLYINREYNDGSTESGLSSFIMRFMVAGFKQNYMKWGRSDNKRDIFWSGNADDGVVKI

SELLGWKDDLLRLKKETDSRLNEEFLAKKSQDKTNGQ*

Modeller is a command-line only tool, and has no graphical user interface, so it demands a script file (usually Python) containing Modeller commands. In case of not being familiar with Python, examples/codes can be consulted in the examples directory/modeller directory. TOP language which is Modeller's internal language is used by Modeller for scripting the alignment. Modeller can calculate multiple models for any input. Usually, it's preferred to generate more than one model so that each model can be evaluated independently to choose the best final model on the basis of DOAP (Discrete Optimized Protein Energy) score method. Below is the example of script file-

Align2d.py

```
env = environ()
aln = alignment(env)
mdl = model(env, file='1J8F', model_segment=('FIRST:A','LAST:A'))
aln.append_model(mdl, align_codes='1J8F', atom_files='1J8F.pdb')
aln.append(file='seq.fasta', align_codes='sirtuin_2')
aln.align2d()
aln.write(file='sir2_sm_1J8F.ali', alignment_format='PIR')
aln.write(file='sir2_sm_1J8F.pap', alignment_format='PAP')
```

modeling

```
from modeller import *
```

```
from modeller.automodel import *

env = environ()

a = automodel(env, alnfile='sir2_sm_1J8F.ali',

        knowns='1J8F', sequence='sirtuin_2',

        assess_methods=(assess.DOPE, assess.GA341))

a.starting_model = 1

a.ending_model = 10

a.make()
```

How to Run Modeller

Modeller is run by giving the command 'mod scriptname'. If you name your script fyz.top, the command is mod fyz.

- To run Modeller you basically need to-
- Open a command line prompts (according to your Operating system; Linux/Unix or Windows or Mac OS X)
- Change to the directory containing the script and alignment files you created earlier, using the 'cd' command.
- Run Modeller itself by typing the commands, full details of which can be found in documentation / manual of Modeller.

An example of the SmHDAC1[5] and SmSirt2 [6] of Schistosoma mansoni are shown in the figure 1 below:

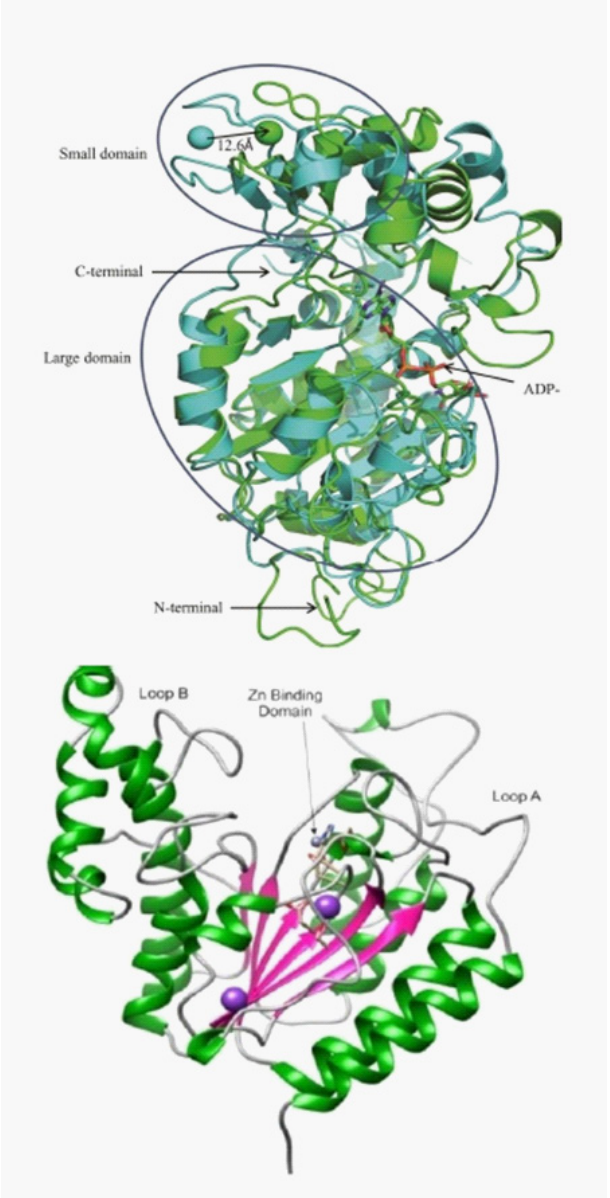


Figure1: An example of the template-based modeling of SmSirt2 (left) and SmHDAC1 (right) by Modeller.

PREDICTING THE MODEL ACCURACY

Estimating the accuracy of a model is the most important factor in the absence of the known structure. A model, calculated using a template structure sharing more than 30% sequence identity is indicative of an overall accurate structure. However, when the sequence identity is lower, it is preferred to check the template used for modeling. When similarities are low, there is high

probability of the error with the alignment step also, making it difficult to distinguish between an incorrect template and an incorrect alignment with a correct template as well. There are several methods that can check whether or not the correct template was used for the modeling. These methods use 3-D profiles and statistical potentials [7][8][9] to access this information. Some examples of these programs include VERIFY3D [8], HARMONY [10], Prosa2003 [11][12], TSVMod [13], ANOLEA [14], DFIRE [15], DOPE [16] , QMEAN local [17], and SOAP [18].

It is notable that a model based on sequence identity >30% does not guarantee its accuracy. Here other factors, including the environment, can strongly affect the accuracy of a model. We can take the example of some calcium-binding protein that undergoes large conformational changes when bound to calcium. If a calcium-free template is used to model the calcium-bound state of the target, the model would be prone to be incorrect even with the good target-template similarity or accuracy of the template structure [19].

Additionally, there are programs such as PROCHECK [20] and WHATCHECK [21] that evaluate the stereo-chemistry of the model which includes bond-lengths, bond-angles, backbone torsion angles, and non-bonded contacts. PROCHECK can check the overall as well as residue by residue geometry. Tools of WHATCHECK produce easy to understand report. Homepages of both the program can be consulted for more information.

SUMMARY

Modeller calculates comparative models and achieves all necessary steps of homology modeling. Apart from model building, Modeller can also perform auxiliary tasks, including fold assignment , alignment of two protein sequences or their profiles, multiple alignment of protein sequences and/or structures [22], calculation of phylogenetic trees, and de novo modeling of loops in protein structures [23]. Specifically, users have access to 'ModBase' which is a comprehensive database of comparative models for all known protein sequences detectably related to at least one known protein structure, a web server 'ModWeb' for automated comparative protein structure modeling; and 'ModLoop' which is a web server for automated modeling of loops in protein structures.

Over the past few years, improvements in the techniques and increment in number of known protein sequences and structures resulted in better output of Modeller. There has been a great increase in the accuracy of comparative models [24][25][26], along with the decrement in magnitude of errors in fold assignment, alignment, and the modeling of side-chains and loops. Nevertheless, there`s always a scope for future methodological improvements. Modeling of distortions and rigid-body shifts, as well as detection of errors in a given protein structure model still demand more accuracy.

References

1. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993; 234: 779-815.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. The Protein Data Bank. Nucleic Acids Research. 2000; 28: 235–242.
3. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 2000; 29: 291-325.
4. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B. 1998; 102: 3586-3616.
5. Singh R, Pandey PN. Molecular docking and molecular dynamics study on SmHDAC1 to identify potential lead compounds against Schistosomiasis. Molecular biology reports. 2015; 42: 689-698.
6. Singh R, Singh S, Pandey PN. In-silico analysis of Sirt2 from Schistosoma mansoni: structures, conformations, and interactions with inhibitors. J Biomol Struct Dyn. 2015;.
7. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol. 1990; 213: 859-883.
8. Lüthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature. 1992; 56: 283–285.
9. Melo F, Sánchez R, Sali A. Statistical potentials for fold assessment. Protein Sci. 2002; 11: 430-448.
10. Topham CM, Srinivasan N, Thorpe CJ, Overington JP, Kalsheker NA. Comparative modelling of major house dust mite allergen Der p I: structure validation using an extended environmental amino acid propensity table. Protein Eng. 1994;7:869–894
11. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins. 1993; 17: 355-362.
12. Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 2007; 35: W407-410.
13. Eramian D, Eswar N, Shen MY, Sali A. How well can the accuracy of comparative protein structure models be predicted? Protein Sci. 2008; 17: 1881-1893.
14. Melo F, Feytmans E. Assessing protein structures with a non-local atomic interaction energy. J Mol Biol. 1998; 277: 1141-1152.
15. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. 2002; 11: 2714–2726.
16. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006; 15: 2507-2524.
17. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics. 2011; 27: 343-350.
18. Dong GQ, Fan H, Schneidman-Duhovny D, Webb B, Sali A. Optimized atomic statistical potentials: assessment of protein interfaces and loops. Bioinformatics. 2013; 29: 3158-3166.
19. PawÅ‚owski K, BierzyÅ„ski A, Godzik A. Structural diversity in a family of homologous proteins. J Mol Biol. 1996; 258: 349-366.
20. Laskowski R, MacArthur M, Moss D, Thornton J. PROCHECK-a program to check the stereochemical quality of protein structures. J. Appl. Cryst. 1993; 26: 283–291.
21. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. Nature. 1996; 381: 272.
22. Madhusudhan MS, Marti-Renom MA, Sanchez R, Sali A. Variable gap penalty for protein sequence-structure alignment. Protein Engineering, Design & Selection. 2006; 19: 129–133.
23. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. Protein Sci. 2000; 9: 1753-1773.
24. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 2000; 29: 291-325.
25. Baker D, Sali A. Protein structure prediction and structural genomics. Science. 2001; 294: 93-96.
26. Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, et al. ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res. 2011; 39: 465–474.

Austin Publishing Group

Protein Structure Prediction Using Molecular Homology Modeling

Barbosa LCB*[1] and Carrijo RS[2]

[1]Institute of Natural Resources, Federal University of Itajubá (UNIFEI), Brazil

[2]Institute of Chemistry, São Paulo State University (UNESP) at Araraquara, Brazil

***Corresponding author:** Barbosa LCB, Institute of Natural Resources, Federal University of Itajubá (UNIFEI), Av. BPS, 1303, Pinheirinho, 37500903, Itajubá-MG, Brazil, Tel: 55-16-982088781; Email: luiz_cbb@hotmail.com

Published Date: December 01, 2016

INTRODUCTION

Modern large-scale DNA sequencers currently produce massive amounts of biological sequence data. However, genome sequencing itself does not allow for full understanding of the biochemical and molecular mechanisms involved in a cell. In this context, knowledge about the three-dimensional structure of proteins is highly valuable, allowing for biological processes to be investigated more directly and at higher resolution and with finer detail [1]. Despite community-wide efforts in structural biology, the high degree of difficulty in determining the three-dimensional structure of proteins has generated a large discrepancy between the volume of data generated by genome projects and the number of three-dimensional structures of proteins that are currently known [2].

Currently, three main experimental methods are used to solve three-dimensional structures of proteins: X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and electron microscopy (EM). In X-ray crystallography, crystalline atoms are exposed to a beam of incident

X-rays that diffract in many specific directions. By measuring the angles and intensities of these diffracted beams, a three-dimensional picture of the density of electrons is obtained and the mean positions of the atoms in the crystal as well as their chemical bonds can be determined [3,4]. NMR spectroscopy utilizes the quantum mechanical properties of the atomic nucleus to determine how the atoms are linked chemically and how close they are located to each other [5-7]. NMR spectroscopy is a method that can identify three-dimensional structures of proteins molecules in the solution phase, allowing for investigation of time-dependent chemical phenomena such as reaction kinetics and intramolecular dynamics [6,7]. Despite substantial progress in the methodologies for structural determination of proteins, limitations remain for these two experimental techniques. For example, protein crystals are essential for the use of crystallography and many proteins either do not crystallize or generate crystals that are inadequate for analysis. In addition, NMR spectroscopy has restrictions regarding the size of the protein to be studied and has been limited to relatively small proteins or protein domains. For both methodologies, structure determination can fail due to problems of aggregation and reduced solubility. In electron microscopy, the sample is exposed to a beam of electrons, and the emerging electrons are detected and used to map out the structure of the materials they smashed into [8,9]. Until only a few years ago, electron microscopy was usually not the first choice for many structural biologists due to its limited resolution. Recent works, however, have been changed this scenario producing high resolution models using electron cryo-microscopy (cryo-EM) [9-12]. For any choice among the three techniques discussed, in general, experimental determination of three-dimensional protein structures is expensive and time consuming.

In the absence of experimental methods, computational approaches for predicting three-dimensional structures have been used to obtain information about the structure of proteins [2,13–17]. Structural bioinformatics is an area of computational biology focused on the structure of macromolecules, including DNA, RNA, and proteins [1]. Elucidation of three-dimensional structures of proteins is undoubtedly one of the main areas of research in structural bioinformatics [1]. Currently, computational approaches for predicting three-dimensional protein structures can be divided into four main classes [2,17]: 1) first principle methods without database information [15]; 2) first principle methods with database information [18,19]; 3) fold recognition and threading methods [20-23]; and 4) homology (or comparative) modeling methods [24,25]. A biennial community-wide Critical Assessment of protein Structure Prediction (CASP) experiment evaluates the progress and challenges in state-of-the-art of protein structure modeling techniques [26-29]. The CASP is a competition where researchers are given a set of protein sequences that have known but unreleased three-dimensional structures to use as the input for modeling programs. Three-dimensional solutions are submitted, evaluated, and compared with the known protein structures, which are released after the contest concludes.

This chapter briefly reviews protein structure prediction using molecular homology modeling, with a focus on conceptual methodology.

BACKGROUND

In homology (or comparative) modeling, previously solved structures of related proteins are used as templates [24,25]. This approach is based on the premise that evolutionarily related proteins tend to be similar in their three-dimensional structures (Figure 1). The prediction process consists of fold assignment and template selection, alignment of the target and template sequences, model building, and model evaluation and refinement [16,24,25,30]. If necessary, alignment and model building are repeated until a satisfactory result is obtained. A general flowchart illustrating generic steps in the construction of a model is shown in Figure 2. Several useful programs and servers for performing these steps are listed in Table 1. Protein structure homology modeling has become a routine method for providing structural models in cases where no experimental structures are available and there is at least one closely related protein with an experimentally determined three-dimensional structure.

Homology modeling is the most frequently used methodology in protein structure prediction because it is a very precise and accurate prediction method when a reasonable evolutionary relationship is present [16,25,31]. Furthermore, the reliability and quality of the predicted structures can be estimated. At same time, homology modeling is limited due to the inability to perform prediction of new folds since this methodology can only predict structures of protein sequences that are closely related to other protein sequences of known structures [2].

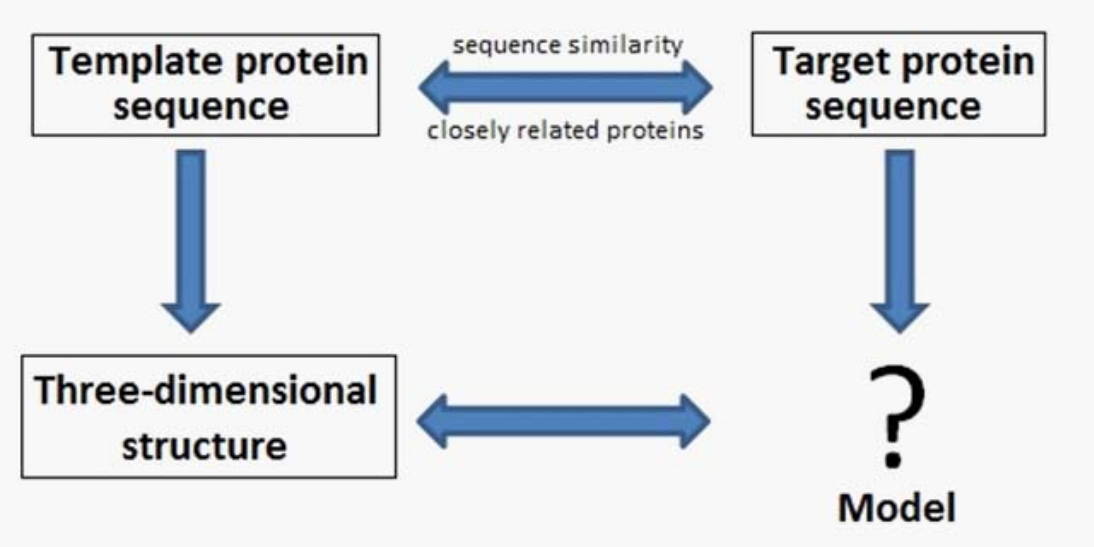


Figure 1: **Premise of molecular homology modeling.** Evolutionarily related proteins have some differences with respect to their amino acid sequences but retain high degrees of structural similarity.

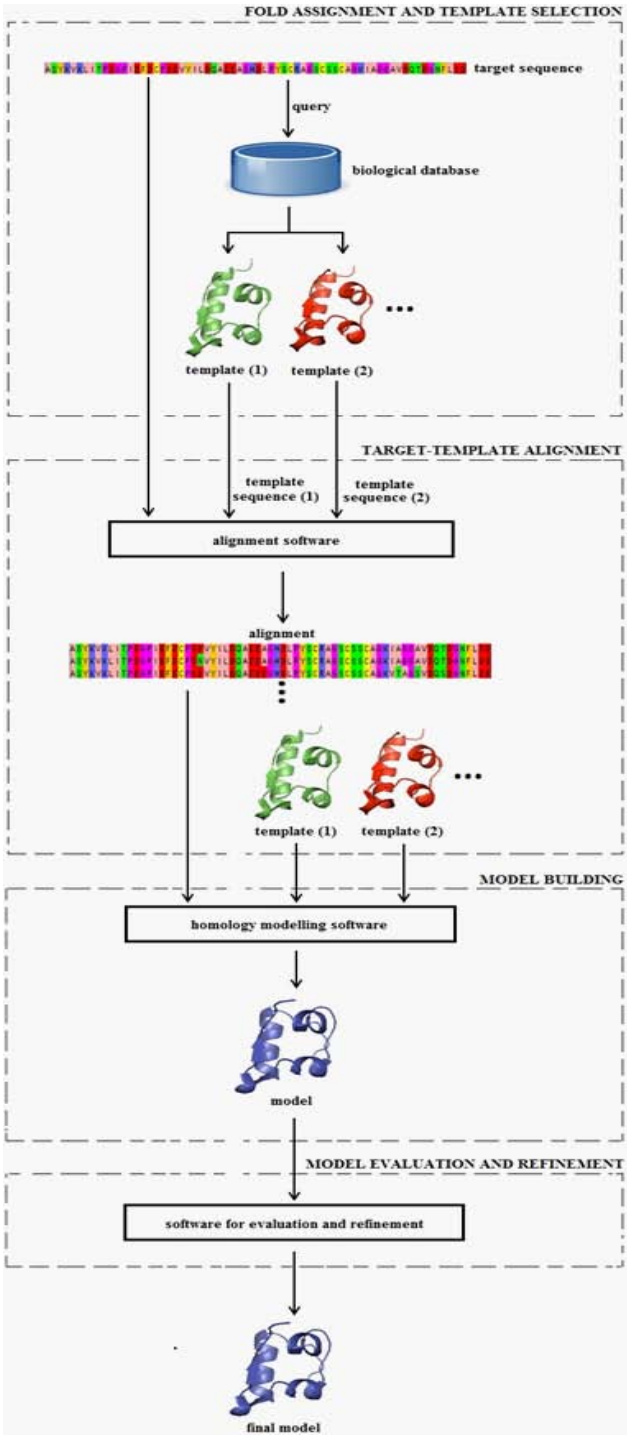


Figure 2: A flowchart illustrating generic steps in the construction of a protein structural model using molecular homology modeling.

Table 1: Useful programs/servers in homology modeling.

Programs / Servers	Main characteristic	Availability
Protein Data Bank - PDB	Archive about the 3D shapes of proteins	http://www.rcsb.org/pdb/home/home.do
Protein Data Bank in Europe - PDBe	Archive about the 3D shapes of proteins	http://www.ebi.ac.uk/pdbe/services
FASTA	Database-scanning software	http://www.ebi.ac.uk/Tools/sss/fasta/
BLAST	Database-scanning software	http://www.rcsb.org/pdb/home/home.do#Subcategory-search_sequences http://blast.ncbi.nlm.nih.gov/Blast.cgi
EMBOSS Needle	Pairwise sequence alignment	http://www.ebi.ac.uk/Tools/psa/emboss_needle/
ClustalW2	Multiple sequence alignment	http://www.ebi.ac.uk/Tools/msa/clustalw2/
MUSCLE	Multiple sequence alignment	http://www.ebi.ac.uk/Tools/msa/muscle/
T-Coffee	Multiple sequence alignment	http://www.ebi.ac.uk/Tools/msa/tcoffee/
3D-JIGSAW	Web server using rigid-body assembly	http://bmm.cancerresearchuk.org/~3djigsaw/
SWISS-MODEL	Web server using rigid-body assembly with loop modeling	http://swissmodel.expasy.org/
SEGMOD	Homology modeling by segment matching	http://csb.stanford.edu/levitt/segmod/
JACKAL (NEST)	Homology modeling by artificial evolution	http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:Jackal
MODELLER	Homology modeling by satisfaction of spatial restraints	https://salilab.org/modeller/
ModLoop	Loop modeling	http://modbase.compbio.ucsf.edu/modloop/
RAPPER	Loop modeling	http://mordred.bioc.cam.ac.uk/~rapper/
FALC-Loop	Loop modeling	http://falc-loop.seoklab.org/
SuperLooper	Loop modeling	http://bioinf-applied.charite.de/superlooper/
SCWRL	Side-chain modeling	http://dunbrack.fccc.edu/scwrl4/
SCCOMP	Side-chain modeling	http://www.sheba-cancer.org.il/cgi-bin/sccomp/sccomp1.cgi
SCAP	Side-chain modeling	http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:Scap
RAMP	Main-chain and side-chain modeling	http://www.ram.org/computing/ramp/ramp.html
PROCHECK	Model assessment	http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/index.html
WHATCHECK	Model assessment	http://swift.cmbi.ru.nl/gv/whatcheck/
ProSA-web	Model assessment	https://prosa.services.came.sbg.ac.at/prosa.php
VERIFY3D	Model assessment	http://services.mbi.ucla.edu/Verify_3D/
ERRAT	Model assessment	http://services.mbi.ucla.edu/ERRAT/
ANOLEA	Model assessment	http://melolab.org/anolea/

TECHNIQUE DESCRIPTION

The first step in homology modeling of a protein of unknown three-dimensional structure (target protein) is identification of proteins that can function as templates. A good candidate template should be a protein closely related to the target protein with a known three-dimensional structure [25,32]. This step is performed using database-scanning software with the target sequence as a query. Since it is necessary to find a related protein with a known three-dimensional structure, the search can be performed by querying the structural database Protein Data Bank

(PDB) [33] using an available tool from this database, such as the Basic Local Alignment Search Tool (BLAST) [34]. The selection of homologues with known structures from the PDB is a straightforward task if the query sequence has high sequence identity (>30%) to a structure. One major challenge in the template search step is the detection of remote homologues. In response, sequence profile methods, such as position-specific profile search methods [35] and hidden Markov models (HMMs) [36,37], have emerged as the primary approaches in distant homology detection and have improved the accuracy of sequence alignments, extending the boundaries of detectable sequence similarity [30].

After identifying at least one template protein, it is necessary obtain a sequence alignment between the target protein and template proteins. The main objective of the alignment is to identify a good correlation between the amino acid residues of each sequence. Correct alignment is the most important step, since errors introduced into the model by misalignment are difficult to remove in the later stages of refinement [30]. When only one template is used, a pairwise alignment method such as Dynamic Programming [38] is applied. Multiple alignment methods can be used when more than two sequences must be aligned [39–47].

Information about the templates and a target-templates sequence alignment is then used to generate a three-dimensional structural model of the target. Four main methods of model generation are employed for this purpose [30,52]: 1) modeling by assembly of rigid bodies [48,49]; 2) modeling by segment matching [50,51]; 3) modeling by artificial evolution [30,52]; and 4) modeling by satisfaction of spatial restraints [53-55]. The assembly of rigid bodies approach begins with the identification of conserved and variable regions of the templates by superposition [48, 49]. A framework for the superimposed templates can be calculated by averaging the atom coordinates of the structurally conserved regions. Using the closest conserved segment (in terms of root-mean-square deviation to the framework), conformations of the residues of the conserved regions are directly transferred to the model, and unconserved regions are then constructed using an ab initio approach or by searching a database for compatible structures [48,49]. In the segment matching approach, a model for the target sequence is built from a database of known structures. The target structure is broken into short fragments that are used to select segments with matching shape in the database [50,51]. Thus, sequence alignment is done over segments rather than over the entire protein. The segment coordinates are fitted into the building target structure until all the atomic coordinates of the target structure are obtained. Several independent models are built and then an average model is obtained [50,51]. Modeling by artificial evolution involves splitting the alignment between the query and template sequence into a list of operations, such as residue mutation, insertion, or deletion, representing a scenario in which the template ("parent structure") evolves into the target [30,52]. The model could be considered a process of evolving the template structure, based on the alignment, so that changes are carried out in a stepwise manner, with each step involving an energy cost [30]. Finally, modeling by satisfaction of spatial restraints uses a procedure that is similar in concept to that used in the determination of protein structures from

NMR-derived restraints. The model is generated after applying many restraints obtained from the alignment of the target with the template structures [53–55]. These restraints are relative to distances and dihedral angles and stereochemical restraints such as bond length and bond angle preferences. The model is built by an optimization method to satisfy spatial restraints [53].

Once a protein structure model is built, it is refined by focusing on tuning alignment and modeling loops and side chains [30,52]. Target sequences often have regions that are structurally different from the related regions in the templates. These segments correspond to loop regions that are characterized by insertions and deletions, which producing gaps in the sequence alignment. The gaps cannot be directly modeled requiring additional loop modeling procedure, which is a very difficult problem in homology modeling and is also a major source of error [56]. There are two main categories of methods for loop modeling [57]: 1) knowledge-based methods, which try to identify a segment of a protein with a known three-dimensional structure that fits the stem regions of the loop; and 2) ab initio (de novo) methods, which are usually based on potentials or scoring functions. Loop modeling is critical because these regions may contribute to specific interactions, such as active and binding sites. After the main chain atoms are built, the positions of side chains that are not modeled must be determined. Due to the side chain geometry is very important in evaluating protein-ligand interactions at active sites and protein–protein interactions at the contact interface, much effort has been dedicated to the development of many side chain packing programs. As a strategy, a side chain can be built by searching every possible conformation at every torsion angle of the side chain to select the one that has the lowest interaction energy with neighboring atoms. However, this approach is computationally prohibitive in most cases and the programs frequently utilize a combinatorial search based on discrete side chain conformations called rotamers. In this regard, all successful approaches to side-chain placement are at least partly knowledge based [57]. Although loop modeling and side chain modeling steps apply potential energy calculations to improve the model, this does not guarantee that the entire raw homology model is free of structural irregularities. To relieve steric collisions and strains without significantly altering the overall structure, an energy minimization and molecular dynamic simulation procedures can be applied on the entire model.

Evaluation of model quality is an important final step in homology modeling. When a model is built, it is necessary to check it for possible errors. Every homology model contains errors that mainly depend on the percentage of sequence identity between template and target and the number of errors in the template. These errors are frequently estimated either from the energy of the model or from the resemblance of a given characteristic of the model to real structures [30,58]. Assessment of homology models includes physicochemical parameters evaluations (such as φ–ψ angles, chirality, bond lengths and bond angles), assigning a score for each residue in its current environment [30,58]. The resolution of the models is another useful evaluation that determines the applications of prediction models. For example, studies involving drug design require high resolution models with a root-mean-square deviation (RMSD) of 1 to 1.5Å [59,60]. It is important to note that a relative low resolution model is still useful for certain purposes [60].

FINAL CONSIDERATIONS AND CONCLUSIONS

The complete understanding of the biological roles of proteins requires knowledge of their structures. However, there are many important proteins for which the sequence information is available, but their three-dimensional structures remain unknown. Experimental methods to determine protein structures are time consuming and limited in their approach. Protein structural prediction offers a theoretical alternative to experimental determination of structures, allowing an efficient way to obtain structural information when experimental techniques are not successful. The process of the molecular homology modeling of proteins is simple in principle, deriving models from close homologs with experimentally determined three-dimensional structure. It is performed by sequential steps involving template selection, sequence alignment, backbone generation, loop building, side chain modeling, model refinement, and model evaluation. Among these steps, sequence alignment is the most crucial step and loop modeling is the most difficult. Although this technique is inability to perform prediction of new folds, because its knowledge-based nature, molecular homology modeling is one of the most frequently used methodology, allowing generation of precise and accurate models.

References

1. Zhang Q, Veretnik S, Bourne, PE. Overview of Structural Bioinformatics. In: Chen YPP, editor. Bioinformatics Technologies. Berlin: Springer. 2005; 15-44.
2. Dorn M, E Silva MB, Buriol LS, Lamb LC. Three-dimensional protein structure prediction: Methods and computational strategies. Comput Biol Chem. 2014; 53PB: 251-276.
3. Rupp B. Biomolecular crystallography: principles, practice, and application to structural biology. Garland Science. 2009.
4. Nespolo M. Crystals, X-rays and Proteins. Comprehensive Protein Crystallography. By Dennis Sherwood and Jon Cooper. Oxford University Press, 2015 (paperback), ISBN 9780198726326, price GBP39.99. Acta Crystallogr D Struct Biol. 2016; 72: 181.
5. Keeler J. Understanding NMR Spectroscopy. 2th edn. New Jersey: Wiley. 2010.
6. Wüthrich K. NMR of Proteins and Nucleic Acids. New Jersey: Wiley-Interscience. 1986.
7. Palmer AG, Fairbrother WJ, Cavanagh J, Skelton NJ, Rance M. Protein NMR Spectroscopy: Principles and Practice. 2th edn. Cambridge: Academic Press. 2006.
8. Unwin PN, Henderson R. Molecular structure determination by electron microscopy of unstained crystalline specimens. J Mol Biol. 1975; 94: 425-440.
9. Callaway E. The revolution will not be crystallized: a new method sweeps through structural biology. Nature. 2015; 525: 172-174.
10. Bai XC, Yan C, Yang G, Lu P, Ma D. An atomic structure of human γ-secretase. Nature. 2015; 525: 212-217.
11. Bartesaghi A, Merk A, Banerjee S, Matthies D, Wu X. 2.2 Å resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor. Science. 2015; 348: 1147-1151.
12. Bai XC, McMullan G, Scheres SH. How cryo-EM is revolutionizing structural biology. Trends Biochem Sci. 2015; 40: 49-57.
13. Bujnicki JM. Protein-structure prediction by recombination of fragments. Chembiochem. 2006; 7: 19-27.
14. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol. 2005; 15: 285-289.
15. Osguthorpe DJ. Ab initio protein folding. Curr Opin Struct Biol. 2000; 10: 146-152.
16. Tramontano A. Protein structure prediction: concepts and applications. Weinheim: Wiley-VCH. 2006.
17. Floudas CA, Fung HK, McAllister SR, Moennigmann M, Rajgaria R. Advances in protein structure prediction and de novo protein design: a review. Chem. Eng. Sci. 2006; 61: 966-988.

18. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol. 2004; 383: 66-93.

19. Srinivasan R, Rose GD. LINUS: a hierarchic procedure to predict the fold of a protein. Proteins. 1995; 22: 81-99.

20. Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science. 1991; 253: 164-170.

21. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature. 1992; 358: 86-89.

22. Bryant SH, Altschul SF. Statistics of sequence-structure threading. Curr Opin Struct Biol. 1995; 5: 236-244.

23. Turcotte M, Muggleton SH, Sternberg MJE. Application of inductive logic programming to discover rules governing the three-dimensional topology of protein structure. In: Page D, editor. Inductive Logic Programming. Springer. 1998; 53-64.

24. Sánchez R, Sali A. Advances in comparative protein-structure modelling. Curr Opin Struct Biol. 1997; 7: 206-214.

25. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 2000; 29: 291-325.

26. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)--round x. Proteins. 2014; 82 Suppl 2: 1-6.

27. Moult J, Fidelis K, Kryshtafovych A, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)--round IX. Proteins. 2011; 79 Suppl 10: 1-5.

28. Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A. Critical assessment of methods of protein structure prediction - Round VIII. Proteins. 2009; 77 Suppl 9: 1-4.

29. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T. Critical assessment of methods of protein structure prediction-Round VII. Proteins. 2007; 69 Suppl 8: 3-9.

30. Xiang Z. Homology-based modeling of protein structure. In: Xu Y, Xu D, Liang J, editors. Computational Methods for Protein Structure Prediction and Modeling. Volume 1: Basic Characterization. Berlin: Springer. 2007; 319-357.

31. Koehl P, Levitt M. A brighter future for protein structure prediction. Nat Struct Biol. 1999; 6: 108-111.

32. Srinivasan N, Guruprasad K, Blundell TL. Comparative modelling of proteins. In: Sternberg M, editor. Protein structure prediction: a practical approach. Oxford University Press. 1997; 111-140.

33. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28: 235-242.

34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215: 403-410.

35. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25: 3389-3402.

36. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol. 1994; 235: 1501-1531.

37. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics. 1998; 14: 846-856.

38. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970; 48: 443-453.

39. Lipman DJ, Altschul SF, Kececioglu JD. A tool for multiple sequence alignment. Proc Natl Acad Sci U S A. 1989; 86: 4412-4415.

40. Hirosawa M, Totoki Y, Hoshida M, Ishikawa M. Comprehensive study on iterative algorithms of multiple sequence alignment. Comput Appl Biosci. 1995; 11: 13-18.

41. Grasso C, Lee C. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. Bioinformatics. 2004; 20: 1546-1556.

42. Kim J, Pramanik S, Chung MJ. Multiple sequence alignment using simulated annealing. Comput Appl Biosci. 1994; 10: 419-426.

43. Edgar RC1. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32: 1792-1797.

44. Brudno M, Chapman M, Göttgens B, Batzoglou S, Morgenstern B. Fast and sensitive multiple alignment of large genomic sequences. BMC Bioinformatics. 2003; 4: 66.

45. Notredame C. Recent evolutions of multiple sequence alignment algorithms. PLoS Comput Biol. 2007; 3: e123.

46. Thompson JD, Plewniak F, Poch O. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. 1999; 27: 2682-2690.

47. Wallace IM, Blackshields G, Higgins DG. Multiple sequence alignments. Curr Opin Struct Biol. 2005; 15: 261-266.

48. Greer J. Comparative model-building of the mammalian serine proteases. J Mol Biol. 1981; 153: 1027-1042.

49. Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM. Knowledge-based prediction of protein structures and the design of novel molecules. Nature. 1987; 326: 347-352.

50. Levitt M. Accurate modeling of protein conformation by automatic segment matching. J Mol Biol. 1992; 226: 507-533.

51. Jones TA, Thirup S. Using known substructures in protein model building and crystallography. EMBO J. 1986; 5: 819-822.

52. Xiang Z. Advances in homology protein structure modeling. Curr Protein Pept Sci. 2006; 7: 217-227.

53. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993; 234: 779-815.

54. Srinivasan S, March CJ, Sudarsanam S. An automated method for modeling proteins on known templates using distance geometry. Protein Sci. 1993; 2: 277-289.

55. Aszódi A, Taylor WR. Homology modelling by distance geometry. Fold Des. 1996; 1: 325-334.

56. Xiong J. Essential Bioinformatics. 1st edn. Cambridge: Cambridge University Press. 2006.

57. Krieger E, Nabuurs SB, Vriend G. Homology modeling. In: Gu J, Bourne PE, editors. Structural Bioinformatics. 2nd edn. Hoboken: Wiley-Blackwell. 2009; 509-523.

58. Laskowski RA. Structural quality Assurance. In: Gu J, Bourne PE, editors. Structural Bioinformatics. 2nd edn. Wiley-Blackwell. 2009; 273-303.

59. Baker D, Sali A. Protein structure prediction and structural genomics. Science. 2001; 294: 93-96.

60. Kihara D, Chen H, Yang YD. Quality assessment of protein structure models. Curr Protein Pept Sci. 2009; 10: 216-228.

Austin Publishing Group

Structure, Shape and Electrostatic Based Virtual Screening to Discover Small Molecule Therapeutics

Parkesh R[1]*, Bhutani I[1] and Madathil R[1]

[1]Institute of Microbial Technology, Chandigarh, India

***Corresponding author:** Parkesh R, Institute of Microbial Technology, Council of Scientific and Industrial Research, Chandigarh 160036, India, Tel: 91-172-6665488; Fax: 91-172-2690585; Email: rparkesh@imtech.res.in

Published Date: December 01, 2016

ABSTRACT

Computer-assisted decision-making contributes to selection and optimization of lead molecules as well as the discovery of small molecule drug development tools for chemical biology. Integration of chemoinformatics, bioinformatics and high throughput virtual screening is increasingly used to address challenging and unexplored drug targets amenable to small molecule perturbations. In this chapter, we will highlight the ability of high throughput virtual screening for the development of potential small molecule therapeutics targeting unexploited but important drug targets. Our particular focus will be on automated docking, three-dimensional shape-based screening, electrostatic complementarity as well as the application of these methods in finding lead molecules. This novel synergistic virtual screening technique has emerged as promising, cost-effective, time saving and even helpful in cases where structural information of the target is not available. Finally, we will provide examples of how high throughput virtual screening encompassing docking, three-dimensional shape-based matching and electrostatic complementarity have helped in the advancement of small molecule therapeutics against unexploited targets such as *Mtb* GlgB, RNA and NAADP-mediated signalling pathway.

Keywords: HTVS; Shape-based screening; Docking; Compound library; Conformer; Database; RNA; *Mtb*; GlgB

INTRODUCTION

The main aim of a drug discovery researcher is to identify a new chemical scaffold (hit molecules) which shows reasonably sufficient biological activity for a particular drug target and optimizing this scaffold through iterative cycles of structure-activity relationship (SAR), thus, yielding a lead molecule with improved potency and favourable absorption-distribution-metabolism-excretion-toxicity (ADMET) properties [1,2]. After successful completion of various phases of clinical trials, this compound can be launched as an FDA approved drug. The novel chemical scaffolds can be identified by random screening, phenotypic and target-based high throughput screening (HTS) and target-based direct design approach (if the high atomic resolution structure is available) [3]. Computational approaches such as high-throughput virtual screening (HTVS) are increasingly developed and refined to identify novel lead molecules in the pharmaceutical industry and academic groups [4]. HTVS is an application based upon various computational methods to "screen" large compound libraries, prioritizing ligands for experimental HTS or chemical synthesis [4]. HTVS typically ranks library molecules (using matrices such as docking score, shape and electrostatic Tanimoto) in the order of decreasing biological activity [5,6].

Can HTVS be complementary to HTS? Traditionally, random screening and now HTS is employed to identify novel small molecule ligands for biological targets [3,7]. With advances in robotics, automation, database management, software development, statistics, assay miniaturization, HTS is a method of choice, especially in pharmaceutical settings, to screen compound libraries comprising of millions of compounds for drug discovery [3]. Combinatorial methods for library design 'churning out' millions of compounds and advances in sophisticated biological assay development have increased the popularity of HTS. However, HTS methods have persistently failed to generate satisfactory results. The low hit rate, higher screening costs and intrinsic errors of HTS data, place severe constrains on the significance of HTS screening platform for developing small molecule ligands for biological targets [7-10]. Moreover, HTS approaches are less flourishing for unexploited targets as the compound libraries are biased towards protein targets with well-defined binding sites. There is a wide gap in translating the HTS identified hits into a preclinical candidate that can be optimized to become a drug molecule. Advances in structural biology, computing resources and development of high-end graphical cards for GPU computation has resulted in HTVS being integrated into the drug discovery and the design of small molecule chemical tools for understanding various biological processes (Figure 1). As HTVS involves pre-filtering of the compound libraries (Figure 2), it is attractive with the promise of higher hit rates, more drug like or lead like molecules [4,11].

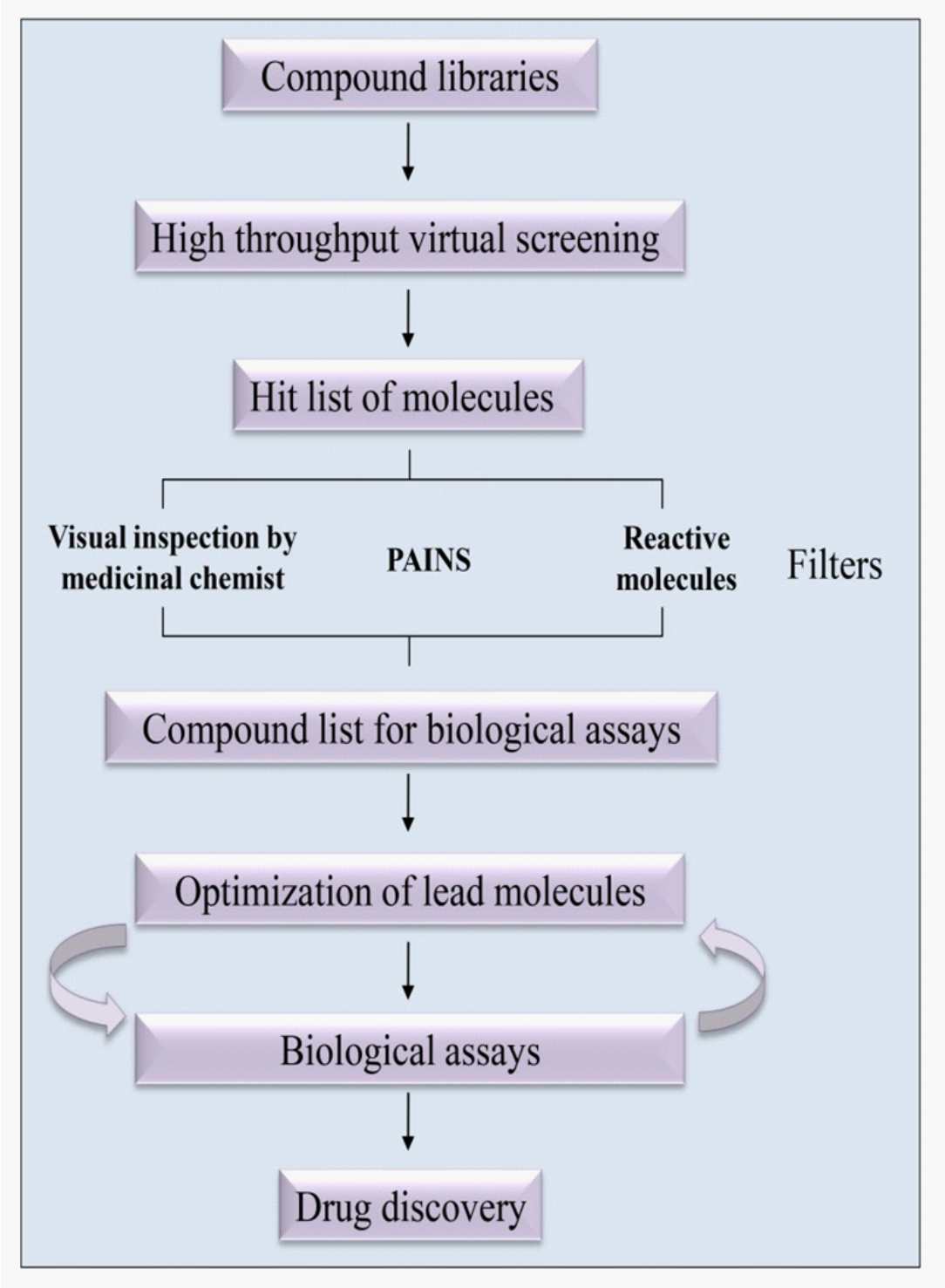


Figure 1: General workflow for HTVS integrated drug discovery process.

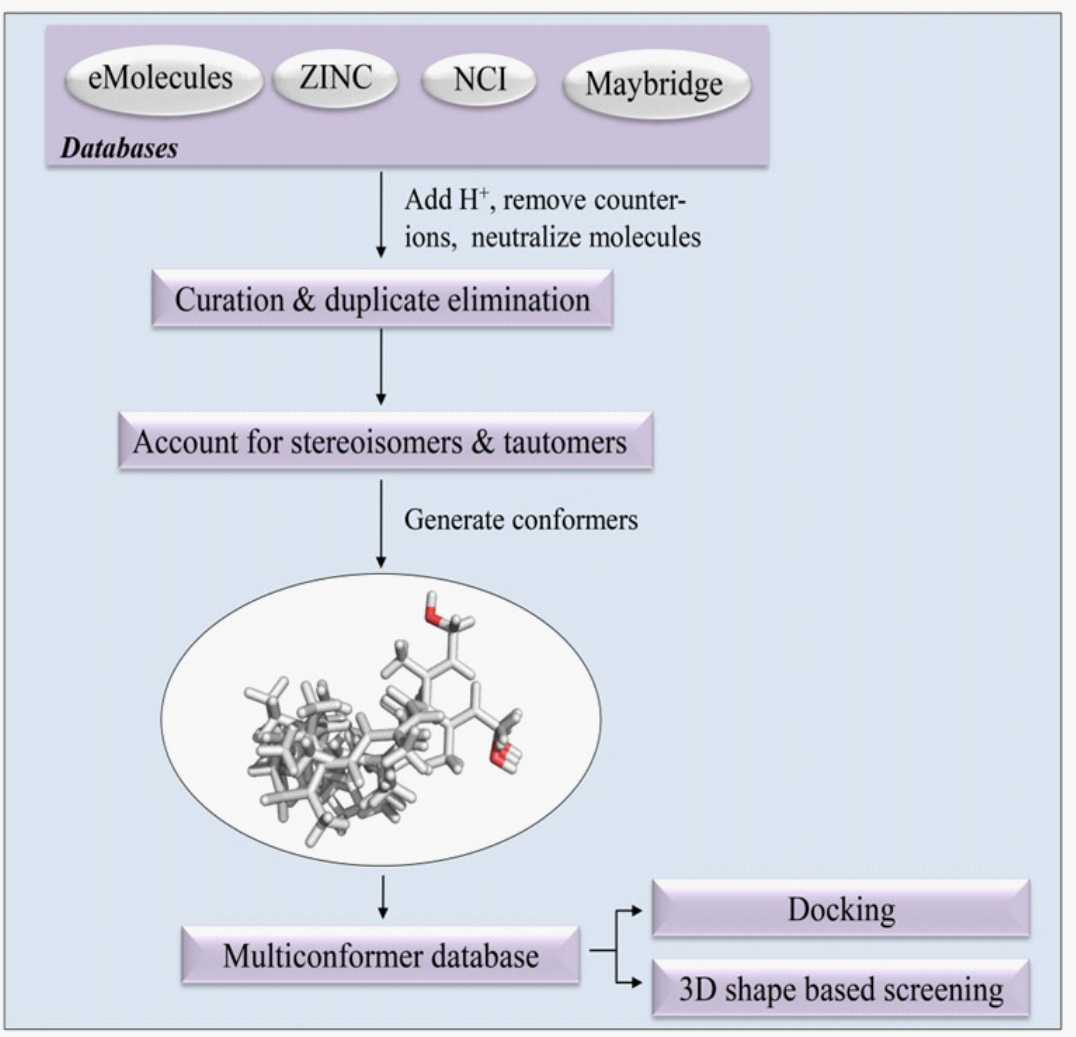


Figure 2: The schematics depicting the procedure for preparing a multi-conformer database.

Broadly HTVS approaches fall into two main classes1:1) structure –based, protein-centric and 2) shape-based, ligand-centric [12]. When the ligand-bound high atomic resolution crystal structure of the target protein is available, virtual screening using docking is often considered as the first choice strategy [13] (Figure 3). Many stand-alone docking programs, each with their strength, weaknesses and scoring functions are available (Table 1). Additionally, several new docking software initiatives that utilize the power of "distributive computing" are available for various drug discovery processes (Table 2). Docking is a useful method of virtual screening with its strength and limitations [13]. The reliability of the scoring functions used to rank the docked molecules is well-documented [12]. Furthermore, the uncertainty to correctly predict the binding mode of the diverse compounds has further limited the use of docking as a validated screening technique.

Table 1: A list of stand-alone docking programs with their web addresses.

Docking Program	Web Address	Algorithm	Description
AutoDock	http://autodock.scripps.edu/	Genetic algorithm, Lamarckian genetic algorithm, simulated annealing	Keeps ligand and protein side chains flexible, provides high quality predictions of ligand conformations, and good correlations between predicted inhibition constants and experimental ones.
DOCK	http://dock.compbio.ucsf.edu/	Shape fitting (sphere sets)	Rigid body docking. It uses geometric matching algorithm to superimpose ligand onto the negative image of the binding pocket.
GOLD	http://www.ccdc.cam.ac.uk/Solutions/GoldSuite/Pages/GOLD.aspx	Genetic algorithm	Ligand is flexible and protein is partial flexible. Provides virtual screening, lead optimization, and identify the correct binding mode of active molecules.
FlexX	http://www.biosolveit.de/flexx/	Incremental construction	Ligand is flexible and protein flexibility is achieved through ensemble of protein structure. Accurately predicts the protein-ligand complex geometry.
FRED	http://www.eyesopen.com/oedocking	Shape fitting (guassian)	Performs a systematic, exhaustive, non-stochastic examination of all possible poses within the protein active site and also filters for shape complementarity and pharmacophoric features.
Glide	http://www.schrodinger.com/Glide	Monte Carlo sampling	Both ligand and protein are flexible. Exhaustive search based docking program.
LigandFit	http://accelrys.com/	Monte Carlo sampling	Ligand conformations are docked into an active site based on shape and minimized using CHARMM.
ICM	http://www.molsoft.com/docking.html	Monte carlo minimization	Both ligand and protein are flexible. Based on pseudo-Brownian sampling and local minimization.
Surflex	http://www.certara.com/products/molmod/sybyl-x/sbd/	Surface based molecular similarity	Offers HTVS and accurate prediction for ligand binding mode and conformation.
FITTED	http://fitted.ca/	Genetic algorithm	It accounts for flexibility of the two molecules and location of water molecules to form potential covalent bonds with the protein side-chains
HYBRID	http://www.eyesopen.com/oedocking	Shape fitting (guassian)	It uses bound ligand information to improve virtual screening performance

Table 2: A few popular distributive computing softwares for docking.

Database	Web address	Description
Docking@home	http://boinc.berkeley.edu/wiki/Docking@Home	Simulate the docking of ligands to proteins
Surflex-Dock	http://www.certara.com/products/molmod/surflex/surflex-dock	Predicts binding pose, screen and prioritize molecules for lead discovery
Rosetta@home	http://boinc.bakerlab.org/	Predict and design protein structures, and protein-protein and protein-ligand interactions
Surflex-sim	http://www.certara.com/products/molmod/surflex/surflex-sim	Helps in finding chemical scaffolds while potentially reducing risks associated with toxicity.
Topomer search	http://www.certara.com/products/molmod/sybyl-x/simpharm/	Screen whole molecules, side chains, or scaffolds using conformationally independent topomer similarity

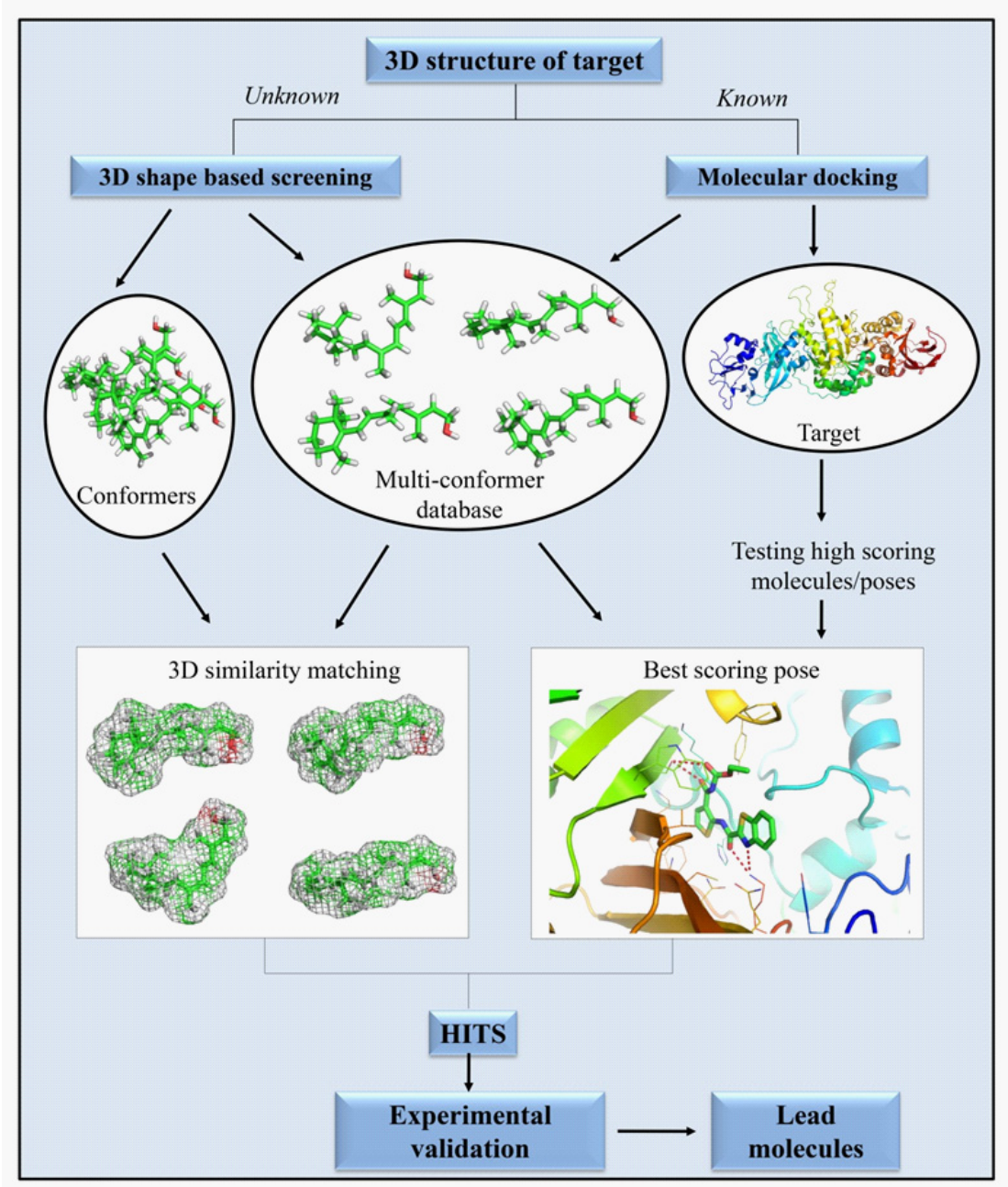


Figure 3: HTVS workflow illustrating two distinct approaches - target-centric (molecular docking, target crystal structure available) and ligand centric (3D shape-based screening, unknown target structure).

Recently, the focus has shifted to virtual screening methodology involving three-dimensional (3D) shape and electrostatic-based screening [14,15]. The shape-based screening is based on the

concept of similarity principle "similar shape molecules show a similar biological response. There are many shape-based algorithms available including Cat-Shape as implemented in CATALYST, the Phase-Shape module as implemented in the Schrodinger product suite, Ultrafast shape recognition (USR) as implemented in ElectroShape, Rapid Overlay of Chemical Structures (ROCS) as implemented in the Openeye software and XED field points as implemented in Blaze from Cresset technology [5,16-18]. Shape-based virtual screening methods are becoming increasingly popular, and the literature data show that a shape-based screening methodology is more reliable and performs better than docking [15]. Electrostatic interactions are long-range forces that favour the formation of non-covalent bonds for partners with oppositely charged groups [14]. Many drug like molecules contain charged groups, such as carboxylates or aliphatic amines, which interact with complementary partners in the binding site. Thus, electrostatic complementarity plays an important role in exploring the relationship between the polarity of the molecule and its relative promiscuity [14].

Herein we will discuss various strategies for the chemical database use, compound preparation for docking and 3D shape-based screening by illustrating three distinct cases of *Mycobacterium tuberculosis (Mtb)*, RNA and NAADP-mediated signalling pathway [19-22]. We envisaged that the example provided in the case studies discussed here would encourage researchers from diverse backgrounds to apply HTVS for in drug discovery or for the development of small molecule chemical toolkit for studying various biological processes.

Successful Applications of HTVS

Case: Where the X-ray crystal structure of the target molecule is known

The α-1,4-glucan branching enzyme (GlgB) is critical for the biosynthesis of α- glucan and an essential component of *Mtb* cell wall [23]. Inhibition of GlgB with potent chemical scaffolds would prevent the branching of α-glucan, resulting in linear forms, thus causing cell wall lysis. The crystal structures of human (4BZY) and *Mtb* (3K1D) exhibit clear differences in structural aspects [24]. As a result, the inhibitory targeting of *Mtb* GlgB would be highly beneficial. The study successfully unearthed potent GlgB inhibitors using ligand and structure based drug design methodologies [19]. To identify hits that are selective for *Mtb* GlgB, HTVS was performed on both human and *Mtb* GlgB. Automated docking was used in combination with 3Dshape-based screening to pull out diverse hits targeting *Mtb* GlgB [19]. From the generated hits of *Mtb* GlgB, 17 compounds that are specific for *Mtb* GlgB are selected for further evaluation. These compounds were further analysed for favourable *in silico* pharmacokinetic and enzyme inhibition assay. The 3D conformers of the selected ligands were used as queries to find 29 additional hits from the in-house database. The binding conformation and interaction of these two compounds (after biological assays) in human and *Mtb* GlgB were compared by using AutoDock [25] (Figure 4).

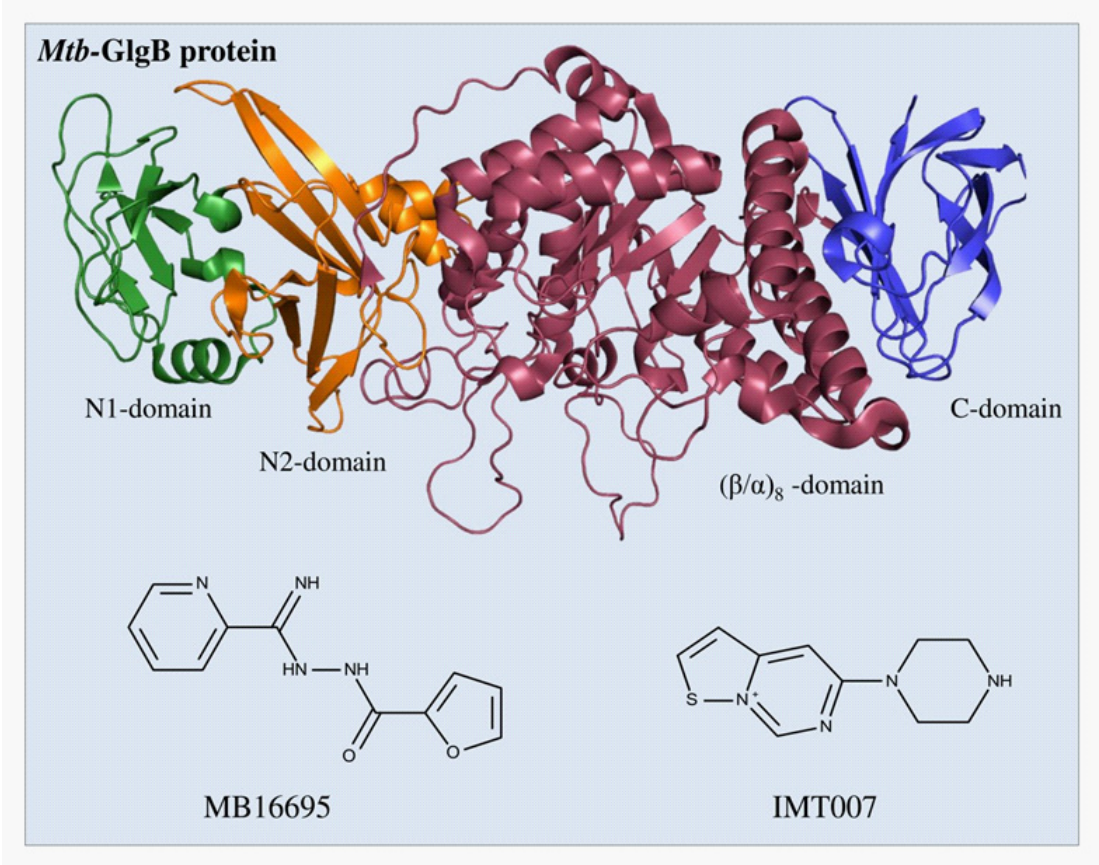


Figure 4: a) *Mtb* GlgB crystal structure. b) and c) are chemical structures of inhibitors identified using HTVS approach.

Outcome: Two potent inhibitors of *Mtb* GlgB were discovered and experimentally validated using enzymatic and *in vitro* biological assay [19].

The detailed steps followed in the HTVS methodology are summarized below:

1. The *Mtb* GlgB crystal structure was downloaded from PDB (www.pdb.org) (PDBID: 3KID) [24].

2. The homology model of human GlgB was generated using SWISS-MODEL interface, a web integrated service (http://swissmodel.expasy.org/) [26]. The target sequence was used as input for automated modelling of the protein.

3. Downloaded diverse ligand databases: Maybridge from www.maybridge.com and ZINC from http://zinc.docking.org/.2 Stepwise procedure to perform docking using Glide:

a. To run any Schrodinger program on Linux- The Schrodinger environment variable is set to the installation directory by the command:

csh/tcsh: setenv SCHRODINGER *installation-directory*. The Maestro interface can be started by using the command:

SCHRODINGER/maestro&.

For windows: start → All programs → Schrodinger → Maestro.

b. In the graphical interface Maestro, protein preparation wizard was used to add missing hydrogen's and water, assign bond orders and minimize models using OPLS-AA in Schrodinger package. The energy minimization comprises of rigid-body translations and rotations to optimize the model.

c. "Receptor grid generation" application was used to prepare a grid around the active sites of the GlgB protein.

d. The database molecules were prepared using "LigPrep" to add missing hydrogen's, remove counter ions, neutralize charged groups, generate ionization states, tautomers and optimize ligand geometries.

e. In ligand-based VS, the prepared ligands were docked into the generated receptor grid.

f. In the end, the results are shown in the project table as pose viewer file. The top hits were selected by considering selective binding towards *Mtb* GlgB.

4. The selected ligands were assessed for their ADMET properties using QikProp. To run QikProp:

a. Choose Applications QikProp

b. Choose Project Table from the Use structures from option menu.

c. From the incorporate option menu, choose Replace existing entries.

d. Click save and run.

5. These hits generated from HTVS were used to find additional hits from our *in-house* database using 3Dshape-based screening. The structures of query molecules were drawn in ChemDraw and energy minimization was performed using MMFF94.

6. Generation of 100 three-dimensional conformers of each query ligand molecule and each molecule in the in-house database using Omega [27]. Command line option:

Omega2 –in query molecule name.sdf –out conformers.oeb.gz –maxconfs 100

(Option maxconfs is used to set the maximum number of conformers to be generated)

7. 3Dshape comparison was performed using ROCS [27] based on the shape Tanimoto coefficient [6] and color scores (chemistry alignment overlap) [28]. The commands used for screening using ROCS are:

a. Firstly the output is changed to a compact form:

rocs –dbase <database name.oeb.gz> -query <query name.sdf> -prefix out_oeb –besthits 500 –oformat oeb

b. Command for 3Dshape matching:

rocs –dbase <database name.oeb.gz> -query <query name.sdf> -prefix tversky –besthits 500 –oformat oeb.gz rankby tanimoto

(*Tversky* measure ranks molecules, biased towards the query molecule; Tanimoto quantifies the 3Dshape-based match score)

c. Command for 3D chemistry alignment:

rocs –dbase <database name.oeb.gz> -query <query name.sdf> -prefix colour -besthits 500 –chemff ImpilcitMillsDean

d. The ranking is done on the basis of the total score as the sum of shape Tanimoto coefficient [6] and color score [28]. The Tanimoto coefficient ranges from 0 (no similarity) to 1 (complete shape similarity) [6]. The color score "1" and "0" represents no overlap and complete chemistry overlap, respectively [28].

8. Molecules with the highest score were output in rank order as potential hits.

9. To determine the binding conformation and interaction of these molecules, AutoDock [25] was used to dock the compounds into *Mtb* GlgB and subsequently check their binding to human GlgB.

10. Biological assay validation.

Case: To use known ligands as templates in virtual screening for finding novel chemical scaffolds

RNA is a versatile supramolecule that plays an important role in various cellular processes. RNA has been shown important in the progression of many infectious, metabolic and genetic disorders directly or indirectly [29-32]. There has been a tremendous advance in the structural biology of RNA and thus, targeting RNA by small molecules is an attractive area of research. In this case study, we have provided examples of RNA molecules that are directly involved as causative agents for disease progression. For example, RNA with expanded repeats of CUG, CAG, etc. are responsible for many debilitating diseases (such as myotonic dystrophy, Huntington's and Spinocerebellar ataxia type 3) [33-38]. The structure of the RNA molecule with fewer repeats of CAG, CUG, etc. has been solved, but the pathological RNA molecule with hundreds of repeats remains unknown [39,40]. Additionally RNA is a "floppy" molecule, which makes docking problematic, despite of thefew advances made [41-43]. In this context, we will highlight an alternative approach, where the known ligands Hoechst 33258, pentamidine and DAPI have been exploited to mine various high quality databases by 3Dshape-based and electrostatic screening [20,21] (Figure 5).

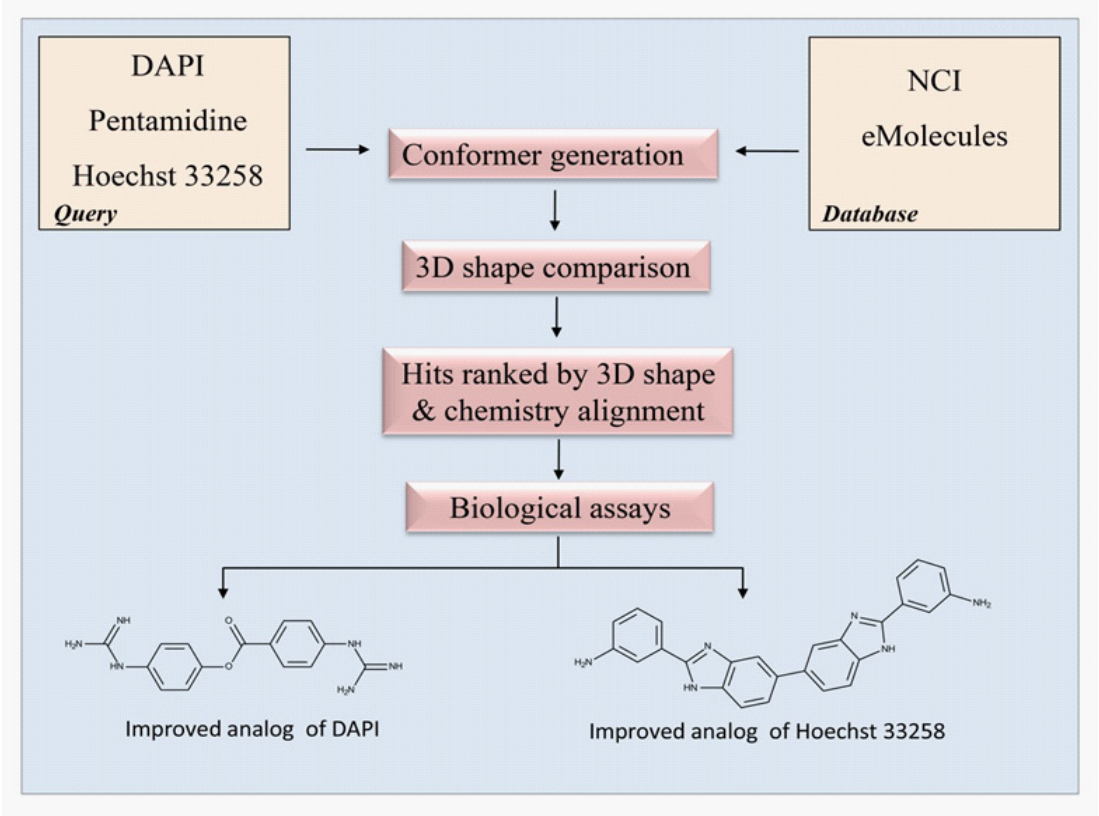


Figure 5: Schematics depicting ligand centric, shape-based screening strategy applied for RNA repeats of CUG and CAG.

Outcome: The study reported a series of inhibitors that are approximately twenty folds more potent than the query molecules pentamidine, Hoechst 33258 and DAPI [20,21].

To perform 3Dshape-based screening, the essential steps as reported in the case studies [20,21] are listed below:

Download diverse ligand databases such as NCI from cactus.nci.nih.gov and eMolecules from www.emolecules.com. [3]

Generation of 100 3D conformers of each query ligand molecule as well as each molecule in the NCI and the eMolecules database using Omega [4,27].

Omega2 –in query molecule name.sdf –out conformers.oeb.gz –maxconfs 100

(Option maxconfs is used to set the maximum number of conformers to be generated)

1. The structures of DAPI, pentamidine and Hoechst 33258 were drawn in ChemDraw and energy minimized with MMFF94 force field and the query molecules were entered as neutral molecules in the screen.

2. 3Dshape comparison was performed using ROCS [27] based on the shape Tanimoto coefficient [6] and color scores (chemistry alignment overlap) [28]. Command line to perform shape-based matching and chemistry alignment.

a. Firstly the output is changed to a compact form:

rocs –dbase <database name.oeb.gz> -query <query name.sdf> -prefix out_oeb –besthits 500 –oformat oeb

b. Command for 3Dshape matching:

rocs –dbase <database name.oeb.gz> -query <query name.sdf> -prefix tversky –besthits 500 –oformat oeb.gz rankby tanimoto

(*Tversky* measure ranks molecules, biased towards the query molecule; Tanimoto quantifies the 3Dshape-based match score)

c. Command for 3D chemistry alignment:

rocs –dbase <database name.oeb.gz> -query <query name.sdf> -prefix colour -besthits 500 –chemff ImpilcitMillsDean

d. The ranking is based on the total score (sum of shape Tanimoto coefficient and color score) [28]. The Tanimoto coefficient or color score of "1", "0.5" and "0" represents complete overlap, 50% overlap and no overlap respectively.

3. The top 500 molecules with the highest score were output in rank order as potential hits [6,27,44].

4. A visually inspected chemically diverse subset of molecules was selected from the ROCS [5] selected molecules.

Case: Shape and electrostatic based screening to identify analogs of a biologically relevant molecule NAADP

The discovery of two novel calcium signalling second messenger, cyclic ADP-ribose (cADPR) and nicotinic acid adenine nucleotide phosphate (NAADP) has created an unprecedented and unexpected understanding of Ca2+ signalling in living organisms [45-48] . Medicinal chemists have solved many technological challenges to allow the synthesis of various analogs of cADPR and NAADP. These analogs have provided invaluable pharmacological insights into the cell signalling events in living cells. We will highlight how HTVS particularly 3Dshape-based screening, and electrostatic complementarity have lead to the discovery of NED-19, a potent nanomolar antagonist of NAADP [22] (Figure 6).

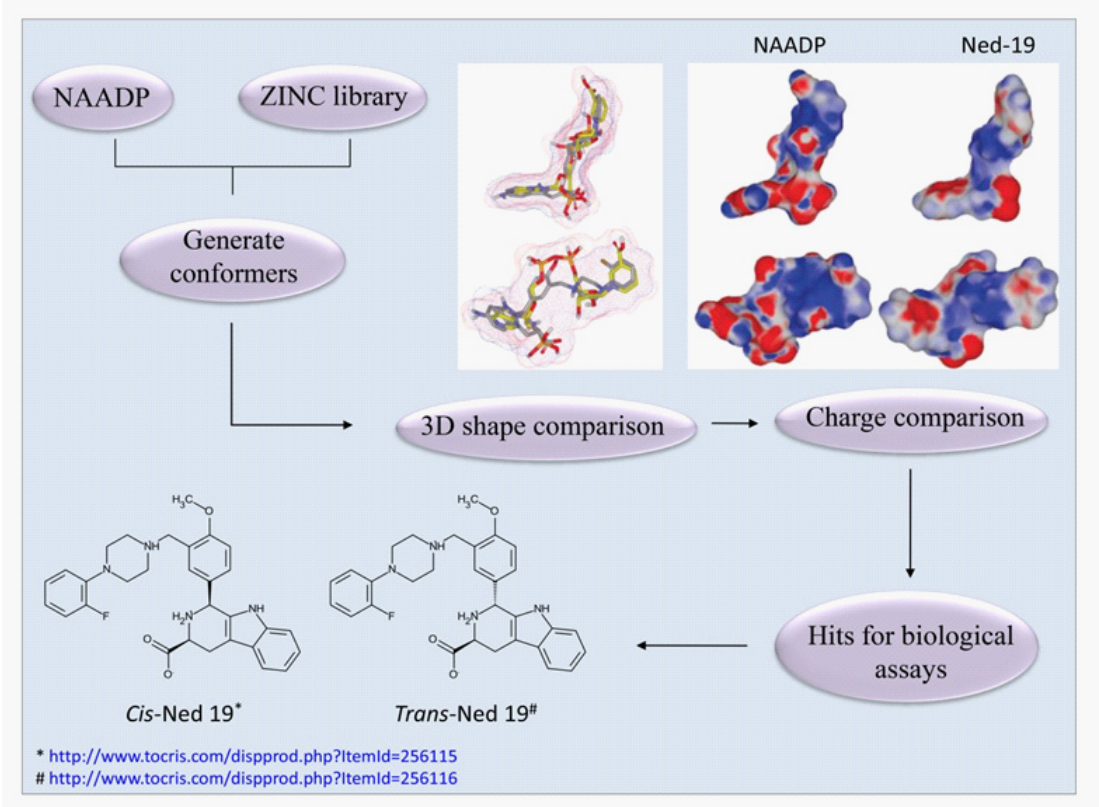


Figure 6: Depiction of ligand based virtual screening process used to develop NAADP antagonist. (The figure is partly adapted from reference [22]).

Outcome: NED-19 is one of the rare success story, where the integration of biology, virtual screening and medicinal chemistry have lead to the development and commercialization of novel chemical tools for calcium signalling [22].

The steps followed for HTVS are as follows:

1. Generation of 40 three-dimensional D conformers of the NAADP molecule (query) and 100 conformations of each ligand molecule in the ZINC database (zinc.docking.org)5 were generated using Omega [27]. Command line options for omega;

Omega2 –in query molecule name.sdf –out conformers.oeb.gz –maxconfs 100

(Option maxconfs is used to set the maximum number of conformers to be generated)

2. The structures of NAADP were made in ChemDraw and energy minimized with MMFF94 force field. For NAADP, all oxygen-phosphate bonds were set to single, approximating phosphate resonance shape.

3. The 3Dshape comparison was performed using ROCS [27] based on the shape Tanimoto coefficient [6].

a. Firstly the output is changed to a compact form:

rocs –dbase <database name.oeb.gz> -query <query name.sdf> -prefix out_oeb –besthits 500 –oformat oeb

b. Command for 3Dshape matching:

rocs –dbase <database name.oeb.gz> -query <query name.sdf> -prefix tversky –besthits 500 –oformat oeb.gz rankby tanimoto

(*Tversky* measure ranks molecules biased towards the query molecule; Tanimoto quantifies the 3Dshape-based match score)

c. Command for 3D chemistry alignment:

rocs –dbase <database name.oeb.gz> -query <query name.sdf> -prefix colour -besthits 500 –chemff ImpilcitMillsDean

4. Ranking is done on the basis of the total score as the sum of shape Tanimoto coefficient [6] and color score [28] ("1" shows a complete overlap (same shape) and "0.5" means 50% overlap)

5. The top 500 molecules with the highest score were output in rank order as potential hits [6,27,44].

6. These 500 molecules from ROCS were used for electrostatic comparison to NAADP molecule using EON software. Two EON runs were performed. In the first run, lowest energy conformer of NAADP was electrostatically compared with 500 top molecules from ROCS hit molecules. In the second run, all 40 conformations were electrostatically compared with top 500 hit molecules obtained from ROCS. For NAADP, all phosphates were protonated and modelled as neutral molecules, and oxygen-phosphate single bonds were converted to double bonds. Command line options for using EON:

eon –dbase rocs_hits.oeb –query rocsquery.sdf –besthits 100.

(ROCS query is used as the EON query and ROCS hits file is used as the database file for EON; the hits are saved in EON_hits.oeb file). The ranking was done based on electrostatic Tanimoto score, which ranges from 1 (identical) to negative values resulting from the overlap of positive and negative charges.

7. The top 10 hits after the initial EON screen and the top 15 Ned hits (5 new compounds) in the second EON screen were selected for *in vitro* testing. This study leads to the identification of NED-19 as a potent nanomolar NAADP antagonist.

8. Biological assay validation.

CONCLUSION AND OUTLOOK

In summary, the present chapter describes the usefulness of HTVS in identifying potential hits for unique and unexploited targets. HTVS, to date, has provided novel chemical scaffolds for drug discovery as well as small molecule pharmacological tools for the fundamental understanding of biological processes. The effectiveness of HTVS can be further enhanced by introducing better docking scoring function and similarity predictions. Furthermore, development of better force-field parameters for small molecules would be helpful. Recent success in fragment-based HTVS is promising. There are still many years of fruitful research waiting before we can fully appreciate the applications of HTVS in drug discovery. Recent high profile investments of pharmaceutical giant Sanofi and philanthropist Bill Gates will provide necessary boost in computational drug discovery informatics to develop cutting edge virtual screening technologies for drug discovery efforts [49].

NOTES

1. In the ligand-centric approach, some research groups include pharmacophore matching and QSAR matching but we consider these as another variant of either shape or electrostatic based.

2. Currently, the Maybridge database consists of 53000 diverse molecules and can be downloaded from www.maybridge.com. NCI version 2.2 has >250000 compounds at present and is accessible from cactus.nci.nih.gov and USA Food and Drugs Administration (FDA) approved drugs is available at http://www.fda.gov/Drugs/).

3. NCI version 2.2 has >250000 compounds at present and is accessible from cactus.nci.nih.gov whereas eMolecules database consist of 5.9 million compounds and is accessible from www.emolecules.com

4. In rigid molecules, please note that conformational sampling could be achieved before the cut-off 100 is exhausted.

5. The Zinc database is available at zinc.docking.org and presently comprises of 35 million molecules.

ACKNOWLEDGEMENT

The authors acknowledge the Institute of Microbial Technology (grant Infra 62 to R.P.) and DST (SB/SO/BB/023/2014 to R.P.) and DBT (BT/Bio-CARe/05/9923/2013-14 to R.M.) for financial support. We thank Dr. Girish Sahni (Director, Institute of Microbial Technology) for providing constant support and research facilities. IB is grateful to IMTECH for a Research Intern fellowship. We would like to thank Mr. Saurabh for help with the figures.

References

1. Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. Br J Pharmacol. 2011; 162: 1239-1249.
2. Wang J. Comprehensive assessment of ADMET risks in drug discovery. Curr Pharm Des. 2009; 15: 2195-2219.
3. Liu B, Li S, Hu J. Technological advances in high-throughput screening. Am J Pharmacogenomics. 2004; 4: 263-276.
4. Mestres J. Virtual screening: a real screening complement to high-throughput screening. Biochem Soc Trans. 2002; 30: 797-799.
5. Hawkins PC, Skillman AG, Nicholls A. Comparison of shape-matching and docking as virtual screening tools. J Med Chem. 2007; 50: 74-82.
6. Haigh JA, Pickup BT, Grant JA, Nicholls A. Small molecule shape-fingerprints. J Chem Inf Model. 2005; 45: 673-684.
7. Bolten BM, DeGregorio T. From the analyst's couch. Trends in development cycles. Nat Rev Drug Discov. 2002; 1: 335-336.
8. Fishman MC, Porter JA. Pharmaceuticals: a new grammar for drug discovery. Nature. 2005; 437: 491-493.
9. Ramesha CS1. How many leads from HTS? - Comment. Drug Discov Today. 2000; 5: 43-44.
10. Lahana R1. How many leads from HTS? Drug Discov Today. 1999; 4: 447-448.
11. Shoichet BK1. Virtual screening of chemical libraries. Nature. 2004; 432: 862-865.
12. Cavasotto CN, Orry AJ. Ligand docking and structure-based virtual screening in drug discovery. Curr Top Med Chem. 2007; 7: 1006-1014.
13. Morris GM, Lim-Wilby M. Molecular docking. Methods Mol Biol. 2008; 443: 365-382.
14. Radhakrishnan ML, Tidor B. Specificity in molecular design: a physical framework for probing the determinants of binding specificity and promiscuity in a biological environment. J Phys Chem B. 2007; 111: 13419-13435.
15. Nicholls A, McGaughey GB, Sheridan RP, Good AC, Warren G. Molecular shape and medicinal chemistry: a perspective. J Med Chem. 2010; 53: 3862-3886.
16. Ballester PJ, Richards WG. Ultrafast shape recognition to search compound databases for similar molecular shapes. J Comput Chem. 2007; 28: 1711-1723.
17. Ebalunode JO, Zheng W. Molecular shape technologies in drug discovery: methods and applications. Curr Top Med Chem. 2010; 10: 669-679.
18. Sastry GM, Dixon SL, Sherman W. Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring. J Chem Inf Model. 2011; 51: 2455-2466.
19. Dkhar HK, Gopalsamy A, Loharch S, Kaur A, Bhutani I. Discovery of Mycobacterium tuberculosis Î±-,4-glucan branching enzyme (GlgB) inhibitors by structure- and ligand-based virtual screening. J Biol Chem. 2015; 290: 76-89.
20. Kumar A, Parkesh R, Sznajder LJ, Childs-Disney JL, Sobczak K. Chemical correction of pre-mRNA splicing defects associated with sequestration of muscleblind-like 1 protein by expanded r(CAG)-containing transcripts. ACS Chem Biol. 2012; 7: 496-505.
21. Parkesh R, Childs-Disney JL, Nakamori M, Kumar A, Wang E, et al. Design of a Bioactive Small Molecule That Targets the Myotonic Dystrophy Type 1 RNA via an RNA Motif–Ligand Database and Chemical Similarity Searching. J Am Chem Soc. 2012; 134: 4731-4742.
22. Naylor E, Arredouani A, Vasudevan SR, Lewis AM, Parkesh R. Identification of a chemical probe for NAADP by virtual screening. Nat Chem Biol. 2009; 5: 220-226.
23. Agrawal P, Gupta P, Swaminathan K, Parkesh R. α-Glucan pathway as a novel Mtb drug target: structural insights and cues for polypharmcological targeting of GlgB and GlgE. Curr Med Chem. 2014; 21: 4074-4084.
24. Pal K, Kumar S, Sharma S, Garg SK, Alam MS, et al. Crystal Structure of Full-length Mycobacterium tuberculosis H37Rv Glycogen Branching Enzyme: Insights of n-terminal ß-sandwich in substrate specificity and enzymatic activity. J Biol Chem. 2010; 285: 20897-20903.
25. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009; 30: 2785-2791.
26. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. 2003; 31: 3381-3385.
27. Boström J, Greenwood JR, Gottfries J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. J Mol Graph Model. 2003; 21: 449-462.

28. Mills JE, Dean PM. Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. J Comput Aided Mol Des. 1996; 10: 607-622.

29. Calin GA, Croce CM. MicroRNAs and chromosomal abnormalities in cancer cells. Oncogene. 2006; 25: 6202-6210.

30. Caskey CT, Pizzuti A, Fu YH, Fenwick RG Jr, Nelson DL. Triplet repeat mutations in human disease. Science. 1992; 256: 784-789.

31. Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. Genes Dev. 2003; 17: 419-437.

32. Kalnina Z, Zayakin P, Silina K, Linē A. Alterations of pre-mRNA splicing in cancer. Genes Chromosomes Cancer. 2005; 42: 342-357.

33. Kanadia RN, Johnstone KA, Mankodi A, Lungu C, Thornton CA. A muscleblind knockout model for myotonic dystrophy. Science. 2003; 302: 1978-1980.

34. Kanadia RN, Shin J, Yuan Y, Beattie SG, Wheeler TM. Reversal of RNA missplicing and myotonia after muscleblind overexpression in a mouse poly(CUG) model for myotonic dystrophy. Proc Natl Acad Sci U S A. 2006; 103: 11748-11753.

35. Philips AV, Timchenko LT, Cooper TA. Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. Science. 1998; 280: 737-741.

36. Mankodi A, Takahashi MP, Jiang H, Beck CL, Bowers WJ. Expanded CUG repeats trigger aberrant splicing of ClC-1 chloride channel pre-mRNA and hyperexcitability of skeletal muscle in myotonic dystrophy. Mol Cell. 2002; 10: 35-44.

37. Paul S, Dansithong W, Kim D, Rossi J, Webster NJ. Interaction of muscleblind, CUG-BP1 and hnRNP H proteins in DM1-associated aberrant IR splicing. EMBO J. 2006; 25: 4271-4283.

38. Mankodi A, Logigian E, Callahan L, McClain C, White R. Myotonic dystrophy in transgenic mice expressing an expanded CUG repeat. Science. 2000; 289: 1769-1773.

39. Kiliszek A, Kierzek R, Krzyzosiak WJ, Rypniewski W. Structural insights into CUG repeats containing the 'stretched U-U wobble': implications for myotonic dystrophy. Nucleic Acids Res. 2009; 37: 4149-4156.

40. Kiliszek A, Kierzek R, Krzyzosiak WJ, Rypniewski W. Atomic resolution structure of CAG RNA repeats: structural insights and implications for the trinucleotide repeat expansion diseases. Nucleic Acids Res. 2010; 38: 8370-8376.

41. Azam S, Abbasi S, Batool M. Structure modeling and docking study of HCV NS5B-3a RNA polymerase for the identification of potent inhibitors. Med Chem Res. 2014; 23: 618-627.

42. Barbault F, Ren B, Rebehmed J, Teixeira C, Luo Y. Flexible computational docking studies of new aminoglycosides targeting RNA 16S bacterial ribosome site. Eur J Med Chem. 2008; 43: 1648-1656.

43. Scotti L, Oliveira Lima Ed, da Silva MS, Ishiki M, Oliveira Lima lD, et al. Docking and PLS studies on a set of thiophenes RNA polymerase inhibitors against Staphylococcus aureus. Curr Top Med Chem. 2014; 14: 64-80.

44. Grant JA, Gallardo MA, Pickup BT. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. J Comput Chem. 1996; 17: 1653-1666.

45. Galione A, Morgan AJ, Arredouani A, Davis LC, Rietdorf K. NAADP as an intracellular messenger regulating lysosomal calcium-release channels. Biochem Soc Trans. 2010; 38: 1424-1431.

46. Guse AH, Lee HC. NAADP: a universal Ca2+ trigger. Sci Signal. 2008; 1: re10.

47. Nebel M, Schwoerer AP, Warszta D, Siebrands CC, Limbrock AC, et al. Nicotinic acid adenine dinucleotide phosphate (NAADP)-mediated calcium signaling and arrhythmias in the heart evoked by ß-adrenergic stimulation. J Biol Chem. 2013; 288: 16017-16030.

48. Fliegert R, Gasser A, Guse AH. Regulation of calcium signalling by adenine-based second messengers. Biochem Soc Trans. 2007; 35: 109-114.

49. Stuckler D, Basu S, McKee M. Global health philanthropy and institutional relationships: how should conflicts of interest be addressed? PLoS Med. 2011; 8: e1001020.

Austin Publishing Group

Introduction to the Molecular Dynamics of Biomolecules

Morton-Blake DA[1]*

[1]School of Chemistry, Trinity College, Ireland

***Corresponding author:** Morton-Blake DA, School of Chemistry, Trinity College, Dublin 2, Ireland, Tel: 353-18961943; Fax: 353-16712826; Email: tblake@tcd.ie

Published Date: December 01, 2016

ABSTRACT

An introductory course is presented for the aspiring user of Molecular Dynamics **(MD)**. No specialist terms are used in the discussion, which employs the language of basic chemistry and physics. The dynamics of the particles are introduced through their mutually interactive forces and the laws of motion. The forces on the atoms are explained as also are the means of specifying them. The use of quantum chemical methods to calculate partial atomic charges and molecular geometries is outlined. Some concepts relevant to the appreciation of MD results - radial distribution, mean force and velocity autocorrelation - are introduced and comments are made on the probable future of the subject.

Keywords: Molecular dynamics; Simulation; Relaxation; Migration; Protein folding; Nucleic acids.

INTRODUCTION

Human beings have a passion for modeling. Both fictional and non-fictional literature are projections, or models, of the real world on the pages of a book, and the resulting editing of the original data refines the events or message that the author wishes to convey. In the same way, toy soldiers, space fleets and dolls houses are projected into the world of the young - images which active imaginations endow with properties and powers. Receptive minds then observe and manipulate them in a way denied to the objects from which they were derived.

As researchers we wish to investigate the behavior of atoms and molecules as they re-conform, transverse a condensed medium, progress in channels penetrating a cell membrane, dock on a substrate or intercalate into DNA [1]. But their small sizes make the direct observation of their behavior as hard to attain as it would for a young modeller to direct objects in the real world. So we reprise our modeling activity.

Molecular Dynamics **(MD)** emulates the motions of atoms and molecules in real time. Since it is conducted for atoms or molecules or small groups of them, it can provide details of events which elude experimental measurement and elucidate various types of molecular motions such as transport (Figure 1) or the motion and folding of protein chains [2]. It is widely used to model ions or small molecular species as they migrate through a medium or to elucidate the passage of Na^+ or Cl^- ions through a cell membrane, in a natural [3] or synthetic [4] ion transporter (Figure 1).

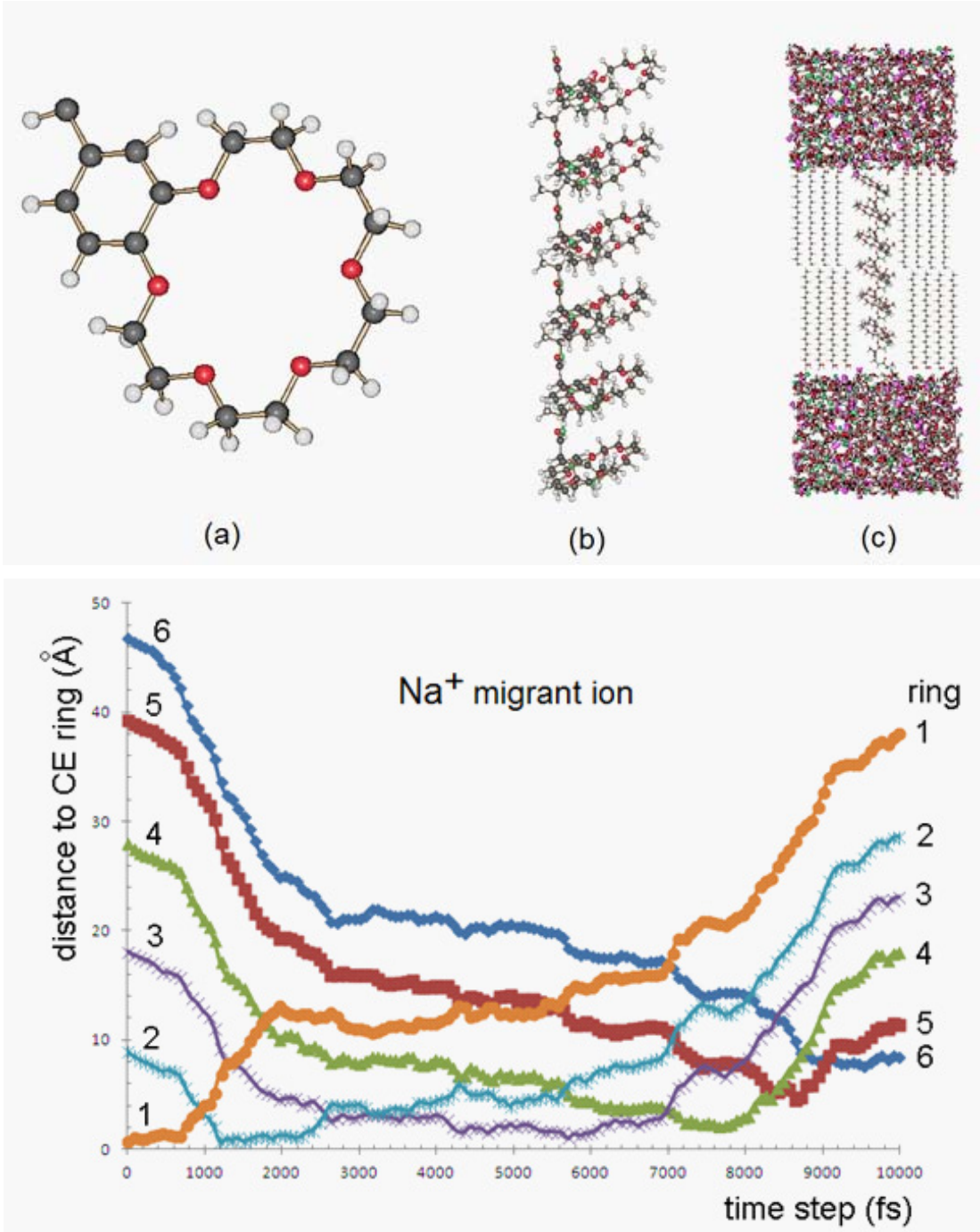


Figure 1: Migration of Na^+ in a synthetic ion channel of six Crown Ether **(CE)** rings mounted on a peptide chain shown symbolically in (a) to (c). In (c) some of the peptide alkyl chains have been removed to reveal the crown ether channels. In the lower diagram the traces show the distance of the ion migrant from each CE ring.

THEORY AND DATA INPUT

General

An assembly of N atoms or molecules is represented by charged point masses and its components referred to as *particles*. They will occupy an orthogonal cell each of whose six faces is surrounded by an identical cell (Figure 2) which is in turn in contact with a sextuplet of identical cells and the image is extended indefinitely. In this model we see that a particle crossing a face into an adjoining cell is immediately replaced by an identical one arriving from the opposite face.

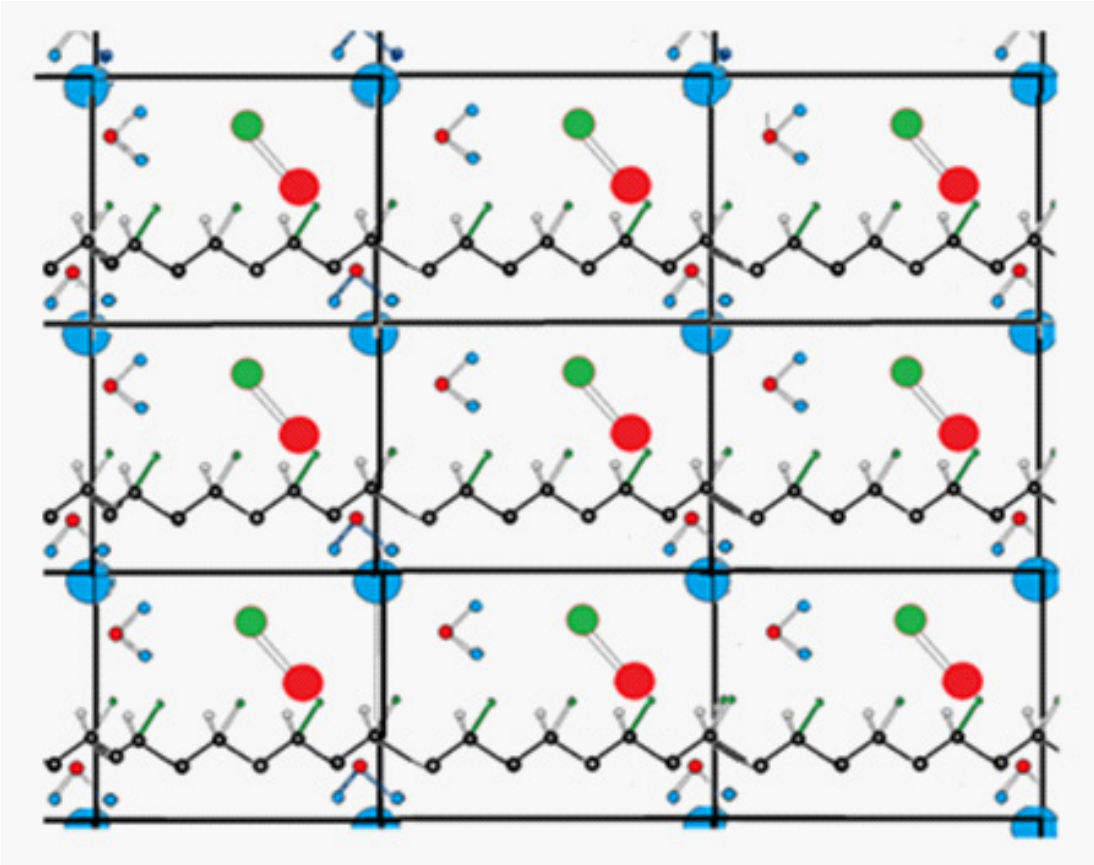


Figure 2: MD unit cell.

The particles' mutual interactions constitute a *force field* which is generated by a force on each particle A, given by FA= fAB where fAB is the force exerted on particle A by particle B. We first consider the unrealistic scenario in which the whole particle assembly is at instantaneously rest. Now particle A is not at equilibrium as it is subject to forces from the remaining *N-1* particles. In fact A responds to the net force by embarking on an accelerating trajectory for a femtosecond-long *timestep* (10^{-15} s). During this brief interval all the other particles also simultaneously undergo displacements to positions from which they impose a new set of forces on the remainder, which

in turn move off to new positions in the course of the second timestep. The particle system has become *dynamic*.

In a normal MD calculation the particles do not start from rest but are assigned a set of initial velocities which match those of the Maxwell-Boltzmann distribution appropriate to that temperature. Consider the particle's altered environment between instants t_i and t_{i+1}. In the course of that interval, (timestep Δt) the particle changes its position to $\mathbf{x}_{i+1}$, its velocity to $\mathbf{v}_{i+1}$ and its to acceleration $\mathbf{a}_{i+1}$. Application of newtonian dynamics to a small interval Δ*t* gives

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \mathbf{v}_i \Delta t + \tfrac{1}{2}\mathbf{a}_i \left(\Delta t\right)^2 \qquad (1)$$

$$\mathbf{v}_{i+1} = \mathbf{v}_i + \tfrac{1}{2_i}\left(\mathbf{a}_i + \mathbf{a}_{i+1}\right)\Delta t \qquad (2)$$

The Forces

The forces generated not only cause the atoms to accelerate but the molecules to *deform* in ways that are calculated from the set $\left\{\mathbf{f}_B^A\right\}$ described in the previous section. The force between a pair of atoms A and B separated by a distance x_{AB} is derived from a set of *atom-pair energy functions* $\left\{V\left(x_{AB}\right)\right\}$:

$$\mathbf{f}_B^{\mathbf{A}} = \left(\frac{\partial V\left(x_{AB}\right)}{\partial x_{AB}}\right) \qquad (3)$$

The functions {*V*} from which these forces (which may extended from pairs AB, to triads ABC, etc.) are derived are commonly referred to as *atom potentials*.

To describe the set of contributions $\left\{V\left(x_{AB}\right)\right\}$ for this chapter presents a dilemma of choosing between (a) explicit atom potentials whose functional forms transparently reflect the interaction expressed, and which can be applied to any atom species in a molecular system and (b) non-explicit 'black box' force fields designed for the investigations of specific systems such as proteins or nucleic acids. For reasons of generality and transparency we follow course (a); however the user may retain the option of selecting or formulating implicit force fields for course (b) which may carry claims of reliability as they are purpose-built for specific classes of molecular system and which may be incorporated in the MD software.

The atom potential functions describing interactions involving atoms A, B, C, … constituting the molecular system that is to be simulated may be classified into (I) *bonded* potentials if the group of atoms are sufficiently close for some of them to be associated with chemical (valence) bonds; otherwise (II) their interaction would be by *non-bonded* potentials.

(I) Bonded Potentials

1. Bond stretching potential $V_b^{AB}(r)$

The force accompanying the compression or extension of a valence bond A-B from r_0 to r may be derived from an atom-pair harmonic potential:

$$V_b^{AB}(r) = \tfrac{1}{2} k_{AB}(r_0 - r)^2 \qquad (4)$$

An alternative bond pair distortion potential is the Morse function eqn. (5), which allows the rupture of the bond, an event not permitted in eqn. (4):

$$V_b^{AB}(r) = D\left[\left\{1 - e^{-a(r-r_0)}\right\}^2 - 1\right] \qquad (5)$$

where D is the depth of the energy well (roughly the bond energy) and a (a function of D) contains the bond's stretching force constant k_{AB}: $a = \sqrt{k_{AB}/2D}$

2. Bond angle potential $V_{ba}^{ABC}(\theta)$

A harmonic potential is often employed to describe the distortion of a bond angle A-B-C from its equilibrium value θ_0 to θ:

$$V_{ba}^{ABC}(\theta) = \tfrac{1}{2} K_{ABC}(\theta_0 - \theta)^2 \qquad (6)$$

where K_{ABC} is the bending force constant associated with the atom triad ABC.

3. Torsional potential $V_{tors}^{ABCD}(\phi)$

This potential measures the energy change when a bond is *twisted* and can be used in simulating π bonds. In a bond B-C let atom B be bonded also to A, and atom D to C. Then the atoms A-B-C-D are in the intersecting planes ABC and BCD. When the group is viewed along the B-C line a twist of the A-B and C-D bonds about B-C is measured by the *dihedral angle* ϕ between the planes. Such torsion about B-C causes ϕ to trace a periodic variation of the potential energy. If ϕ_0 is the minimum-energy angle, the torsional energy for an m-fold barrier of magnitude v in the range $0 < \phi < 360°$ could be described by

$$V_{tors}^{ABCD}(\phi) = v\left[1+\cos(m\phi-\delta)\right] \qquad (7)$$

where δ is a phase angle defining the angular positions of the V_{tors}^{ABCD} minima.

(II) Non-bonded potentials

5. Coulomb potential $V_{coul}^{AB}(r_{AB})$

The potential energy contributed by the electrostatic forces on a pair of atoms A,B separated by a distance r_{AB} with partial charges q_A and q_B is

$$V_{coul}^{AB}(r_{AB}) = \frac{q_A q_B}{4\pi\varepsilon_0 r_{AB}} \qquad (8)$$

Here any solvent molecules present are assumed to be explicitly included in the MD system so that the principal charge-screening effects are accounted for in the responses of the solvent atoms. If the solvent is described as a *bulk medium effect* an effective dielectric constant *D* is included in the denominator of eqn. (8).

Van der Waals potential $V_{vdW}^{AB}(r_{AB})$

The condensation of gases shows that atoms exert mutual attraction at large separations. But due to the lack of accurately known interaction energies for widely separated atom pairs this contribution is probably the least reliably known component of any force field. Commonly used formulations are based on structure-optimizing methods that reproduce experimentally-derived molecular geometries or, for simple solvent molecules that permit thermodynamic properties to be derived which are in accord with measurement. This has resulted in the two approaches mentioned at the start of Section 2.1.

Firstly, simple recipes have been proposed to calculate $V_{vdW}^{AB}(r_{AB})$ with the ambitious purpose of providing 'universal force fields' for all the atoms of the Periodic Table. Two common forms for the interaction between a pair of non-bonded atoms A and B are the Buckingham and Lennard-Jones potentials, which have the respective forms

$$V_{vdW}^{AB}(r_{AB}) = D\exp(-Er_{AB}) - \frac{C}{r_{AB}^6} \qquad (9)$$

and

$$V_{vdW}^{AB}(r_{AB}) = \varepsilon\left[\left(\frac{r_m}{r_{AB}}\right)^{12} - 2\left(\frac{r_m}{r_{AB}}\right)^{6}\right] \qquad (10)$$

both of which describe energy wells rising sharply at diminishing separations r_{AB} compared with the weakly negative energies at larger r_{AB}. Of the two formulations, eqn. (10) is the more transparent, with ε measuring the depth of the potential well (from zero energy at $r_{AB} = \infty$) and r_m the value of r_{AB} at this point. The first term on the right hand side in each of eqns. (9) and (10) describes the strong Pauli repulsion as A and B approach to small separations and the second uses the London dispersion energy r^{-6} dependence of the attractive interaction (which appears also in the van der Waals theory of interaction between gaseous molecules). Figure 3 shows a comparison of the interaction energies in a pair of bonded and non-bonded carbon atoms.

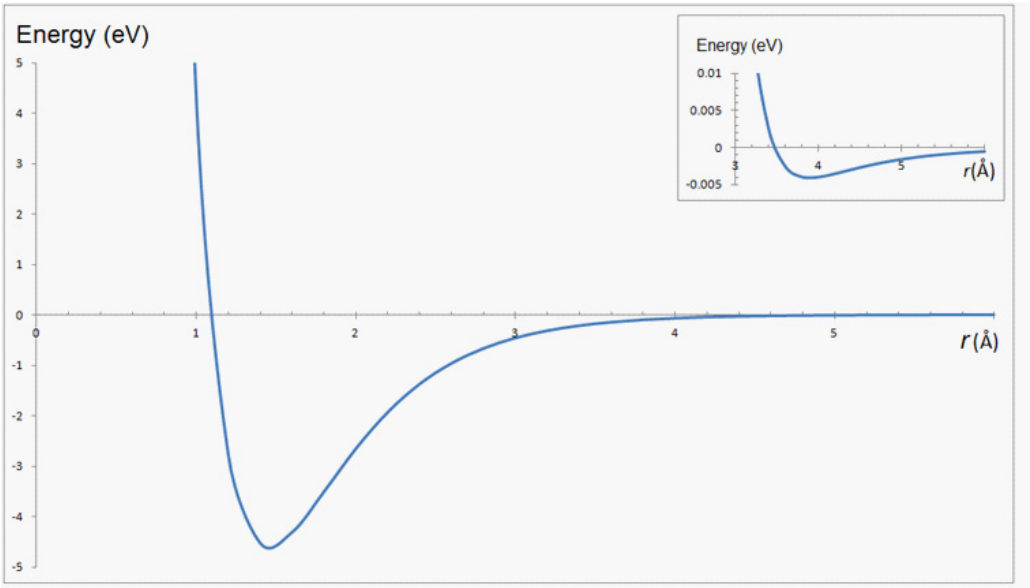


Figure 3: Atom-pair potentials for two bonded carbon atoms (main) and for two non-bonded carbons at various separations *r* (inset). The bonded potential is from a Morse function (eqn. (3)) and the non-bonded is from a Lennard-Jones equation (eqn. (10)).

Secondly, the extensive application of MD to biological systems has resulted in the proposals of several sets of atomistic force fields applicable to specific 'atom types' in this category. Different parameters would then be assigned to atoms of the same species but in different chemical environments (for example a C atom which may be in a -CH_2-, -CO- or a phenyl group, or an oxygen present in hydroxyl, carboxylic etc.). If the system being simulated is large as for a protein system, containing thousands of atoms, for economy of computing times the simulation may be 'coarse-grained', i.e. not all the atoms are explicitly treated in the simulation. Typically the two H atoms in methylene or the three in methyl might be subsumed into the carbon atom and the resulting -CH_2- or -CH_3 'bead' treated as a single pseudo-carbon atom with its own set of atomistic parameters [5].

The total interaction energy between atom A and all the others is the sum of the contributions expressed in eqns. (4) to (10):

$$V^A = \sum_B \left[V_b^{AB} + \sum_c \left[V_{ba}^{ABC} + \sum_D \left[V_{tors}^{ABCD} \right] \right] + V_{coul}^{AB} + V_{vdW}^{AB} \right] \quad (11)$$

To describe *hydrogen bonding* e.g. =O....H–, although pair potentials V_{nb}^{OH} have been proposed, it is usual to describe the association using the Coulomb charges on O and H and applying eqn. (8).

Available Force Fields for Simulation

As there are too many of these for concise inclusion in this chapter the interested reader is encouraged to find the items in the literature. They fall into two groups - those that were developed for use with proteins, polypeptides and nucleic acids (AMBER, CHARM, GROMOS etc.) and others of more general application (DREIDING, OPLS, UFF etc.).

For water as a solvent we refer to a common potential TIP3P [7] which are taken to be a rigid molecule with atomic charges that reproduce the molecular dipole moment and other bulk properties.

Range of the Force Field

The periodic boundary conditions in eqn. (11) in principle require the calculation of an infinite number of the non-bonded terms V_{coul}^{AB} and V_{vdW}^{AB} terms in (8) to (10). The r_{AB}^{-6} factor in V_{nb}^{AB} in eqns. (9) and (10), however, ensures a rapid convergence of this term with distance (and a realistic cutoff of about 9$\mathring{A}$ is usually imposed around atom A). The convergence of the Coulomb potential with r_{AB}^{-1} in eqn. (8), on the other hand, is too slow for a cutoff to be appropriate. The problem is overcome by the incorporation of an Ewald summation in the MD program. This consists of computing a short-range summation in real space combined with a long-range contribution evaluated employing a Fourier transform into reciprocal space, resulting in tractable summations.

Partial Atomic Charges and Molecular Geometry

The Coulomb energy using eqn. (8) requires a set of partial electrostatic charges $\{q_i\}$ on the atoms. These may either be inferred from recipes for formal charges and atomic electronegativities [8] or better, by performing a quantum chemical calculation on the static molecular species [9], [10].

The initial geometries of the molecules in the MD input can be taken from diffraction or spectroscopic data on the species or, if unavailable, on a simpler molecule with similar structural features. Alternatively the structure may be obtained from quantum chemical calculations on the molecule (or its model) with the 'geometry optimization' option in place. Of course the geometries into which the system is relaxed by the MD is consistent with the force field potentials in eqn. (11) but not necessarily with the 'gas phase' optimum quantum chemical geometry produced by the quantum calculations which were performed as a guide to conformations.

Since electronic wave functions are used to calculate charges and energies, the user must select a suitable method and basis set [11]. Quantum chemical code packages are available, from which Density Functional Theory **(DFT)** at a B3LYP level may be deployed. Because of the lower criticality of the *molecular geometry* referred to in the previous paragraph, a 6-31G (d,p) basis set is probably sufficient, but to obtain reliable atomic charges for use in eqn. (8) diffuse functions are recommended e.g. 6-31+G (d,p). There is little consensus on which *definition* of partial atom charge should be used - common ones are Mulliken populations, Natural Population Analysis **(NPA)**, Atoms in Molecules **(AIM)** etc. The properties calculated using these definitions and tested against experiment (e.g. pKa) seem in fact to be best predicted by NPA charges [9].

Run Times

An MD run would involve the sequence of a large number of femtosecond timesteps to simulate a real time event on a molecular scale. The Einstein-Smoluchowski equation $\lambda = \sqrt{2Dt}$ [12] predicts that the mean time for a particle to migrate a distance $\lambda = 1\mathring{A}$ in an aqueous medium with diffusivity D is about 4 ps. If the molecular event of interest were the passage of a diffusant through a pore or molecular channel embedded in a $40\mathring{A}$ thick bilayer membrane of a cell, the minimum simulation time would then be of the order of 0.1 ns, or 10^5 timesteps. In channel migration consisting of a series of simple *passive* Stokes diffusion episodes in an aqueous medium depending only on temperature, density and migrant concentration gradients the simulation, a satisfactory MD might require about a million timesteps. If the ion transporter or membrane were more complex, and the migration a series of specifiable *active* processes simulation of the transport processes would take some orders of magnitude longer.

Input Data for a Molecular Dynamics Program

The atomic positions in the unit cell are defined by Cartesian coordinates which are obtained by calculation or from a molecular visualization program. The user selects a force field according to Section 2.1 and the partial atomic charges from calculations as described in Section 2.4. Additional input items include the required total running time for the MD and the system's temperature and pressure. While a thermally coupled thermostat ensures that the runs are performed at constant temperature, the user must specify whether the volume or the pressure is to be held constant. (If the systems involve membranes other restrictions too may be applicable.) Time intervals are specified at which the structure and other episodes of the atomic system are to be monitored.

Having performed the required number of timesteps some ensemble-averaged and time-averaged properties are calculated. Output files include the atomic positions, velocities and forces at the monitoring times. From these, the program calculates diffusion coefficients of migrants and other features (Section 3) that characterize the dynamics, which may be derived from the atom trajectories. Some quantities, particularly those associated with the time-evolving structure, may be recorded as 'snapshots' or animated graphics.

MOLECULAR DYNAMICS OUTPUT

Radial Distribution Function *g(r)*

The Radial Distribution Function **(RDF)** describes the time-averaged structure of the system. Atom B in one molecule is part of the environment of atom A which is taken to be the origin of concentric spherical shells with separation Δr. If there is an average of $n_B(r)$ B atoms in the shell of radius r, the function

$$g_{AB}(r) = \frac{n_B(r)}{4\pi r^2 \rho \Delta r} \qquad (12)$$

measures the probability of encountering atom B at a distance r from A. At large r the rdf in eqn. (12) tends to unity. The two rdf curves $g_{AB}(r)$ in Figure 4 reflect the degree of order of two species of B atom surrounding A with B as a solvent- or ionic- atmosphere around A. Here a Na^+ ion has entered a crown-ether based channel (mounted on the helical oligopeptide shown in Figure 1) [4]. The red and blue traces respectively denote the probability of the Na^+ migrant being found near the O and C atoms of the crown ether ring, showing that as expected for the electronegative oxygen (that bears an appreciable negative charge) the Na^+ has a greater association with it than with the carbon atoms of the ring; the subsidiary O peaks are for atoms on further CE rings. At higher temperatures the reduced structural order results in shorter, broader peaks.

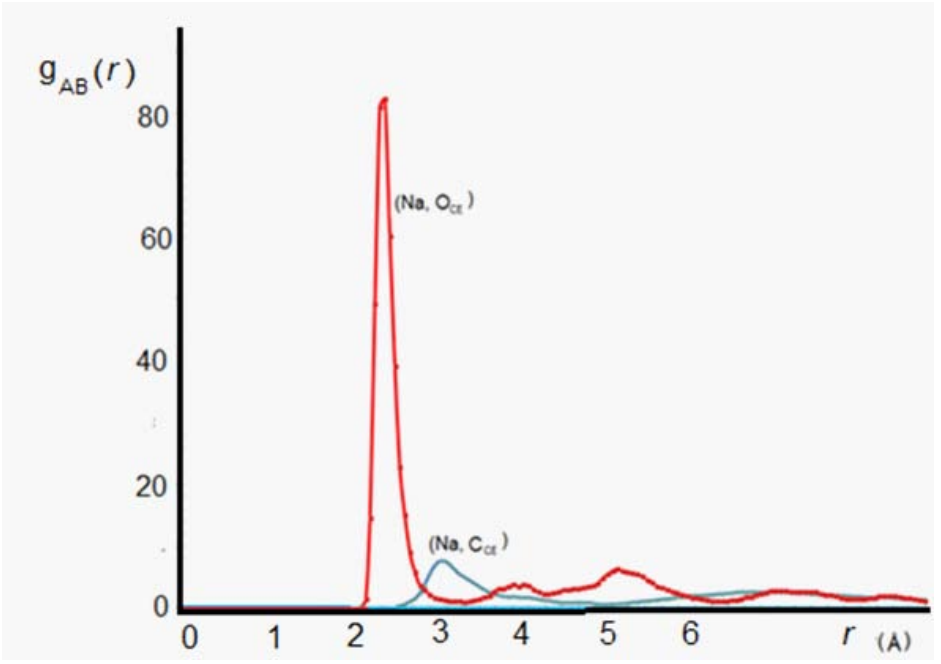


Figure 4: Radial distribution curves for Na^+ ... O and Na^+ ... C in a crown ether channel.

Potential of Mean Force *w(r)*

This quantity describes the dynamics by keeping one part of the system fixed (constraining some element of the atomic configuration) [5]. The *average force* exerted by all the rapidly varying configurations of another component of the system is then monitored to derive the varying potential at these positions. The constrained quantity might be the instantaneous position of a migrant species in the ion transporter and the averaged quantities then constitute the potential of mean force from (for instance) the solvent molecules or bilayer membrane. Alternatively if the constraints were specified as atom-pair separations in the process investigated, *w(r)* could be used to monitor the process as the average work needed to bring the two atoms from infinite separation to one that at separation *r*. It is a free energy quantity, implicitly taking into account the redistribution of its molecular environment and is related to the system's radial distribution function *g(r)*, by

$$w(r) = -kT \ln g(r) \qquad (13)$$

The potential *w(r)* has been invoked as a tool to understand ion permeation and selectivity in membrane channels [13], and the dynamics of confined polymers [6].

Velocity Autocorrelation Function *G(t)*

The dynamics may be characterized by following the change of a quantity - in this case the velocity of a particle species - with time [14]. Let the velocity of a particle be $\mathbf{v}_0$ at time 0 and $\mathbf{v}_t$ at time *t*, where **v** is a three-component vector (v_x, v_y, v_z). If the particle's trajectory during this interval were free then $\mathbf{v}_t = \mathbf{v}_0$ for all t and the scalar product ($\mathbf{v}_0$, $\mathbf{v}_t$) would be $v_t^2 = v_{x0}^2 + v_{y0}^2 + v_{z0}^2 = \left|v_0^2\right|$. However due to effects from the other particles the particle's velocity varies with time and the value of ($\mathbf{v}_0$, $\mathbf{v}_t$) depends on the particle's trajectoral history up to time *t*. As *t* becomes large the randomness of the magnitudes and directions of the other *N* particles causes scalar product ($\mathbf{v}_0$, $\mathbf{v}_t$) to diminish, sometimes with brief oscillations (Figure 5). After attaining zero at long times the scalar product indicates that $\mathbf{v}_t$ has lost all 'memory' of its time-evolution. In practice the velocity autocorrelation function $G_v(t)$ is defined as

$$G_v(t) = \frac{1}{N} \sum_{i=1}^{N} v_i(t_0).v_i(t) \qquad (14)$$

by averaging over 'time windows' with different end points t_0 and t_N. The function $G_v(t)$ in eqn. (14) can be a sensitive parameter for a particle's dynamic interaction with its environment up to time *t*.

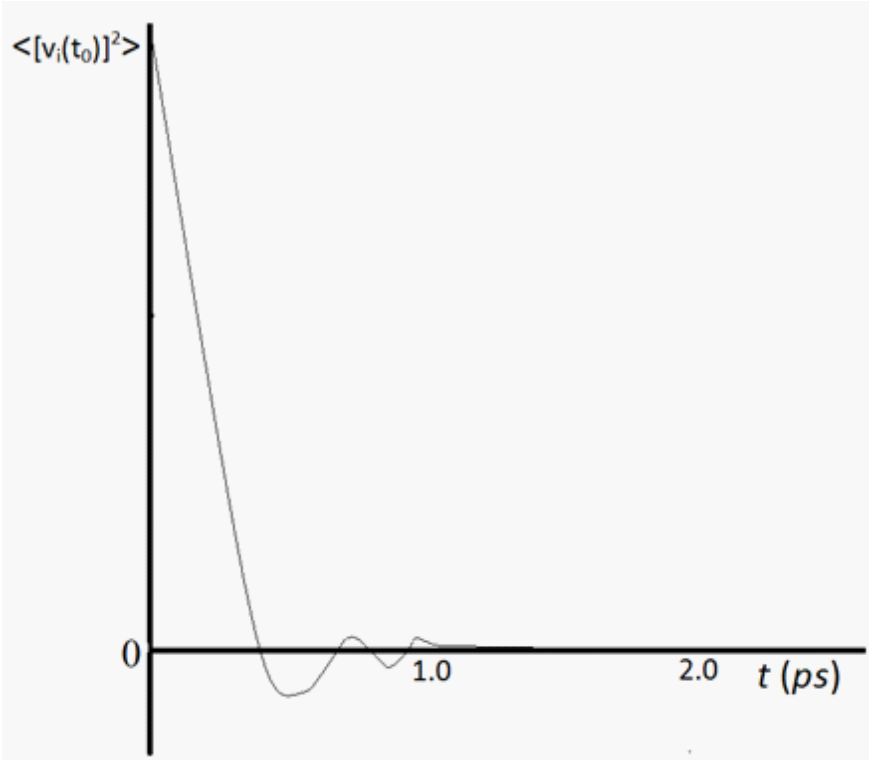


Figure 5: Velocity autocorrelation function.

SOME CURRENT APPLICATIONS AND THE FUTURE TREND OF MOLECULAR DYNAMICS

Although other computational methods e.g. Molecular Mechanics and Monte Carlo simulations also emulate chemical and physical changes for systems which may be inaccessible to experiment, the fact that MD does so in real time (albeit on scales expanded by $\sim 10^{16}$ for running times!) enables it to propose mechanisms for the transport of ions in synthetic monolayers [15], to reveal the dynamical behavior of lipid bilayers [16] and their permittivity to the migration of species [3]. These studies include the energetics of ion conduction through K^+ channels [17], the docking of drug molecules [18], DNA-ligand interactions and charge migration and the folding/unfolding of nucleic acid duplexes [19,20] and ion transporters including synthetic molecules in model biological systems [4].

A limitation of current MD is the unavoidable shortness of the femtosecond timestep that is necessary to integrate the equations of motion in Section 1. Consequent difficulties are its application to the common chemical and physical processes in the micro- or millisecond timescales of conformational changes in bio-macromolecules, for which extremely long running times would be required. Brute force methods using high computational power have indeed been applied to electrochemical processes involving enzymes on electrodes with times of the order of

microseconds [21], and even millisecond long protein-foldings have been simulated [2]. But the timescale problem has also been countered by other remedies. Consider for example a system in which the particle dynamics are normal but the special interest to the user (for example the passage of a migrant through the mouth of a channel) is a rare event. Then 'Steered MD' [22] or Interaction MD [23] methods may be deployed, both of which employ an interactive approach to ensure that the particle system is close to the configuration at which the special effect (channel entry) occurs.

The quantum behavior of mobile hydrogen atoms (and other small atoms) renders the particles invalid for description by the laws of motion of 'classical' particles. The dynamics of a hydrogen bond binding the base pairs in DNA or that may be involved in a [=O...H–O– ↔ –O–H... O=] tautomerism defy a description by MD. They must instead be treated by 'hybrid' methods embodying quantum theory [24].

As computing power and methodologies develop it is envisaged that molecular dynamics will eventually be subsumed into all-quantum *ab initio* methods. This will obviate the component terms in eqn. (11) as the electronic charges, the binding and the current uncertain non-bonding terms eqns. (9) and (10), all of which actually continuously change within the MD time period. The dynamics of the atoms too, free of Born-Oppenheimer restrictions, will be fully described by quantum theory, a procedure which hitherto has been initiated for simple systems [25], but which will ultimately constitute a fundamental principle of molecular dynamics.

References

1. McCammon JA, Harvey SC, Dynamics of proteins and nucleic acids, Cambridge: Cambridge University Pres. 1987.
2. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How fast-folding proteins fold. Science. 2011; 334: 517-520.
3. de Groot BL, Grubmüller H. The dynamics and energetics of water permeation and proton exclusion in aquaporins. Curr Opin Struct Biol. 2005; 15: 176-183.
4. Morton-Blake DA. Diffusion phenomena in engineering materials. Belova I, Murch G, Öchsner A. editors. In: Diffusion Foundations 4, Aug. 2015; Chapter VI.
5. Müller-Plathe F. Coarse-graining in polymer simulation: from the atomistic to the mesoscopic scale and back. Chemphyschem. 2002; 3: 755-769.
6. Eslami H, Varzaneh HAK, Müller-Plathe F. Coarse-grained computer simulation of nano confined polyamide-6,6, Macromolecules. 2011; 44: 3117-3128.
7. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. 1983; 79: 926-935.
8. Abraham RJ, Grant GH, Haworth IS, Smith PE. Charge calculations in molecular mechanics. Part 8. Partial atomic charges from classical calculations. J Comput Aided Mol Des. 1991; 5: 21-39.
9. Gross KC, Seybold PG, Hadad CM. Comparison of different atomic charge schemes for predicting pKa variations in substituted anilines and phenols, International Journal of Quantum Chemistry. 2002; 90: 445-458.
10. Thompson JD, Xidos JD, Sonbuchner TM, Cramer CJ, Truhlar DG. More reliable partial atomic charges when using diffuse basis sets. Phys Chem Comm. 2002; 18: 117-134.
11. For basis sets see https://www.shodor.org/chemviz/basis/teachers/background.html.
12. Atkins P, de Paula J. Physical Chemistry, 8th edition, Oxford: Oxford University Press. 2006, Chapter 21.

13. Allen TW, Andersen OS, Roux B. Molecular dynamics - potential of mean force calculations as a tool for understanding ion permeation and selectivity in narrow channels. Biophys Chem. 2006; 124: 251-267.

14. JJ Erpenbeck and WW. Wood. Molecular-dynamics calculations of the velocity-autocorrelation function, Phys. Rev A. 1982; 26: 1648-1675.

15. Morton-Blake DA, Leith D. Molecular dynamics of ions in two forms of an electroactive polymer. Ann N Y Acad Sci. 2009; 1161: 105-116.

16. Lyubartsev AO, Rabinovich AL. Recent development in computer simulation of lipid bilayers. Soft Matter. 2011; 7: 25-39.

17. Bernèche S, Roux B. Energetics of ion conduction through the K^+ channel. Nature. 2001; 414: 73-77.

18. Gupta J, Nunes C, Vyas S, Jonnalagadda S. Prediction of solubility parameters and miscibility of pharmaceutical compounds by molecular dynamics simulations. J Phys Chem B. 2011; 115: 2014-2023.

19. Barnett RN, Cleveland CL, Joy A, Landman U, Schuster GB. Charge migration in DNA: Ion-gated transport, Science. 2001; 294: 567-571.

20. Pérez A, FJ, Orozco M. Frontiers in the Molecular Dynamics Simulations of DNA, Accounts of Chemical Research. 2012; 45: 196-205.

21. Oteri F, Ciaccafava A, de Poulpiquet A, Baaden M, Lojou E. The weak, fluctuating, dipole moment of membrane-bound hydrogenase from Aquifex aeolicus accounts for its adaptability to charged electrodes. Phys Chem Chem Phys. 2014; 16: 11318-11322.

22. Patel JS, Berteotti A, Ronsisvalle S, Rocchia W, Cavalli A. Steered molecular dynamics simulations for studying protein-ligand interaction in cyclin-dependent kinase 5. J Chem Inf Model. 2014; 54: 470-480.

23. Grayson P, Tajkhorshid E, Schulten K. Mechanisms of selectivity in channels and enzymes studied with interactive molecular dynamics. Biophys J. 2003; 85: 36-48.

24. Nunthaboot N, Pianwanit S, Parasuk V, Ebalunode JO, Briggs JM, et al. Hybrid quantum mechanical/molecular mechanical molecular dynamics simulations of HIV-1 integrase/inhibitor complexes, Biophysical Journal. 2007; 93: 3613-3626.

25. Car R, Parrinello M. Unified approach for molecular dynamics and density-functional theory. Phys Rev Lett. 1985; 55: 2471-2474.

Austin Publishing Group

GPU Accelerated Molecular Dynamics Simulations in Predicting the Protein-Protein Binding Affinity from Residues Interactions within the Binding Surface

Chong WL[1], Gautam V[1], Zain SM[1], Rahman NA[1] and Lee VS[1]*

[1]Department of Chemistry, Faculty of Science, University of Malaya, Malaysia

***Corresponding author:** Vannajan Sanghiran Lee, Department of Chemistry, Faculty of Science, University of Malaya, Kuala Lumpur 50603, Malaysia, Tel: 603-79677022 ext. 2142; Fax: 603-79674193; Email: vannajan@gmail.com

Published Date: December 01, 2016

ABSTRACT

Graphic Processing Units (GPUs) that were first introduced for visualization particularly in gaming industry are now widely exploited in computational and theoretical research when CPU has reached its limitation to satisfy the speed demand if a macromolecular system is studied. It is a revolution of parallel computation that has been crucial to stimulate the growth of the computational study for various purposes. GPU-accelerated workstation surpasses that of conventional CPU as molecular properties of one nano second of a macromolecular system is now collectible within hours. Hence, it is the solution to the bottleneck of simulating a very large biological system like membrane. Apart from high speed, GPU-enabled machine is proven to save more energy and meeting the demand of green technology. With the advances available and increasing popularity, the cost involved in setting up a workstation that runs on GPUs would no longer be a burden. Our group has successfully transformed the CPU workstations to run on GPUs by replacing the graphic card to a higher-end one together with some simple installation steps. GPUs have become a more powerful tool when CUDA is implemented and complementary to the

machine. MD simulations that run on pmemd.cuda have been showing its outstanding speed and efficiency in finishing the submitted jobs. It would be useful if one needs to simulate a large system or a great number of systems. It has also made the protein design from MD simulations possible and of course the predictions would be more reliable and agreeable with observations reported from experimentalists. We have demonstrated some examples in text, which describes how a high-speed machine becomes essential and crucial in dissecting the protein-protein interactions (PPIs) which is normally a missing piece for experimentalists and also calculating the binding affinity that correlated well with reported findings.

INTRODUCTION

In the past 20 years, molecular modeling has advanced from simulating small system (<300 atoms) to the routine modeling of entire proteins in solution with lipids and explicit water (30,000-100,000 atoms). This remarkable achievement has been in large part the result of the use of high-performance computing (HPC). However, simulations are still far from realizing the full potential of computational molecular biology because of the limited time scale (usually sub 1 μs) that they can practically achieve. The recent evolution of graphics processing units (GPUs) into general-purpose, fully programmable, high performance processors represents an important technological innovation that may realize the full potential of atomistic molecular modeling and simulation. To exploit the computational power of GPUs, it is necessary to redesign and reprogram algorithms to suit the architectures of these devices. There are over a hundred GPU-accelerated applications in computational chemistry field (Quantum Mechanics, Molecular Mechanics, Molecular dynamics) recently available and growing. McCammon and colleagues have reported the first molecular dynamics (MD) simulations on an enzyme in year 1977 [1] and since then MD simulations have evolved to become an important tool in rationalizing the behavior of biomolecules. The field is continuously progressing such that the molecular properties of a small enzyme with 500 atoms could be collected on the microsecond time scale [2–4] from the initially 10-ps-long simulation and simulations containing millions of atoms can be considered routine [5,6]. Nonetheless, simulations are numerically very intensive, and employing conventional CPU centric hardware it requires access to large-scale supercomputers or well-designed clusters with expensive interconnects that are beyond the reach of many research groups.

Numerous attempts have been made over the years to accelerate classical MD simulations by exploiting alternative hardware technologies such as ATOMS by AT&T Bell Laboratories [7], FASTRUN by Columbia University and Brookhaven National Laboratory [8], MDGRAPE by RIKEN [9], and most recently Anton by DE Shaw Research LLC [10]. It is however the mentioned approaches have failed to make an impact on mainstream research because of their excessive cost. Additionally, these technologies have been based on custom hardware and do not form part of what would be considered a standard workstation specification. Such technologies have made the experiments difficult and hence the respective development and innovation are not sustained. It further limits them from being ubiquitous community-maintained research tools.

Graphics processing units (GPUs), on the other hand, have been an integral part of personal computers for decades, and a strong demand from the consumer electronics industry has resulted in significant sustained industrial investment in the stable, long-term development of GPU technology. In addition to low prices for GPUs, this has led to a continuous increase in the computational power and memory bandwidth of GPUs, significantly outstripping the improvements in CPUs. As a consequence, high-end GPUs can be considered as standard equipment in scientific workstations, which means that they either already exist in many research laboratories or can be purchased easily with new equipment. This makes them readily available to researchers and thus attractive targets for acceleration of many scientific applications including MD simulations. The nature of GPU hardware has recently made their use in general purpose computing challenging to all but those with extensive three-dimensional (3D) graphics programming experience. The development of application programming interfaces (APIs) targeted at general purpose scientific computing has reduced this complexity substantially such that GPUs are now accepted as serious tools for the economically efficient acceleration of an extensive range of scientific problems [11,12]. The computational complexity and fine grained parallelism of MD simulations of macromolecules makes them an ideal candidate for implementation on GPUs. Modern MD algorithms implemented on GPUs is capable of performing for both, implicit and explicit solvent models [13], and exceeds, what is achievable with any current CPU-based supercomputer. A number of studies conducted previously have investigated the use of GPUs to accelerate MD simulations [14–20]. For a better insight of the use of GPUs for acceleration of condensed phase bio-molecular MD simulations, we refer to the recent review [12].

Availability of such high performance GPU implemented with implicit solvent generalized Born (GB) MD for the AMBER [21] and CHARMM [22], speeds up the simulation time, particularly in protein-protein system. We also aim to use high-performance GPU implementation pairwise additive force fields on CUDA enabled NVIDIA GPUs which are implemented within the AMBER [23,24] PMEMD dynamics engine in a manner to be as transparent to the user as possible. The processing power of GPUs can be used both in serial and multiple and can achieve performance in comparison to conventional CPU clusters. Previous testing study revealed that use of a GPU was faster than using locally available CPUs (simulations using 1 GPU were ~50% faster than those using 32 Xeon 3.4 GHz CPU cores). Moreover, GPUs provide promising systems for energy efficient scientific computing. It was found that use of a GPU consumed significantly less energy (~3 MJ per nanosecond of dynamics using 1 GPU compared to 10 MJ per nanosecond of using 32 CPU cores). Due to its high performance and proven efficiency in earlier studies, GPU was made the choice for the study.

In this study, we aim to accelerate several computational researches and engineering applications on molecular dynamics simulation with NVIDIA® Quadro® GPUs for the large time scale simulation for several novel bio-molecular / material systems such as membrane, antibody-antigen, proteins, DNA, carbon nanotube, and ionic liquids. The expected outcome is to investigate the behavior of biological and material system under the large time scale simulation with graphics processing units GPU.

GPU BUILDING AND SOFTWARE INSTALLATION FOR MOLECULAR DYNAMICS SIMULATION WITH AMBER

Several GPU-accelerated applications in computational chemistry field (Quantum Mechanics, Molecular Mechanics, Molecular dynamics simulation (MDs)) are listed in the NVIDIA website. (http://www.nvidia.com/object/gpu-applications.html, July 25, 2015). Our interest is in the use of molecular dynamics simulation with AMBER software that was introduced in version 11, with the ability to use NVIDIA GPUs to massively accelerate PMEMD for both explicit solvent PME and implicit solvent GB simulations with further extend improvement in AMBER 12 and AMBER 14, we now can perform MDs with ~30% performance improvement for single GPU runs and have an additional support for multi-GPU runs providing enhanced multi-GPU scaling. Some considerations have to be made for MDs simulation on AMBER (http://ambermd.org/gpus/, July 25, 2015) for the supported features, supported GPUs and recommended hardware (http://ambermd.org/gpus/recommended_hardware.htm, July 25, 2015), system size limits for implicit and explicit solvent simulations, accuracy considerations. With AMBER 14 the single and double precision models (SPSP, SPDP and DPDP) have been depreciated and replaced with a new hybrid model, SPFP that combines single precision calculation with fixed precision accumulation. Its accuracy is as good as or better than the original SPDP model. [25]

Detailed information on the implementation specifics, methods used that could perform with controlled accuracy, and respective validation are available for routine microsecond molecular dynamics simulations with AMBER - Part I: Generalized Born and Part II: Particle Mesh Ewald [26,27]

Practically, GPU building and software installation for molecular dynamics simulation with AMBER can be done with three main steps for the modern desktop computer by adding the external GPU card. Quick steps and commands in setting up the GPU and software can be found in Appendix A while overview of the procedures could be found in Figure 1. With the introduction of CUDA, GPU has attracted more attention [28]. It brought a critical change in the field as CUDA could actually perform well with pmemd (AMBER) and it reported an impressive high speed, about 11 times faster than conventional CPU at the very first place [29]. CUDA has been widely exploited since then as GPU power could be increased and optimized. Pmemd program in AMBER is now evolving to pmemd.cuda that makes collection of molecular properties for 1 nano second possible in few hours with our less sophisticated yet economical machine (specifications are described in Table 1). Even though our GPUs may not able to outperform those high-end GPU that cost a few ten thousands USD dollars but they have moved our research a few steps forward with limited budget. As recommended by AMBER in using GPU cards manufactured by NVIDIA, we have performed some speed tests to compare Quadro card against their closest GeForce equivalents. Of the two higher-end GPUs we have, one runs on NVIDIA Quadro K2000 while another one on Asus Geforce GTX680. The price comparison to set-up simple desktop GPU for MDs simulation was listed in Table 1.

Table 1: Price comparison for economical GPU for MDs simulations.

HP workstation Z220 (RM 3990)	Customised Intel workstation (RM 5995)
Power 400W Intel core i7-3770 3.4GHz 8GB (2x4GB) DDR3-1600 nEcc RAM 4 cores 2 x NVIDIA K2000 GPU cards with 2GB RAM (128 bits) (Note: Single NVIDIA K2000 = RM 2090)	Power 650 w Intel core i7 4770K 3.5 GHz 16GB (2x8GB) DDR3 RAM 4 cores Asus GTX 680 with 2GB DDR5 (256 bits) (Note: Single GTX 680 = RM 1930)

(As of Nov 11, 2015, 1 RM = 0.23 USD).

We have simulated a protein-protein complex that made up by 47502 atoms in total after water molecules and salt were added using the two workstations that run independently of each other. The required simulation time was 8.75 ns/day and 4.70 ns/day for workstation that runs on Asus Geforce GTX680 card and Quadro K2000, respectively. In contrast, the simulation speed for similar complex was 0.11ns/day when it runs on sander (CPU). Apart from protein-protein system, we have also observed that the performance in term of speed of the GPUs is in the order such that GTX680> K2000 > Q2000D in protein-ligand systems. Coupling the advanced GPU card with CUDA code, it opens a door for performing sophisicated study particularly mutation study which normally a time consuming and expensive process. Computational assay that investigated drug binding of neuraminidase H5N9 [30] has been reported and it indeed excites us as we can now gain insight into mutation that could either enhance or harm the binding activity. Probing the protein-protein interactions and to reveal the binding affinity have been a struggle before the GPU-enabled machine is made available. Kodchakorn et al. [31] had probed the binding activity between ankyrin repeat protein and maltose binding protein with pmemd.cuda code and were able to predict the absolute binding affinity which is comparable to experimental results. However, the data produced would be questionable if it is not concluded from long time-scaled MD simulations. Apart from speed, energy conservation would be another concern as some calculations are relatively expensive. As reported elsewhere, GPUs are observed and proved to be more energy saving compared to CPUs with improved performance. In short, transfroming a CPU to GPU is simple with few steps as decribed earlier and it is much more afforadable. Wherener there is a demand of very high speed calculation, one needs to consider to have to a higher-end graphic card installed. Together with pmemd.cuda, a GPU will absolutely empower the machine to accelerate the calculations effectively.

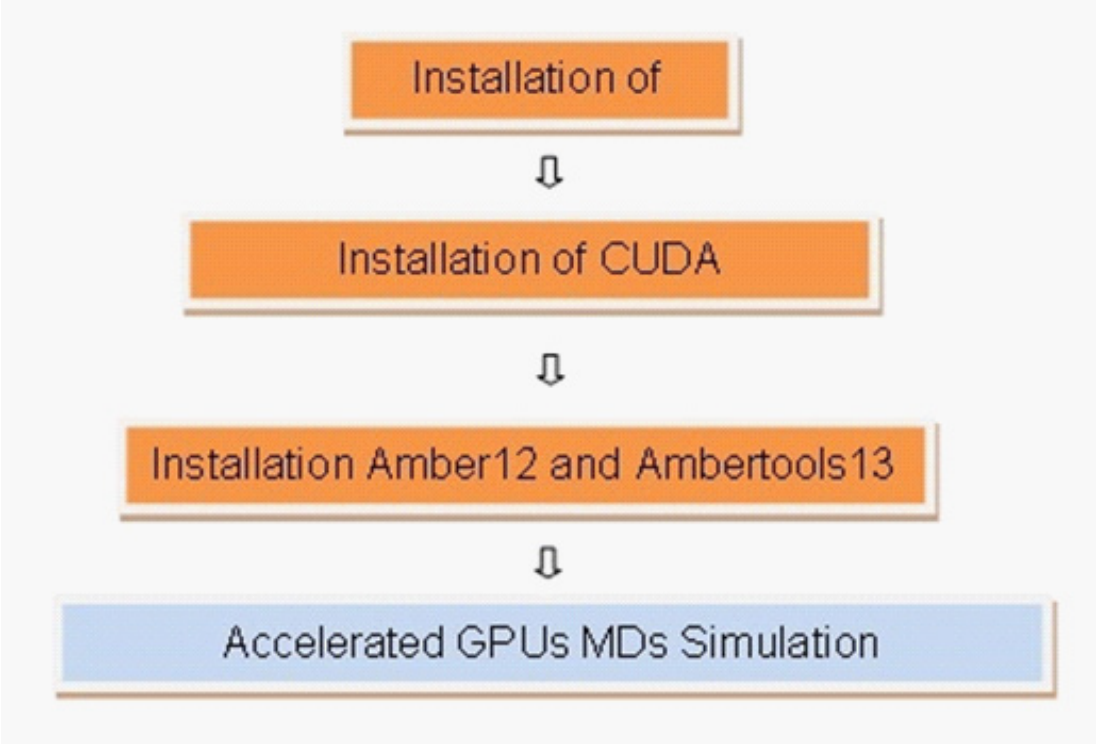


Figure 1: Steps in GPU building and software installation for molecular dynamics simulation with AMBER. For more information, 1) Installation of NVIDIA driver. (http://www.nvidia.com/Download/index.aspx?lang=en-us) 2, July 25, 2015) Installation of CUDA programming. (https://developer.nvidia.com/cuda-downloads) 3) Installation Amber12 and Ambertools13. (http://jswails.wikidot.com/installing-amber12-and-ambertools-13, July 25, 2015).

RESIDUE INTERACTION IN THE BINDING SURFACE OF PROTEIN-PROTEIN COMPLEX THAT CONTRIBUTED TO BINDING AFFINITY FROM ACCELERATED GPU MOLECULAR DYNAMICS SIMULATIONS

The binding free energy and identification of the hot-spot residues of proteins can be achieved using the Molecular Mechanics–Poisson-Boltzmann Surface Area/Generalized Born Surface Area (MM-PPSA/GBSA) protocols from molecular dynamics simulation which can be run in longer time scale via accelerated GPU. The aforementioned protocols have been exploited to study different protein-ligand [32-35] and protein-protein interactions [36]. In addition, calculation for each energy term can be performed without a large training set that fits various parameters under Molecular Mechanics–Poisson-Boltzmann Surface Area/Generalized Born Solvent Area (MM-PBSA/GBSA) protocol [37] has made it an advantage to be more efficient to that of free energy perturbation (FEP) and thermodynamic integration (TI) methods [38]. In general, the complex structure for protein-peptide and protein-protein complexes need to be prepared. The initial structure of the specific complex system can be taken from the X-ray structure from protein

data bank (PDB). However, if no complex structure is available, the initial structure for protein-peptide and protein-protein complexes can be predicted and generated using molecular docking software eg. Autodock [39], Zdocker [40] and SwarmDock [41]. All missing hydrogen atoms for each system can be added using the LEaP module in AMBER later. The predicted ionization states for amino acid residues with potentially charged side chains will be calculated. All systems can be solvated using TIP3P water in a 20Å^3 box using Na+ or Cl– as the neutralizing counterion. The more forcefields should be added like ff12SB [42] are used to model the protein. Figure 2 shows the flow chart for the molecular dynamics simulations for protein-peptide and protein-protein binding affinity via accelerated GPU.

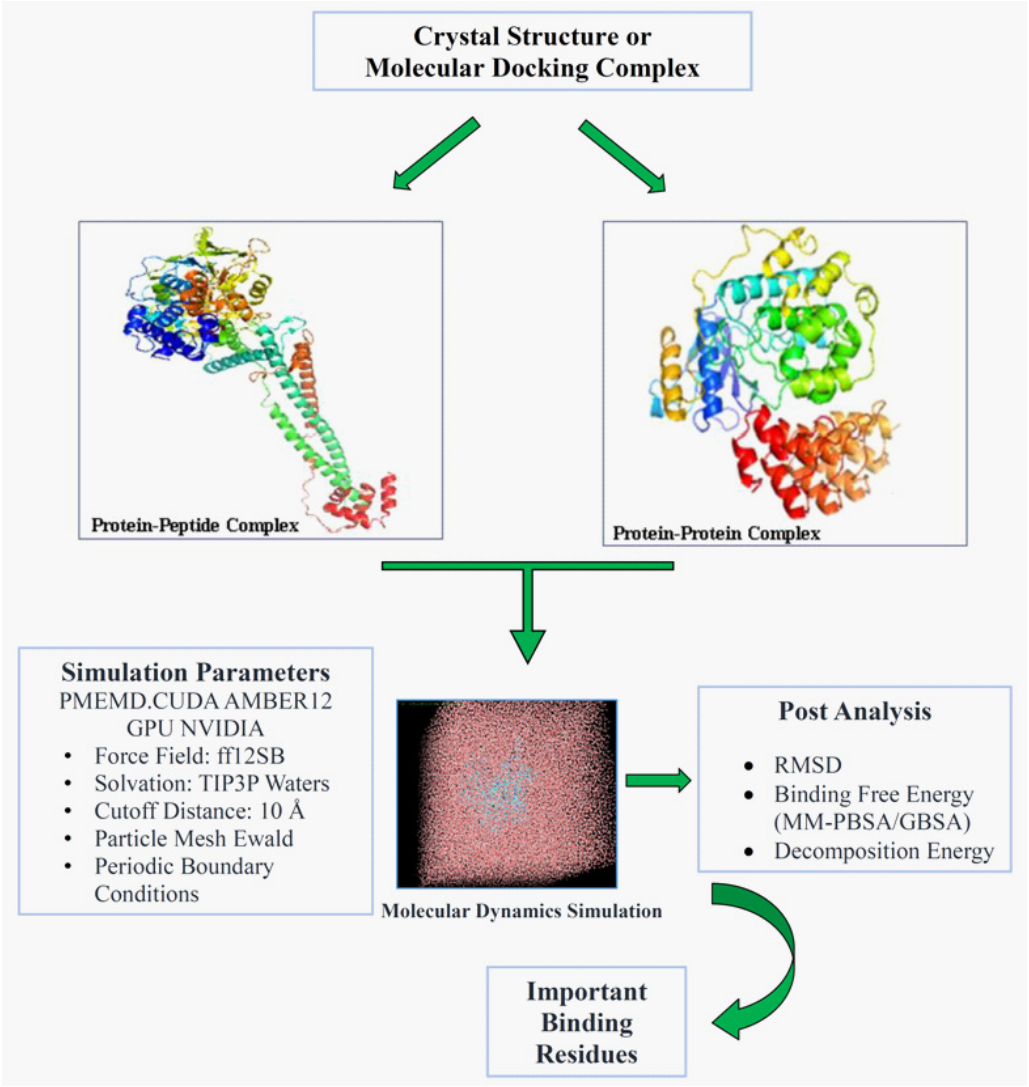
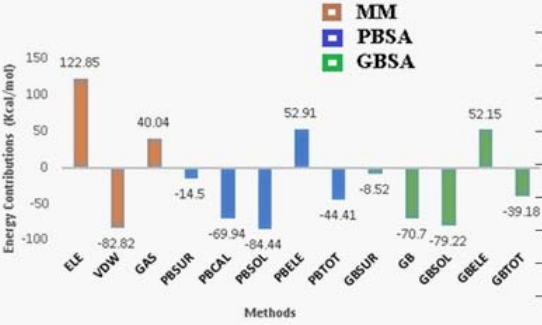


Figure 2: The flow chart for the molecular dynamics simulations for protein-peptide and protein-protein complex.

To elaborate in details how GPUs have assisted our work, we have taken one of our systems as example. Complex of designed ankyrin repeat (AR) structures of integrin-linked kinase (ILK) with PINCH1 were simulated to probe the binding affinity, protein-protein interactions and important residues that are involved in the interaction with target [43]. It has been found that the heterotrimeric complex between integrin-linked kinase (ILK), PINCH, and parvin is an essential platform for signaling and serve as a convergence point for integrin and growth-factor signaling and regulating cell adhesion, spreading, and migration. The molecular basis of ILK-PINCH interactions was revealed by a crystal structure of the ILK ankyrin repeat domain bound to the PINCH1 LIM1 domain and providing a structural description of this region of ILK. Five ankyrin repeats in ILK have been identified by this structure. The data provides an atomic-resolution description of a key interactions in the ILK–PINCH–parvin scaffolding complex [44]. The initial structure of the ILK Ankyrin repeat domain bound to the PINCH1 LIM1 domain complex was taken from the x-ray crystallography structure with PDB ID of 3F6Q [44]. MD simulations at the molecular mechanics level were employed using ff12SB force field as implemented in the AMBER12 suite of program. The ankyrin-kinase complex were solvated in a cubic box of TIP3P water extending at 10Å in each direction from the solute with 9 Na+ ions added so as to neutralize counterions and the cut-off distance was kept to 20Å to compute the non-bonded interactions. The protein-protein complex was simulated using PMEMD.CUDA from AMBER12 [45] on graphical processors (GPUs) Quadro 2000D produced by NVIDIA which speed up the simulation wall time required to obtain the trajectory files from each simulation. All simulations were performed under periodic boundary conditions, and long-range electrostatics were treated by using the particle-mesh-Ewald method. Initially, the temperature of each system was increased gradually from 0 to 310 K over a period of 60 ps of NVT dynamics. This was followed by 300 ps of NPT equilibration at 310.15 K and 1 atm pressure and then 10,000 ps of NPT-MD simulation was performed for the collection of properties. The structural properties and intermolecular interactions of the ankyrin-kinase and PINCH1-LIM1 were analyzed from the MD trajectories of 10 ns . The binding free energy of the complex was calculated based on the MM-PBSA/GBSA protocol. In this study, the binding free energy of each system was calculated from 6-10 ns of the trajectories. 500 snapshots have been taken into the binding free energy and decomposed binding free energy calculations. The interaction energy profiles of ILK-AR and PINCH1-LIM1 were generated by decomposing the total binding free energies into residue-residue interaction pairs by the MM-GBSA decomposition process in the MM-GBSA program of AMBER12. The values of the interior and exterior dielectric constants in MM-GBSA were set to 1 and 80, respectively. The exploration of hot spots on ILK-AR and assessment of effects of amino acid residues on the binding affinity of Ankyrin and PINCH1 in aqueous solution were discussed after long time scale molecular dynamics. Several important residues with energy < -2 kcal/mol on AR-ILK have been deduced to be critical in binding activity as described in Figure 3.



Amino-Acid Residues	Decomposed Energy (Kcal/mol)	Amino-Acid Residues	Decomposed Energy (Kcal/mol)
HIS33	-2.16	HIS99	-0.97
PHE35	-1.4	ASN101	-0.7
ARG43	-7.21	HIS105	-0.59
ARG65	-1.69	TYR106	-1.2
GLY66	-0.78	PHE109	-6.86
ASP68	1.6	TRY110	-4.11
SER76	- 3.74	CYS129	0.01
HIS77	-1.01	ASN130	-0.02
ASN97	-0.47	TYR132	-1.32
GLU98	0.52	GLU134	0.87

Figure 3: The MM-PBSA/GBSA calculation (left) and the decomposition energy (kcal mol^{-1}) versus significant amino acids residues of AR ILK in binding with PINCH during 6-10 ns simulations (right).

In another investigation in comparing the binding activity between the experimental and theoretical value was carried out with the aid of high-speed GPU molecular dynamics simulations, in designing ankyrin that binds specifically to ERK2 [46]. It is well known that, proteins undergo posttranslational modifications (PTMs) which play a crucial role in signal propagation and regulation of their functions. Up to 30% of all human proteins may be modified by kinase activity, and kinases are known to regulate the majority of cellular pathways, particularly those involved in signal transduction [47]. Kinase enzyme modifies other proteins by chemically adding phosphate groups to them [48]. Out of the many families of protein kinase, MAPKs are one of the most widely studied classes of signalling proteins [49], moreover, their own activity is controlled by a specific phosphorylation event. Extracellular signal-regulated kinase 2 (ERK2), a member of MAPK family; exists in two forms viz. Phosphorylated (active) and the non-phosphorylated (inactive) Figure 4. ERK2, undergoes phosphorylation to regulate several physiological and pathological phenomena, including inflammation, apoptotic cell death oncogenic transformation, tumor cell invasion and metastasis [50]. Overexpression of kinase associated phosphatases are implicated in many different cancers [51-53]. Hence, it is much desired to get binders for inhibition of overexpressed kinases for the control of cancer. Earlier most of the PTMs were studied using antibodies but owing to their improved stability and better binding, DARPins (Designed Ankyrin Repeat proteins) are replacing the conventional monoclonal antibodies [54]. In an attempt to explore the binding interactions of Ankyrins with the ERK2 forms, several DARPins from synthetic library were screened for their binding with ERK2, out of which DARPin E40 was found to have a good binding with ERK2 while pE59 was inactive towards it. The two DARPins differ in one repeat shown as highlighted residues (green) in figure 4. E40 consists 3 repeats while pE59 consists 2 repeats. These DARPins showed no binding with any other kinase tested [55].

```
E40     MRGSHHHHHHGSDLGKKLLEAARAGQDDEVRILMANGADVNAHDDQGSTPLHLAAWIGHP 60
pE59    MRGSHHHHHHGSDLGKKLLEAARAGQDDEVRILMANGADVNALDE--------------- 45
        **************************************** *:

E40     EIVEVLLKHGADVNARDTDGWTPLHLAADNGHLEIVEVLLKYGADVNAQDAYGLTPLHLA 120
pE59    ------------------DGLTPLHLAAQLGHLEIVEVLLKYGADVNAEDNFGITPLHLA 87
                         ** *******: ******************:* :*:******

E40     ADRGHLEIVEVLLKHGADVNAQDKFGKTAFDISIDNGNEDLAEILQKLN 169
pE59    AIRGHLEIVEVLLKHGADVNAQDKFGKTAFDISIDNGNEDLAEILQKLN 136
        * ***********************************************
```

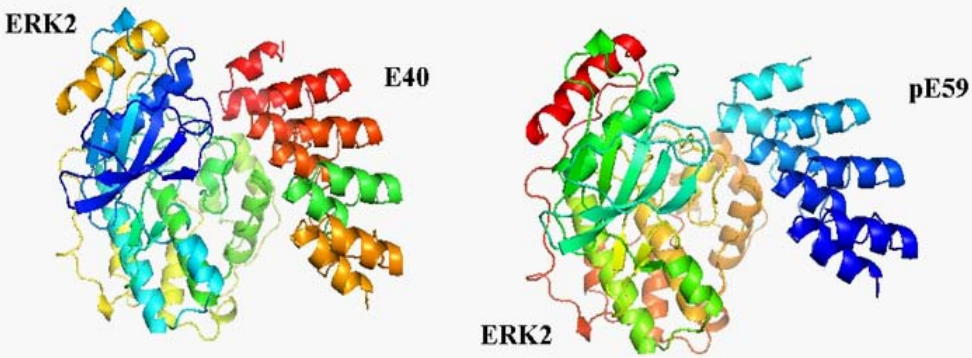


Figure 4: Sequence Alignment of DARPins E40 and pE59 by ClustalW, showing similarity. Structures of E40/ERK2 and pE59/ERK2. Residues different in DARPins, E40 are shown in green.

In order to find a theoretical justification of the above observations, crystal structure of E40 and pE59 in complex with ERK2, were studied for their interaction. Starting structure of the complex was taken from the PDB database (PDB ID 3ZU7). We have applied similar MD simulations protocol on the protein-protein system. In this study, the binding free energy of each system was calculated from 6-10 ns of the trajectories. The values of the interior and exterior dielectric constants in MM-GBSA were set to 1 and 80, respectively. Two DARPin-kinase complexes were studied in order to find out the binding interactions. Binding energy data (GBTOT) indicates, good binding affinity between E40 and ERK2 while that is not found between pE59 and ERK2. The theoretical results expressed as Kd (Dissociation constant) correlates well with the experimental findings (GBTOT) as tabulated in Table 2. This finding leads to results where theoretical calculations produced reliable binding affinity prediction of DARPin-kinase. It was concluded that DARPin E40 has a better binding towards kinase, ERK2 while pE59 showed poor activity. These observations suggest that E40 could be a promising binder for ERK2.

Table 2: Comparison of the experimental and theoretical binding affinity of E40/ERK2 and pE59/ERK2.

DARPins/Kinase	Methods	
	Experimental (KD)	Theoretical GB_{TOT} (Kcal/mol)
E40/ERK2	6.6×10^{-9}	-51.98 ± 6.64
pE59/ERK2	$>8.7 \times 10^{-6}$	-21.73 ± 4.54

Performing computational investigations allow us to view the interactions established between the proteins at molecular level, which is not possible to be observed in laboratory. The predictive power of computational approach has been improved and simulation results found are as close as the reported laboratory results. Therefore, computational tool is not only producing merely a prediction but also explanation of the mechanism involved in a particular system.

ACKNOWLEDGEMENT

This research is partly funded with the support of University Malaya Research Grant under This research is partly funded with the support of University Malaya Research Grant under Frontier Science Research Cluster (UMRG-Project no. RP020C-14AFR and RP027B-15AFR) and Computation and Informatics (C+i) Research Cluster/High Performance Scientific Computing Program (UMRG Project no. RP001C-13ICT).

APPENDIX A: EXAMPLE ON GPU AND SOFTWARE INSTALLATION FOR MOLECULAR DYNAMICS SIMULATION WITH AMBER

The example here is for Ubuntu 12.04, NVIDIA-Linux-x86_64-310.32, CUDA 5.0, and AMBER12.

1. Install Nvidia driver
2. Install CUDA programing
3. Install AMBER12 and Ambertools12

Install NVIDIA driver

- Get the driver for your GPU from the following web

  http://www.nvidia.com/Download/index.aspx?lang=en-us

- Check your GPU spect

  $sudo lshw -short

  or $lspci -v

- Install driver in /home/username

  $chmod +x NVIDIA-Linux-x86_64-310.32.run

  $./NVIDIA-Linux-x86_64-310.32.run

- If error about “must be root”

  $sudo -i

  and reinstall again

  $sudo ./Nvidia.....run

- If error about “exit x server to install nvidia

  $sudo service lightdm stop

  $sudo ./Nvidia.....run

- Restart computer and install nvidia second time

  $sudo shutdown -r now

  $./NVIDIA-Linux-x86_64-310.32.run

  $lsmod|grep nvidia (if install driver correctlly, you should see the nvidia information. If not, reinstall again)

- If the DASH home in ubuntu has gone, do following

1. Switch to a terminal Ctrl+Alt+F1.
2. Login as your username.
3. Install linux headers:

   $sudo apt-get install linux-headers-generic

4. Uninstall nvidia driver - this depends on which version you installed :

   $sudo apt-get remove nvidia-current

   or

   $sudo apt-get remove nvidia-current-updates

   or

   $sudo apt-get remove nvidia-experimental-304

5. Reinstall nvidia driver

   $sudo apt-get install nvidia-current-updates

   When you do this, it must say something like:

   Building initial module for 3.5.0-17-generic Done.

   If it says

Module build for the currently running kernel was skipped since the kernel source for this kernel does not seem to be installed.

then the problem will not be solved. Do not believe the message. It is not asking for linux-source to be install, it does only want the headers but you must install the specific -generic headers for your kernel. Run:

$sudo apt-get install linux-headers-`uname -r`

It will not work with just linux-headers-generic or linux-headers-3.5.0-17 (for example).

$sudo apt-get install linux-headers-3.5.0-17-generic (find out your kernel version by typing in a terminal uname -r)

6. If it successfully installs, restart the computer :

$sudo shutdown -r now

Install CUDA 5.0 on Ubuntu12.04

- Download CUDA 5.0 from https://developer.nvidia.com/cuda-downloads

$chmod +x cuda5.0....._linux_64_ubuntu11.04.run

$sudo ./ cuda5.0....._9_linux_64_ubuntu11.04.run

#Enter until the end of the file 100% and type accept, install in the default directly.

Do not install the driver . The message:.....driver: no,directory: enter ; cuda toolkit: yes, directory enter ; cuda sample: yes, directory: enter

select driver install : no

cuda install : yes and etc : yes

- Fix .bashrc file as following

$cat $PATH

$ export PATH=/usr/local/cuda/bin:$PATH

#for 64-bit machines:

$export LD_LIBRARY_PATH=/usr/local/cuda/lib64:/usr/local/cuda/lib:$LD_LIBRARY_PATH

#for 32-bit machines:

$export LD_LIBRARY_PATH=/usr/local/cuda/lib:$LD_LIBRARY_PATH

Now we want to test did the toolkit install properly:

$nvcc –version

you should see this

nvcc: NVIDIA (R) Cuda compiler driver

Copyright (c) 2005-2012 NVIDIA Corporation

Built on Fri_Sep_21_17:28:58_PDT_2012

Cuda compilation tools, release 5.0, V0.2.1221

#in the .bashrc file you should see

AMBERHOME=/home/username/amber12

export AMBERHOME

export LD_LIBRARY_PATH=$LD_LIBRARY_PATH\:$AMBERHOME/lib

CUDA_HOME=/usr/local/cuda-5.0

export CUDA_HOME

export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:$CUDA_HOME/lib:$CUDA_HOME/lib64

$source .bashrc

Install AMBER12 and Ambertools12

- Extract amber12 and ambertools12

  $pwd

  /home/username

  $tar jxvf AmberTools12.tar.bz2

  $tar jxvf Amber12.tar.bz2

  $cd amber12

  $export AMBERHOME=`pwd`

  $./patch_amber.py --update-tree

  $./configure gnu

  # If there were any updates, see the section above about applying

  # updates to make sure you apply all of them.

  $make install

- Install with single GPU

  $cd $AMBERHOME

$make clean

$./configure -cuda gnu

$make install

You should see the dialogue similar to the one in the video appear. Next we want to install & build the SDK sample files:

$chmod +x cuda-samples_5.0.7_linux.run

$./cuda-samples_5.0.7_linux.run

$cd ~/NVIDIA_CUDA_Samples/

$make

$./C/2_Graphics/volumeRender/volumeRender

More Information: http://ambermd.org/gpus/

- Checking gpu run:

  $ nvidia-smi

  $ nvidia-smi -pm 1

  $ nvidia-smi -c 3

- Running GPU on specific card:

  $ export CUDA_VISIBLE_DEVICES="0"

  $ export CUDA_VISIBLE_DEVICES="1"

References

1. McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. Nature. 1977; 267: 585-590.
2. Duan Y, Kollmann PA. Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. Science. 1998; 282: 740-744.
3. Yeh IC, Hummer G. Peptide Loop-Closure Kinetics from Microsecond Molecular Dynamics Simulations in Explicit Solvent. J. Am. Chem. Soc. 2002; 124: 6563-6568.
4. Klepeis JL, Lindorff-Larsen K, Dror RO, Shaw DE. Long-timescale molecular dynamics simulations of protein structure and function. Curr Opin Struct Biol. 2009; 19: 120-127.
5. Sanbonmatsu KY, Joseph S, Tung CS. Simulating movement of tRNA into the ribosome during decoding. Proc Natl Acad Sci U S A. 2005; 102: 15854-15859.
6. Freddolino PL, Arkhipov AS, Larson SB, McPherson A, Schulten K. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. Structure. 2006; 14: 437-449.
7. Bakker AF, Gilmer GH, Grabow MH, Thompson K. A Special Purpose Computer for Molecular Dynamics Calculations. J. Comput. Phys. 1990; 90: 313-335.
8. Fine R, Dimmler G, Levinthal C. FASTRUN: a special purpose, hardwired computer for molecular simulation. Proteins. 1991; 11: 242-253.
9. Susukita R, Ebisuzaki T, Elmegreen BG, Furusawa H, Kato K, et al. Hardware Accelerator for molecular dynamics: MDGRAPE-2. Comput. Phys. Commun. 2003; 155: 115-131.

10. Shaw DE, Deneroff MM , Dror RO, Kuskin JS, Larson RH, et al. Anton, A Special-Purpose Machine for Molecular Dynamics Simulation. Commun. ACM. 2008; 51: 91-97.

11. Gotz AW, Wolfle TM, Walker RC. Quantum Chemistry on Graphics Processing Units. In: Simmerling C, editor. Annual Reports in Computational Chemistry. Amsterdam: Elsevier. 2010; 21-35.

12. Xu D, Williamson MJ, Walker RC. Advancements in Molecular Dynamics Simulations of Biomolecules on Graphical Processing Units. In: Simmerling C, editor. Annual Reports in Computational Chemistry. Amsterdam: Elsevier. 2010; 21-35.

13. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J Chem Theory Comput. 2013; 9: 3878-3888.

14. Phillips JC, Stone JE, Schulten K, editors. Adapting a message driven parallel application to GPU-accelerated clusters. SC 08: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing; 2008 November 15-21; NJ, USA. Piscataway: IEEE Press. 2008; 1-9.

15. Friedrichs MS, Eastman P, Vaidyanathan V, Houston M, Legrand S. Accelerating molecular dynamic simulation on graphics processing units. J Comput Chem. 2009; 30: 864-872.

16. Eastman P, Pande VS. Efficient nonbonded interactions for molecular dynamics on a graphics processing unit. J Comput Chem. 2010; 31: 1268-1272.

17. Harvey MJ, Giupponi G, Fabritiis GD. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. J Chem Theory Comput. 2009; 5: 1632-1639.

18. Harvey MJ, De Fabritiis G. An Implementation of the Smooth Particle Mesh Ewald Method on GPU Hardware. J Chem Theory Comput. 2009; 5: 2371-2377.

19. Hampton SS, Agarwal PK, Alam SR, Crozier PS. Towards microsecond biological molecular dynamics simulations on hybrid processors. Proceedings of the International Conference on High Performance Computing and Simulation. Jun 28-29, 2010; 98-107.

20. Brown WM, Wang P, Plimpton SJ, Tharrington AN. Implementing Molecular Dynamics on Hybrid High Performance Computers-Short Range Forces. Comput. Phys. Commun. 2011; 182: 898-911.

21. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J. Am. Chem. Soc. 1995; 117: 5179-5197.

22. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B. 1998; 102: 3586-3616.

23. Case DA, et al. AMBER 11; SanFrancisco: University of California. 2010.

24. Case DA, Cheatham TE 3rd, Darden T, Gohlke H, Luo R. The Amber biomolecular simulation programs. J Comput Chem. 2005; 26: 1668-1688.

25. Le Grand S, Goetz AW, Walker RC. SPFP: Speed without compromise - a mixed precision model for GPU accelerated molecular dynamics simulations. Comp. Phys. Comm. 2013; 184: 374-380.

26. Goetz AW, Williamson MJ, Xu D, Poole D, Le Grand S. Routine microsecond molecular dynamics simulations with AMBER - Part I: Generalized Born. J. Chem. Theory Comput. 2012; 8: 1542-1555.

27. Salomon-Ferrer R, Goetz AW, Poole D, Le Grand S, Walker RC. Routine microsecond molecular dynamics simulations with AMBER - Part II: Particle Mesh Ewald. J. Chem. Theory Comput. 2013; 9: 3878-3888.

28. Nvidia: NVIDIA CUDA.

29. Liu W, Schmidt B, Voss G, Muller-Wittig W. Accelerating Molecular Dynamics Simulations Using Graphics Processing Units with CUDA. Comp. Phys. Comm. 2008; 179: 634-641.

30. Woods CJ, Malaisree M, Long B, Mclntosh-Smith S, Mulholland AJ. Computational Assay of H7N9 Influenza Neuraminidase Reveals R292K Mutation Reduces Drug Binding Affinity. Scientific Reports. 2013; 3: 3561.

31. Kodchakorn K, Dokmaisrijan S, Chong WL, Payaka A, Wisitponchai T, et al. GPU Accelerated Molecular Dynamics Simulations for Protein-protein Interaction of Ankyrin Complex. Integrated Ferroelectrics. 2014; 156: 137-146.

32. Hou TJ, Guo SL, Xu XJ. Predictions of binding of a diverse set of ligands to gelatinase-A by a combination of molecular dynamics and continuum solvent models. J. Phys. Chem. B. 2002; 106: 5527-5535.

33. Huo S, Wang J, Cieplak P, Kollman PA, Kuntz IDS. Molecular dynamics and free energy analyses of cathepsin D-inhibitor interactions: insight into structure-based ligand design. J. Med. Chem. 2002; 45: 1412-1419.

34. Wang W, Kollman PA. Computational study of protein specificity: the molecular basis of HIV-1 protease drug resistance. Proc Natl Acad Sci U S A. 2001; 98: 14937-14942.

35. Lepsik M, Kriz Z, Havlas Z. Efficiency of a second-generation HIV-1 protease inhibitor studied by molecular dynamics and absolute binding free energy calculations. Proteins. Struct. Funct. Bioinf. 2004; 57: 279-293.

36. Stoica I, Sadiq SK, Coveney PV. Rapid and accurate prediction of binding free energies for saquinavir-bound HIV-1 proteases. J Am Chem Soc. 2008; 130: 2639-2648.

37. Wang W, Kollman PA. Free energy calculations on dimer stability of the HIV protease using molecular dynamics and a continuum solvent model. J Mol Biol. 2000; 303: 567-582.

38. Beveridge DL, DiCapua FM. Free energy via molecular simulation: applications to chemical and biomolecular systems. Annu Rev Biophys Biophys Chem. 1989; 18: 431-492.

39. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009; 30: 2785-2791.

40. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. Proteins. 2003; 52: 80-87.

41. Chen HM, Liu BF, Huang HL, Hwang SF, Ho SY. SODOCK: swarm optimization for highly flexible protein-ligand docking. J Comput Chem. 2007; 28: 612-623.

42. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J. Comput. Chem. 2003; 24: 1999-2012.

43. Gautam V, Wei Lim Chong WL, Wisitponchai T, Nimmanpipug P, Zain SM, et al. GPU-enabled molecular dynamics simulations of ankyrin kinase complex. AIP Conference Proceedings. 2014; 1621: 112.

44. Chiswell BP, Zhang R, Murphy JW, Boggon TJ, Calderwood DA. The structural basis of integrin-linked kinase-PINCH interactions. Proc Natl Acad Sci U S A. 2008; 105: 20677-20682.

45. Case DA, Darden TA, Cheatham TE, Simmerling CL, Wang J, et al. AMBER 12. San Francisco: University of California. 2012.

46. Gautam V, Zain SM, Rahman NA, Lee VS. GPU accelerated molecular dynamics simulations in the designing of Ankyrin as specific binders of ERK2. Poster Presented at: 8th Conference of Asian Consortium on Computational Materials Science. Taipei, Taiwan. June 16-18, 2015.

47. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The Protein kinase complement of the human genome. Science. 2002; 298: 1912-1934.

48. Salway JG. Metabolism at a Glance, 2nd edition. Oxford: Blackwell Science Ltd. 1999.

49. Chen Z, Gibson TB, Robinson F, Silvestro L, Pearson G. MAP kinases. Chem Rev. 2001; 101: 2449-2476.

50. Lawrence MC, Jivan A, Shao C, Duan L, Goad D. The roles of MAPKs in disease. Cell Res. 2008; 18: 436-442.

51. Paul MK, Mukhopadhyay AK. Tyrosine kinase - Role and significance in Cancer. Int J Med Sci. 2004; 1: 101-115.

52. Lee SW, Reimer CL, Fang L, Iruela-Arispe ML, Aaronson SA. Overexpression of kinase-associated phosphatase (KAP) in breast and prostate cancer and inhibition of the transformed phenotype by antisense KAP expression. Mol Cell Biol. 2000; 20: 1723-1732.

53. Nemoto T, Ohashi K, Akashi T, Johnson JD, Hirokawa K. Overexpression of protein tyrosine kinases in human esophageal cancer. Pathobiology. 1997; 65: 195-203.

54. Binz HK, Amstutz P, Kohl A, Stumpp MT, Briand C. High-affinity binders selected from designed ankyrin repeat protein libraries. Nat Biotechnol. 2004; 22: 575-582.

55. Kummer L, Parizek P, Rube P, Millgramm B, Prinz A. Structural and functional analysis of phosphorylation-specific binders of the kinase ERK from designed ankyrin repeat protein libraries. Proc Natl Acad Sci U S A. 2012; 109: E2248-2257.

Austin Publishing Group

Knowledge Formalization and High-Throughput Data Visualization Using Signaling Network Maps

Kondratova M[1,2,3,4], Barillot E[1,2,3,4], Zinovyev A[1,2,3,4] and Kuperstein I[1,2,3,4]*

[1]Institut Curie, France

[2]INSERM, U900, France

[3]Mines Paris Tech, France

[4]PSL Research University

***Corresponding author:** Kuperstein I, Institut Curie, 26 rue d'Ulm, F-75248 Paris, France, Tel: +33 (0) 1 56 24 69 87; Email: inna.kuperstein@curie.fr

Published Date: December 01, 2016

ABSTRACT

Graphical representation of molecular biology knowledge in the form of interactive diagrams became widely used in molecular and computational biology. It enables the scientific community to exchange and discuss information on cellular processes described in numerous scientific publications and to interpret high-throughput data. Constructing a signaling network map is a laborious process, therefore application of consistent procedures for representation of molecular processes and accurately organized annotation is essential for generation of a high-quality signaling network map that can be used by various computational tools. We summarize here the major aims and challenges of assembling information in a form of comprehensive maps of molecular interactions and suggest an optimized workflow. We share our experience gained while creating a biological network resource Atlas of Cancer Signaling Network (ACSN) that was successfully applied in several studies. We explain the map construction process. Then

we address the problem of user interaction with large signaling maps and suggest to facilitate navigation by hierarchical organization of map structure and by application of semantic zooming principles. In addition, we describe a computational technology using Google Maps API to explore signaling networks in the manner similar to global geographical maps and provide the outline for preparing a biological network for this type of navigation. Nowadays the most demanded application of signaling maps is integration and functional interpretation of high-throughput data. We demonstrate several examples of cancer data visualization in the context of comprehensive signaling network maps.

Keywords: Knowledge formalization; Network map construction: Data model; Network map curation; Network map navigation; Data visualization

INTRODUCTION

Geographic maps, diagrams and flowcharts are examples of graphics that contain lots of information that is intuitive and relativity easy to grasp. Similarly, graphical representation of biological knowledge may allow to show complex processes of living cell in a visual and insightful way. Applying the principle of knowledge representation in the form of a diagram can help for systematic representation and formalization of molecular information distributed in thousands of papers. An additional advantage of representing the biological processes in a graphical form is catching collectively multiple cross-talks between components of different cell signaling processes. This allows understanding the global picture and connectivity between processes that is very difficult to keep in mind just from reading multiple scientific papers. Once the processes are depicted together as diagrams, the relationship between molecular circuits in cells can be appreciated, which makes signaling network maps also didactic tools [1].

Representation of biological processes as comprehensive signaling maps has three objectives: (i) generating a resource containing formalized summary of biochemical mechanisms as elucidated by many research groups, (ii) providing a platform for sharing information and discussing the mechanisms of biological processes, (iii) creating an analytical tool for high-throughput data integration and analysis. To achieve these goals, signaling map construction should become an accessible procedure that can be completed in a reasonable time. During last decade, the molecular biology community came up with several solutions of biological knowledge formalization that we shortly describe in this manuscript. We contribute to this global aim by describing the main principles of our established workflow for manual map construction. In addition, we demonstrate the ways of biological network map navigation facilitated by Google Maps technology and suggest our tools for performing data analysis and visualization on top of the signaling network maps.

There are four main modes of cell processes representation. Each mode uses a different logic to depict molecular information: (i) interaction diagrams showing simple binary relations between molecular entities;(ii) activity-flow or regulatory networks representing the flow of information or influences of one entity on another; (iii) entity relationship diagrams depicting relations in

which a given entity participates regardless of temporal aspects and (iv) process description diagrams (known in chemical kinetics as bi-partite reaction network graphs) where sequential order of biochemical interactions is explicitly represented [2]. Using the aforementioned modes of cell processes representation, several pathway databases have emerged [3]. They serve as biological knowledge information resource and as computational analytical tools for systems-based interpretation of data. Significant number of pathway databases have been developed in the private domain, but the majority of them are free open sources (Table 1).

Table 1: Pathways databases and network resources; navigation tools and high-throughput data visualization support.

Signaling pathways and networks resources			
Name	*Website*	*Description*	*Reference*
STRING	http://string-db.org	Integrated protein-protein interaction database	[39]
BioGRID	http://thebiogrid.org	Integrated protein-protein and genetic interaction database	[40]
MINT	http://mint.bio.uniroma2.it/mint/Welcome.do	Carefully curated PPI resource	[41]
PathwayCommons	http://www.pathwaycommons.org	Biological pathways resource collected from public pathway databases	[42]
TRANSPATH	http://www.biobase-international.com	Database of mammalian signal transduction and metabolic pathways	[43]
ConsensusPathDB	http://consensuspathdb.org	Integrated resource of interaction networks and pathways	[8]
Panther	http://pantherdb.org	Collection of biological pathways and data visualization and analysis tools	[6]
Spike	http://www.cs.tau.ac.il/~spike	Collection of curated, peer reviewed pathways and data visualization tools	[7]
WikiPathways	http://www.wikipathways.org	Collection of community curated signalling pathways	[25]
PID-NCI	http://pid.nci.nih.gov	Curated collection of information about biomolecular interactions and signalling pathways	
KEGG Pathway	http://www.genome.jp/kegg/pathway	Collection of manually drawn pathway maps visualization tool	[4]
Reactome	http://www.reactome.org	Collection of curated, peer reviewed pathways and data visualization/analysis tools	[5]
ACSN	http://acsn.curie.fr	Collection of curated, peer reviewed, interconnected cancer-related signaling networks and data visualization/analysis tools	[9]
Tools for network construction, visualization, navigation, and commenting			
Name	*Website*	*Description*	*Reference*
CellDesigner	http://www.celldesigner.org	Structured diagram editor for drawing gene-regulatory and biochemical networks	[14]
SBGN-ED	http://vanted.ipk-gatersleben.de/addons/sbgn-ed	VANTED add-on for create and edit three types of SBGN maps	[12]
CellPublisher	http://cellpublisher.gobics.de	KEGG database-associated tool for data visualization and analysis in the context of pathway maps	[15]
Cytoscape / BiNoM	http://www.cytoscape.org http://binom.curie.fr	Software platform for manipulation of biological networks represented in standard systems biology formats	[44][24]
Payao	http://payao.oist.jp:8080/payaologue/index.html	Network curation tool for simultaneous map commenting using tag system	[30]
Pathway Projector	http://www.g-language.org/PathwayProjector	Web-based zoomable pathway browser using Google Maps API	[16]
NaviCell	http://navicell.curie.fr	Web-based tool for heterogeneous data visualization and analysis in the context of signalling networks	[17][18]

yEd graph editor	http://www.yworks.com/en/products/yfiles/yed/	Application for generate high-quality diagrams construction	http://link.springer.com/chapter/10.1007/978-3-642-18638-7_8
VisANT	http://visant.bu.edu/	Tool for visual analyses of metabolic networks in cells and ecosystems	[13]
Pathway Map Creator	http://lifesciences.thomsonreuters.com/m/pdf/PathwayMapCreator-cfs-en.pdf	Tool for editing and analysis of canonical pathways maps	
Tools for visualisation of high-throughput data in the context of signalling networks			
Name	*Website*	*Description*	*Reference*
ReactomeFiViz	http://wiki.reactome.org/index.php/Reactome_FI_Cytoscape_Plugin	Cytoscape plugin for data integration into signalling networks	[45]
iPAth	http://pathways.embl.de	Web-based tool for data visualization in the context of pathway maps	[46]
Medusa	http://coot.embl.de/medusa	Tool for data visualization in the context of signalling network and network clustering	[47]
NaviCell	http://navicell.curie.fr	Web-based tool for heterogeneous data visualization and analysis in the context of signalling networks	[17][18]
KEGG Mapper	http://www.kegg.jp/kegg/mapper	KEGG database-associated tool for data visualization and analysis in the context of pathway maps	[4]

The current status of representing cellular mechanisms in such databases as KEGG [4], Reactome [5], Panther [6], Spike [7], Consensus Path DB [8] and others, mainly remains at the level of drawing individual signal transduction pathways, that precludes clear representation of cross-regulations between pathways. The alternative solution is creating the seamless map of biological mechanisms covering multiple cell processes at one canvas as it is done in Atlas of Cancer Signaling Network (ACSN) [9] and KEGG metabolic pathway [10]. This "geography-inspired" approach to biological knowledge representation is a very attractive goal. However, achieving this goal is connected with a number of challenges related to creation, maintenance and navigation of the large signaling network maps. This chapter is partially devoted to discussing these challenges and suggesting solutions from our practical long-term experience.

Variety of uncompatible ways are used to represent biological maps in different pathway databases. This creates difficulties in combining complementary maps from multiple resources. With the aim to join the efforts in the field and create a collection of mergeable and exchangeable signaling maps, common rules of map drawing and standard graphical syntax should be developed and consistently applied. The current solution suggested in the field is Systems Biology Graphical Notation (SBGN) syntax, which is compatible with many pathway drawing and analytical tools, allowing to represent not only biochemical processes, but also cell compartments and phenotypes [11]. In such databases as Reactome, Panther pathways diagrams are represented in the SBGN graphical format. In addition, to enable cross-compatibility, several common pathway exchange formats were suggested such as BioPAX, SBML, PSI-MI etc [2].

For creating signaling maps, there exist several free and commercial tools for signaling map diagram construction. These tools use different syntax and also vary in their accessibility for the end users. Examples are SBGN-ED [12], visANT [13], CellDesigner [14] and others (Table 1).

Visualization and exploring biological network diagrams became an important issue, because size and complexity of molecular networks approach the modern geographic maps. Therefore, several tools such as CellPublisher [15], Pathway Projector [16] and NaviCell [17,18] have adopted the logic of navigation from Google Maps technology. They allow uploading networks diagrams, exploring big networks in a user-friendly manner using such Google Maps features as scrolling, zooming, markers and callouts (Table 1).

In this chapter, we provide a workflow where we briefly describe our methodology to meet challenges of map construction, navigation and data integration. We suggest a methodology that is neither unique nor universal, but provides practical solutions for comprehensive maps generation and manipulation. This methodology served for creating the maps for ACSN resource [9]. The approach was also successfully applied in several studies [19,20]. Each step of the workflow starts from the problem statement and description of the principles followed by a solution suggestion demonstrated on a typical example.

We discuss the following topics:

- Defining the aim and the coverage of knowledge on signaling map
- Literature selection and signaling map drawing in CellDesigner tool using SBGN-like syntax [14]
- Preparation of CellDesigner maps in NaviCell format and generation of a NaviCell web-based pages
- Navigation modes using Google Maps-based NaviCell tool [17]
- High-throughput data visualization on top of the signaling maps using NaviCell Web Service module [18]

The entry point to the detailed description of the procedures is provided in the end of the chapter, and available at https://navicell.curie.fr/pages/guide.html.

The suggested workflow for map generation and exploration is schematically depicted in Figure 1.

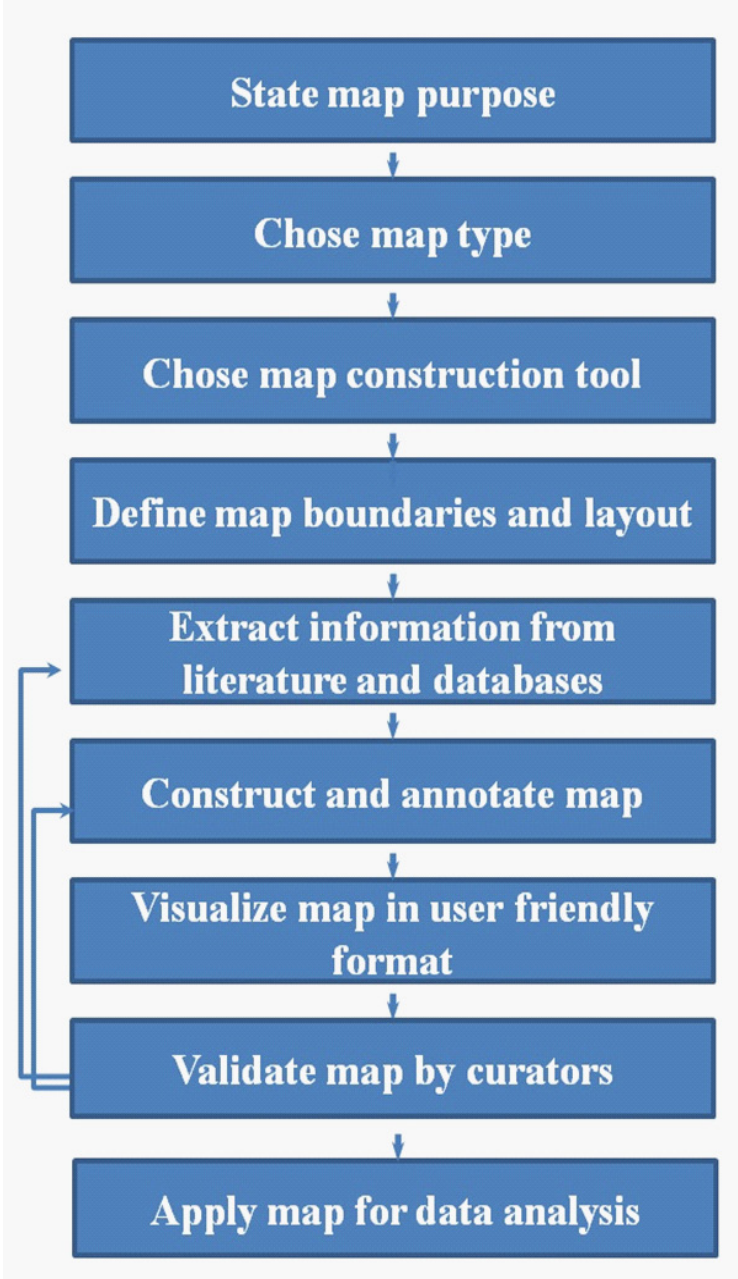


Figure 1: Map construction workflow scheme.

GENERAL PRINCIPLES AND WORKFLOW FOR MAP CONSTRUCTION

Work Organization

Signaling map construction requires an overview of broad scientific literature and a very minute work for correct representation of molecular processes in great details. Given large work content, map construction can be approached as a collective effort. To achieve efficient and coordinated team work, several important decisions should be made prior to the map construction by answering to the following questions: (i) What is the purpose of map construction? (ii) What map type is suitable for proper representation of the knowledge? (iii) What is the appropriate tool to build the map? (iv) What processes to include into the map? (iv) How the map will look like? Once the answers to those questions are found and agreement on the approach is achieved, the way of signaling diagram construction should be strictly followed to ensure generation of a homogeneous and accurate map. An additional important step before constructing a map is consulting similar efforts in the field and clear understanding of added value of a new map.

Map Purpose and Type

As an example, we use the DNA repair map from ACSN resource available at https://acsn.curie.fr/navicell/maps/dnarepair/master/index.html. With purpose to understand how different types of DNA damage are repaired in the cell and how the coordination between various DNA repair mechanisms and the cell cycle takes place, we have decided to construct a comprehensive map of DNA repair and cell cycle signaling. We aim to use this map for detecting the modes of DNA repair machinery rewired in different pathological situations such as cancer or under genotoxic stress. The mechanisms of DNA repair are well studied and information on involved molecules and regulation circuits is available. Therefore, to preserve and depict accurately the processes in whole complexity, we have chosen the process description diagram type were the biochemical reactions can be explicitly depicted.

Map construction tool, graphical standard and data model

Signaling processes are represented as biochemical reactions in CellDesigner diagram editor. CellDesigner uses standard Systems Biology Graphical Notation (SBGN) syntax [11] and is based on Systems Biology Markup Language (SBML) for further computational modeling of the map [14]. The data model, that is applied for our example is schematically depicted in Figure 2. This data model includes such molecular entities as proteins, genes, RNAs, antisense RNAs, simple molecules, ions, drugs, phenotypes, complexes. Biochemical reactions connect reactants to products and various types of reaction regulators are also depicted. Edges on the map represent biochemical reactions or reaction regulations including post-translational modifications, translation, transcription, complex formation or dissociation, transport, degradation, etc. Reaction regulations are catalysis, inhibition, modulation, trigger, and physical stimulation. It is also possible to depict cell compartments such as cytoplasm, nucleus, mitochondria, etc. See http://celldesigner.org/documents.html for CellDesigner tool guide and http://www.sbgn.org/Main_Page for SBGN syntax explanation.

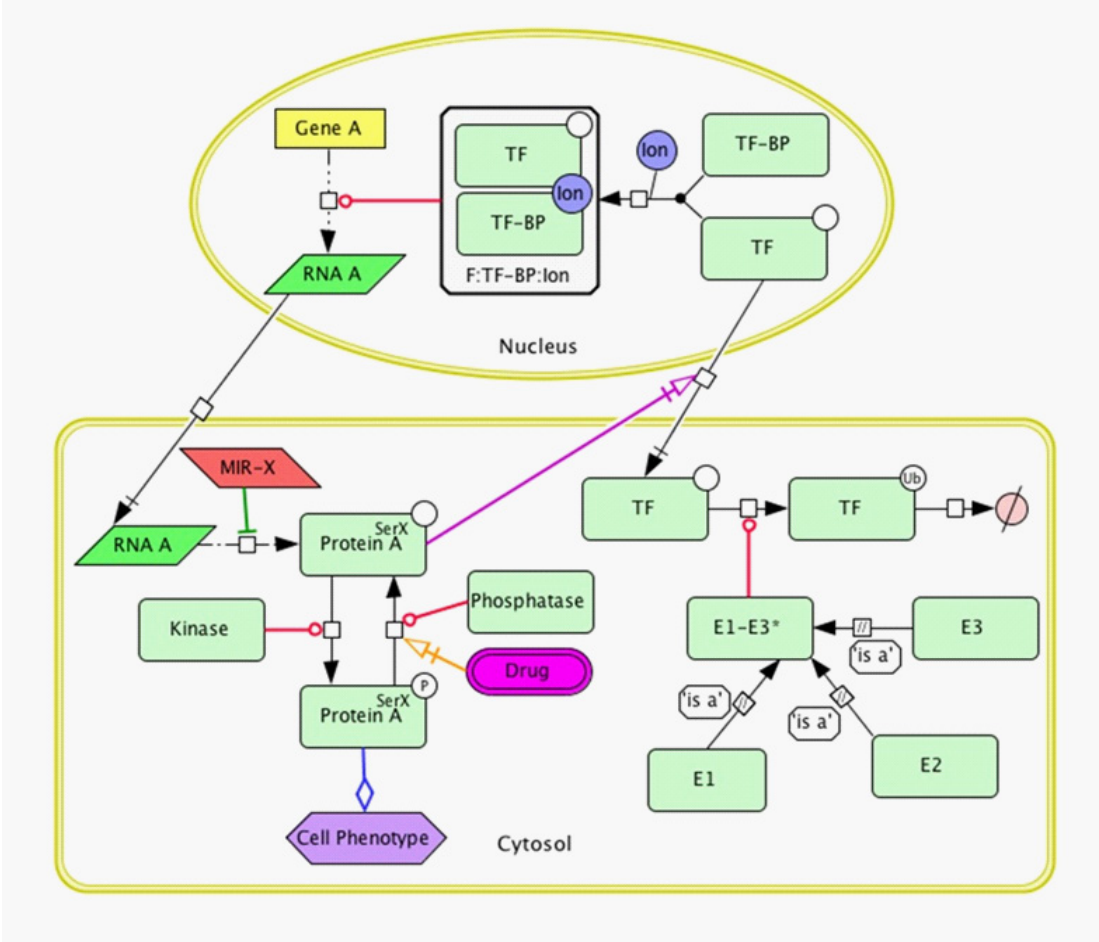


Figure 2: Data model for process description diagram drawing.

MAP BOUNDARIES, LAYOUT AND STRUCTURE

Map Boundaries and Content

Given limitations of graphical tools and difficulties in manipulating large maps, it is recommended to define the boundaries of signaling maps. The most natural way to set map boundaries is to dedicate each map to one biological function (e. g. cell death, DNA replication, immune response) that is a difficult task per se due to 'fuzziness' of borders between processes and overlaps between players and pathways across cell signaling. Therefore, those function-driven maps should be assumed as components of the global atlas where the merging via common players or overlapping parts should be possible due to common standards applied and common identifiers for entities universally used through all maps. The decision about map boundaries highly depends on the opinion of the map creator, thus community-based curation of maps is crucial for making more objective decisions.

In our example, the boundaries of DNA repair map were defined in coordination with several specialists in the corresponding fields and based on the commonly accepted vision of pathways as they are presented in seminal reviews and in well-known databases. Accordingly to the current definition of DNA repair, it is possible to distinguish 10 different modes of repair depending on the type of damage and mechanisms of repair. The DNA damage types are clearly depicted on the map as input initiating DNA repair mechanisms. The ten DNA repair mechanisms with multiple crosstalk [21]; four cell cycle phases and check points including regulatory circuits between cell cycle to DNA repair mechanisms via checkpoints are included into the map [22] (Figure 3).

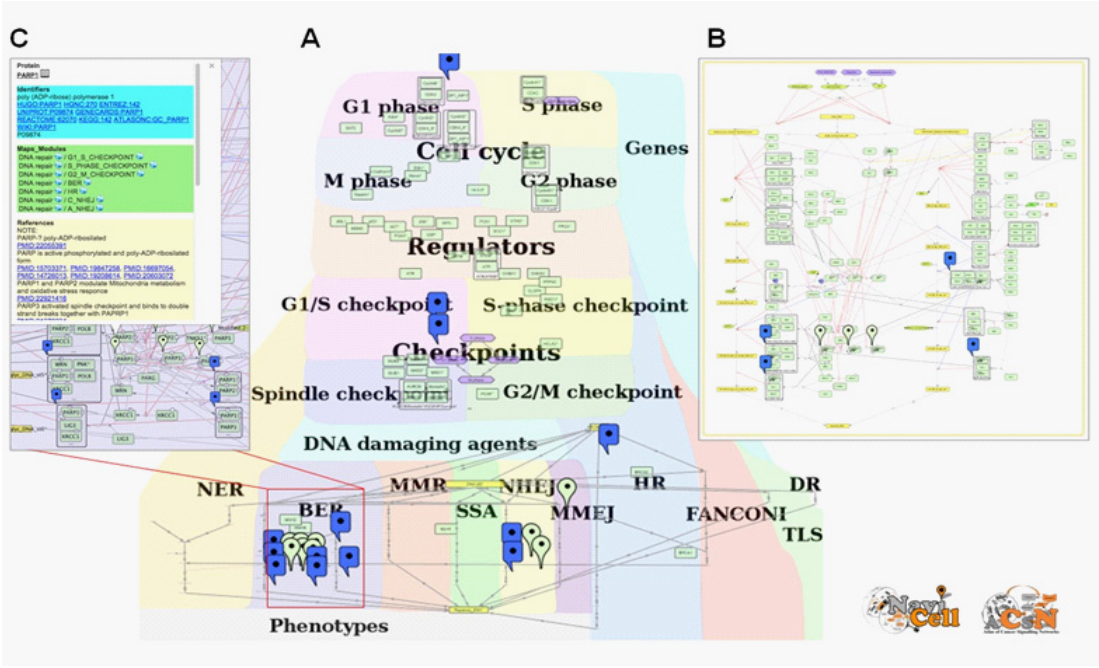


Figure 3: DNA repair map in NaviCell format

(A) Global (top) layout, (B) Individual module layout (BER module of the DNA repair map), (C) Pop-up window with annotation of p53 protein.

Map Layout and Hierarchical Modular Structure

The aim of signaling map construction is not only to summarize molecular mechanisms, but also to allocate processes in a meaningful and biologically relevant way. Careful design of signaling map layout helps for intuitive understanding of 'what is where' and 'what is close and what is distant'. In addition, a bird's eye view on the map can give a general impression about map complexity based on interaction density between players in the network.

The structure of signaling map and its layout are messages by themselves. There are at least three ways of map layout choice: (i) Representing spatial localization of processes in the context of the global cell architecture. Most of the signal transduction maps mimic accepted view of

cell organization, they include cellular compartments and signaling pathways are placed in the corresponding compartments on the map (see examples at https://acsn.curie.fr). (ii) Depicting process propagation in time, for instance, demonstrating propagation of the signaling through four phases of cell cycle (Figure 3A, cell cycle). (iii) Placing processes together according to involvement in a particular biological function. The most representative example is DNA repair pathways, where each pathway is depicting one biological function, these pathways are allocated next to each other, creating together a DNA repair machinery layer (Figure 3A, DNA repair). The combination of layout types at the same map is also possible as in the case of DNA repair map combining the three aforementioned layout types (Figure 3A and https://acsn.curie.fr/navicell/maps/dnarepair/master/index.html).

Type of chosen layout also can guide in separating big maps into sub-maps (modules) and help generating hierarchical modular structure of map. Each such a module can exist as a part of the global map and as an independent map. Exploring these module maps together with the global map can be supported by Google Maps-based map navigation that facilitates understanding of depicted processes (discussed below). Map dimensions and layout should be defined prior to initiating the map drawing. This step is especially crucial in the case of collective map reconstruction approach where final global map is assembled from a number of sub-maps (or module maps).

In our example, the global layout of DNA repair map has been designed to emphasize the cross-regulation between three major blocks of the map: the knowledge of DNA repair machinery and its connection to the cell cycle and to the checkpoints is represented as layers. The upper layer depicts cell cycle, the middle layer represents cell cycle checkpoints coordinating the crosstalk between the cell cycle and the DNA repair machinery which is represented in the lower layer. The DNA map has modular structure composed of 18 functional modules corresponding to ten DNA repair mechanisms, four phases of cell cycle and four checkpoints, all interconnected, with multiple regulatory circuits (Figure 3A).

Typically, in the case of large maps, close up view on individual functional modules within the context of global map is difficult due to the distant location of some players and high density of edges crossing the map in multiple directions. To overcome these constraints, individual module layout can be designed for each modular map that can differ from the original one on the global map and aims to better represent the detailed biochemical reaction flow. An example of a module map with optimized layout for Base Excision Repair (BER) pathway is shown in Figure 3B. For modular map generation instructions see in the Map preparation in NaviCell format procedure (https://navicell.curie.fr/doc/NaviCellMapsPreparationProcedure.pdf).

Similarly, it is possible to generate automatic layouts for module maps or even for maps of individual entity life-cycles with their related edges and reactions. This can be performed in Cytoscape using BiNoM plugin with help of modularization and automatic layout functions [23]. To facilitate integration of separate signaling diagrams, there are at least two methods for map merging: (i) Merge Model plugin in CellDesigner [14] and (ii) BiNoM plugin of Cytoscape which allows to reorganize, dissect and merge disconnected CellDesigner pathway diagrams [23]. See https://binom.curie.fr/docs/BiNoM_Manual_v2.pdf for map merging procedure in BiNoM.

DATA EXTRACTION AND REPRESENTATION

Manual Data Mining

Clear understanding of biological processes and experimental methodologies is essential for constructing a correct and usable signaling map. Different types of molecular interactions can be demonstrated by various experimental methods. Retrieving the correspondence between a method to evidence is needed for correct interpretation of the results described in the papers, followed by suitable graphical representation of the statements. The most common and reliable experimental methods confirming different types of molecular interactions in the cell are summarized in Table 2.

Table 2: Method for studying molecular and genetic interactions.

Interaction	*Method*	*Reference*
Ligand-receptor interactions	Resonance energy transfer [FRET and BRET]	[48][49]
	Flow CytometricAnalysis	[50]
Direct protein-protein interactions	Co-immunoprecipitation [CoIP],	[51][52]
	NMR, X-ray crystallography,	[53]
	GST-pull down assay	[54]
	Tandem affinity purification	[51][49]
	Far Western blotting	[55]
	Phage display	[52]
	Mass-spectrometry	[56]
	Two hybrid assays [yeast and mammalia]	[51]
	Functional mutational analysis	[57]
Direct protein-DNA interaction [transcription regulation and co-regulation]	Chromatin immunoprecipitation [ChIP]	[58][59]
	DNA footprinting,	[58]
	Electrophoretic mobility shift end supershift assays [EMSA]	[58]
	Computational prediction of transcription factors binding sites	[60]
MicroRNA binding	Direct miRNA binding assay,	[61]
	3 'UTR reporter assay,	[61]
	Computational miRNA target prediction	[61]
Regulation of expression [mRNA and protein level]	Reverse transcription polymerase chain reaction [RT-PCR]	[62]
	Reporter assays	[63]
	RNase protection assay	[64]
	Nothern blot	[62]
	Western blot	[65]
	Fluorescence-activated cell sorting [FACS]	[66]
Genetic interactions	Genetic knock-out, knock-down, knock in or overexpression of effector molecules	[67][68]
	Synthetic interaction detection assays	[52][69]

One on the major aims of signaling network maps is to represent biological processes with great precision including post-translational modifications, transport, complex association, degradation etc. Phenotype nodes on the signaling maps normally serve for indication of signaling readouts or cell statuses or biological processes in general. This type of nodes can also serve for schematic representation of observation-type statements when the exact molecular mechanism is still unknown. Some details on cell signaling might be skipped or in contrary, represented rigorously, depending on the purpose of the map drawing and opinion of the map creator. Persisting homogeneity in the presentation of information on the map will ensure correct stepwise appearance of details on different map views (discussed below).

For efficient map construction, we suggest applying systematic literature revision. Hierarchical organization of the map discussed in the previous paragraphs, reflects the principle of literature curation for map generation in our example. To define the map boundaries and content, seminal review papers in the field are used which provide a list of original references. It is also recommended to consult the major pathway databases (Table 1). The canonical pathways retrieved from major reviews and databases reflect the consensus view of the field and serve as a basis for drawing the core processes on the map. Thus, the details can be added to the map to depict information extracted from the recent literature with the requirement that the interactions and processes included into the map are supported by at least two independent investigations. Figure 4 represents an example from the DNA repair map on how the scientific text is translated into the process description diagram. For consistency of text to diagram conversion, we have developed several major rules for standard statements interpretation, summarized in the Map creator guide (https://navicell.curie.fr/doc/NaviCellMapperAdminGuide.pdf) and in the Map preparation in NaviCell format procedure (https://navicell.curie.fr/doc/NaviCellMapsPreparationProcedure.pdf).

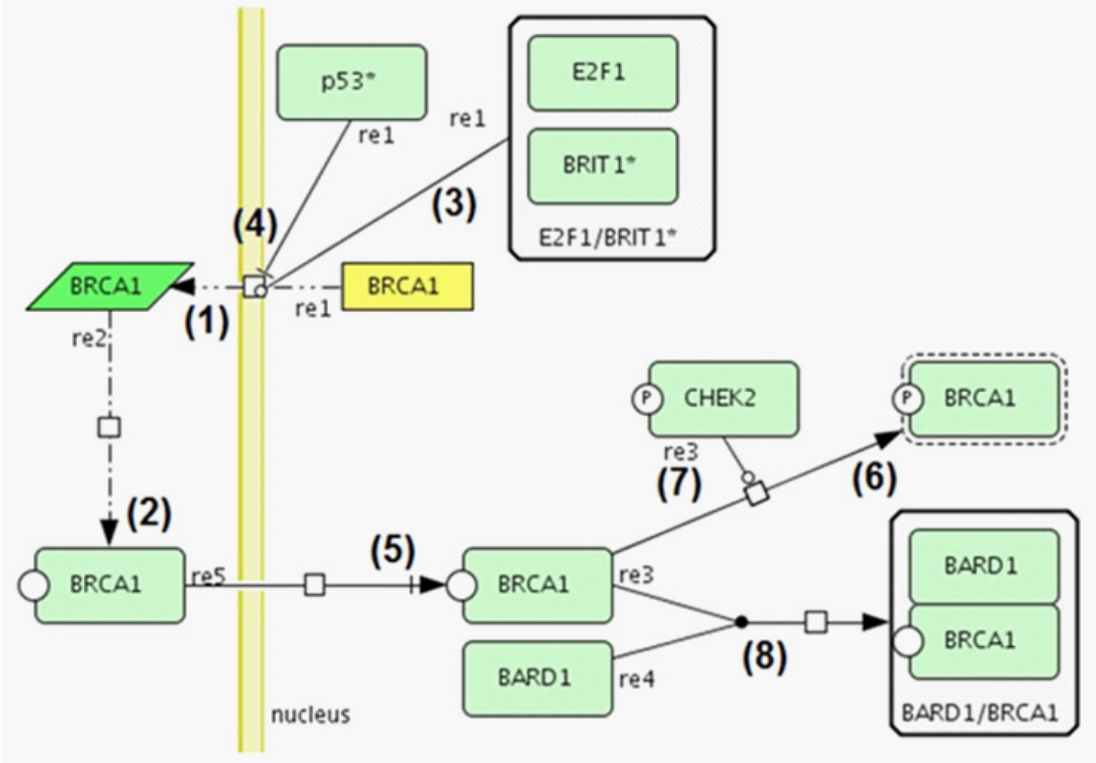


Figure 4: From text to model

Representation of biochemical reactions from the following text from a molecular biology manuscript. Numbers correspond to the reactions in the diagram:*« BRCA1 transcription (1) and translation (2) is positively regulated by E2F1/BRIT1* complex (3) and inhibited by p53 (4). BRCA1 protein is transported into nucleus (5), where CHEK2 kinase activates it by spesific phosphorylation (6) and (7). Additionally, BRCA1 forms a complex with BARD1 (8) and BRCA1 association with BARD1 is essential for the E3 ligase activity of BRCA1»*. References correspondence: reactions 1,2,4 (33); reaction 3 (34); reaction 5 (35); reactions 6,7 (36); reaction 8 (37). Formalized textual description of the diagram, in the BiNoM Reaction Format (BRF) is described in the text.

Processing Formal Statements on Biochemical Reactions

An additional useful method for knowledge to diagram conversion is using a formalized text-based intermediate language. In this method the sentences formulated in the "human" language are first converted into a set of formal statements describing reactions. These statements can be automatically converted into the graphical diagrams, using BiNoM Cytoscape plugin. For example, the set of statements corresponding to the diagram in Figure 4 is:

BRCA1@cytoplasm -/> BRCA1@nucleus

BRCA1@nucleus+BARD1@nucleus -:> BARD1:BRCA1@nucleus

BRCA1-CHEK2|pho@nucleus -> BRCA1|pho|active@nucleus

rBRCA1@cytoplasm -.> BRCA1@cytoplasm

gBRCA1@nucleus-|p53*@nucleus -BRIT1*:E2F1@nucleus -..>rBRCA1@cytoplasm

These statements can be prepared in any simple text-based editor and imported to Cytoscape environment through BiNoM and then converted directly into the CellDesigner diagram presented in Figure 4. The syntax of this BiNoM Reaction Format (BRF) language have been depicted in [19,24] and in the BiNoM manual (https://binom.curie.fr/docs/BiNoM_Manual_v2.pdf).

Map Entities Annotation

Common entity annotation format and consistent integration of stable identifiers for map entities are essential for compatibility of maps with other tools. It also facilitates integration of data into maps and other manipulations as cross-curation of maps by specialists and incorporation of their corrections into the map. We have developed the NaviCell annotation format for each entity that is applied during map construction in CellDesigner (Figure 5). The annotation panel includes sections 'Identifiers', 'Maps_Modules', 'References' and 'Confidence'. 'Identifiers' section provides standard identifiers and links to the corresponding entity descriptions in HGNC, UniProt, Entrez, SBO, Gene Cards and cross-references in REACTOME [5], KEGG [4], Wiki Pathways [25] and other databases. Metabolites and small compounds are annotated by corresponding identifiers and linked to ChEBI [26], PubChem Compound [27] and KEGG Compound [28] databases. 'Maps_Modules' section includes links to modules of the DNA map were the entity participates. In addition, since our example, DNA repair map, is part of Atlas of Cancer Signaling Network (ACSN) resource [1], the links to maps and modules of ACSN where the entity participates, are also provided. 'References' section includes notes added by the map manager and links to relevant publications (Figure 5A). Each entity annotation is represented as a post in the web-blog generated when the NaviCell map is generated from CellDesigner file (described below). The web-blog is providing a possibility of communication between map users and map managers. Comments can be submitted at the blog post page in text form together with files of any type (Figure 5A). The extended description of annotation formats for each type of entities on the map is provided in https://navicell.curie.fr/doc/NaviCellMapsPreparationProcedure.pdf.

Figure 5: Entity and complex annotation.

(A) NaviCell annotation of p53 protein, (B) NaviCell annotation of p53/PARP complex.

Confidence Scores

We introduced two simple confidence scores for proteins complexes and reactions that are provided in "Confidence" in a form "five-star" diagram and calculated automatically while

conversion of CellDesigner map to NaviCell format. Both scores represent integer numbers varying from 0 (undefined confidence) to 5 (high confidence). The reference score, marked by 'REF' indicates both the number and the 'weight' associated to publications found in the annotation of a given reaction, with weight equal 1 point for an original publication and 3 points for a review article. The functional proximity score, marked by 'FUNC' is computed based on the external network of protein-protein interactions (PPI), Human Protein Reference Database (HPRD) [29]. The score reflects an average distance in the PPI graph between all proteins participating in the reaction as reactants, products or regulators. The functional proximity is computed using BiNoM Cytoscape plugin [23] (Figure 5B).

GENERATION OF NAVICELL MAP USING NAVICELL FACTORY AND EXCHANGE FORMATS

CellDesigner maps annotated in the NaviCell format or un-annotated, can be converted into a NaviCell web-based front-end, which represents a set of html pages with embedded JavaScript code that can be launched in a web browser locally or put on a web-server for further online use. Use of identifiers in the annotation of proteins ("HUGO:XXX" tag) will allow using NaviCell data visualization functionality. The NaviCell factory is embedded in the BiNoM Cytoscape plugin and also available as a stand-alone command line package (https://github.com/sysbio-curie/NaviCell). The detailed guide of using the NaviCell factory is provided at https://navicell.curie.fr/doc/NaviCellMapperAdminGuide.pdf.

The maps generated in CellDesigner and exposed in NaviCell format can also be provided in common exchange formats to ensure compatibility of maps with other computational tools. Currently, maps can be generated in BioPAX and PNG formats. In addition, the module composition of maps can be provided in a form of GMT files. The description of map preparation in various formats using the BiNoM Cytoscape plugin is available in the BiNoM manual https://binom.curie.fr/docs/BiNoM_Manual_v2.pdf.

MAP NAVIGATION

Comprehensive signaling maps, as DNA repair map, contain large number of nodes and edges that makes navigation through the map difficult. To solve this problem, maps can be represented as clickable web pages with a clear graphic user interface. We have developed user-friendly NaviCell web-based environment empowered by Google Maps engine for visualization and navigation of the comprehensive maps [17]. These features are demonstrated in our example: scrolling, zooming, markers, callout windows and zoom bar are adopted from the Google Maps interface (Figure 3C). The map in NaviCell is interactive and all components of the map are 'clickable'. For finding the entity of interest, querying for single or multiple molecules using the search window is possible. Alternatively, the entity can be found in the selection panel or directly on the map.

We describe several solutions to optimize navigation through the maps. First, 'horizontal' navigation that is facilitated by the hierarchical modular structure of DNA repair map, as it is

described in previous paragraphs. This modular structure allows shuttling between the global map to the functional module maps and observe different processes on separate modular maps, containing only limited number of entities, as it is demonstrated in the DNA repair map example in Figure 3B.

Second, 'vertical' navigation, is facilitated by the semantic zooming feature of maps, especially prepared in NaviCell format. Semantic zooming simplifies navigation through the large maps of molecular interactions, showing readable amount of details at each zoom level. Gradual appearance of details allows exploration of the map content from the top-level toward detailed view. To prepare maps for this type of navigation, pruning of maps is performed in order to eliminate non-essential information for each zoom level. We recommend to prepare four zoom levels, although number of zooms is unlimited in NaviCell.

In our example of DNA repair map, the first, top-level view, shows modules of the map depicted as colored background shapes (Figure 3A). At the next level, a more detailed level of zooming shows canonical cell signaling pathways. These pathways are defined by intersecting the content of the map with the corresponding pathways in other databases (Figure 6A). Next zoom hides names of complexes, entities and reaction (Figure 6B) and the last zoom is the most detailed view where all map elements are present (Figure 6C). The module background coloring appears as a context layer in the background of all levels of zooming. For detailed instructions on map zoom levels creation see in the Map creator guide (https://navicell.curie.fr/doc/NaviCellMapperAdminGuide.pdf)and in the Map preparation in NaviCell format procedure (https://navicell.curie.fr/doc/NaviCellMapsPreparationProcedure.pdf).

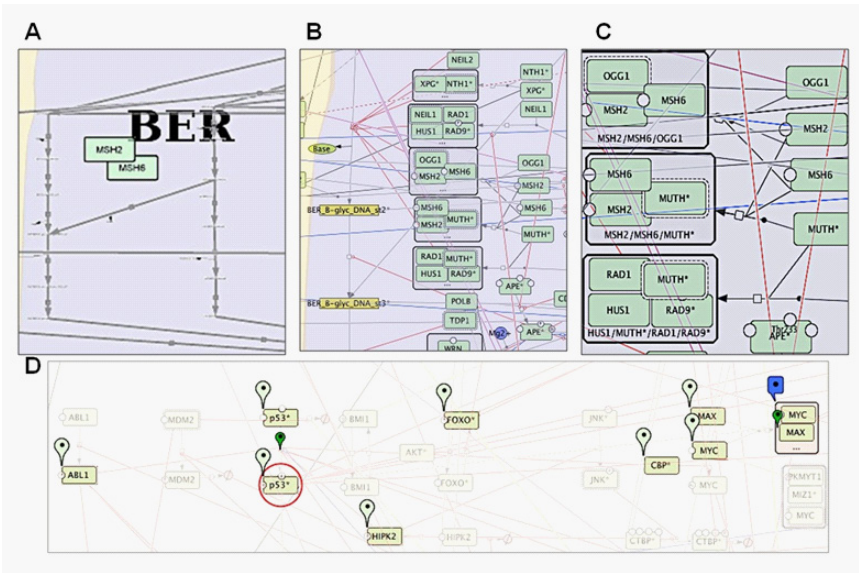


Figure 6: Semantic zooming and entity visualization on DNA repair map.

(A) Canonical pathways view, (B) Hide-details view, (C) Detailed view, (D) "Highlighting" p53 neighbours.

A third way to facilitate exploration of a map is focusing on individual entities of the map by highlighting them. We have developed a NaviCell function of selecting and highlighting species of interest or neighbors of a molecular species of interest. This function allows step-wise enlarging of the neighborhood coverage to understand propagation of signaling on the map as shown for p53* molecule on the DNA repair map in Figure 6D.

MAP MAINTENANCE AND CURATION

Given the fact that biological knowledge about the majority of signaling pathways is not yet solid and continuously grows, one of the major problems of signaling maps is their fast obsolescence. To overcome the problem, permanent maintenance and updating of maps is essential. The community of users is the most reliable and trustable contributor to map maintenance, because specialists could support and update maps from the area of their own research that would ensure highest quality of maps update. To enable such a community-based effort, efficient curation tools should be created. To our knowledge, there is only one community curation tool for comprehensive maps, the Payoa plugin of CellDesigner (30).

We recommend to carry out map curation in the context of NaviCell environment. The process of map curation and maintenance in NaviCell involves map managers that regularly examine the comments posted in the blog of the maps (Figure 5A), check the latest scientific literature and update the maps and the annotations accordingly. An automated procedure supports the map updating and archives older versions of posts including comments, thus providing traceability of all changes on the maps and all discussions in the blog [17].

VISUALIZATION OF OMICS DATA IN THE CONTEXT OF SIGNALING NETWORK MAPS

To make data visualization a straightforward and easy task, we have developed a built-in toolbox for visualization and analysis of high-throughput data in the context of - comprehensive signaling networks. The integrated NaviCell web-based toolbox allows importing and visualizing heterogeneous omics data on top of the maps and performing simple functional data analysis. It is also suitable for computing aggregated values for sample groups and protein families and mapping this data onto the maps. The tool contains standard heatmaps, barplots and glyphs as well as the novel map staining technique for displaying large-scale trends in the numerical values along the map. The combination of these flexible features provides an opportunity to adjust the modes of visualization to the data type and achieve the most meaningful picture [18]. An extended documentation, tutorial, live example and guide for data integration using NaviCell is provided at https://navicell.curie.fr/pages/nav_web_service.html.

To illustrate data visualization, the DNA map was used for analysis of omics data from breast cancer patients. To grasp the difference in the data distribution on top of the map and rewiring of signaling processes across different stages of breast cancers, the module activity was calculated based on transcriptomic data. The colors represent the average contribution of

all module components (Figure 7A). There is a clear difference in pattern between non-invasive stage 1 group of breast cancer patients to stages groups 4 invasive group, indicating a major shift in signaling involvement while the cancer cells transformation from non-invasive to invasive. In addition, closer look at the cell cycle checkpoints shows that spindle checkpoint is activated in the stage 4, whereas all other cell cycle checkpoints are down regulated. The observation is consistent with the 'checkpoint addiction' phenomena, when the tumors develop dependence on spindle checkpoint to allow chromosomes separation despite accumulated genomic instability and ensures cell proliferation regardless of the status of DNA damage [31].

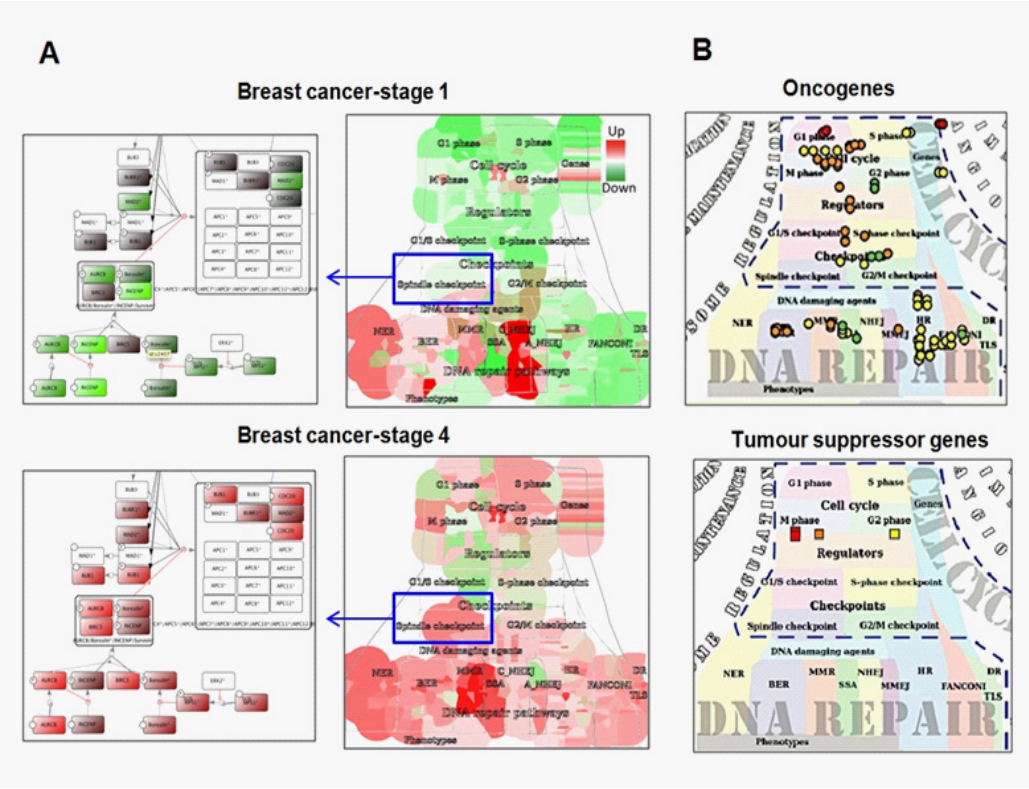


Figure 7: Visualization of high-throughput data in the context of DNA repair map.

(A) Visualisation of difference in gene expression between two breast-cancer grades (Red – upregulation, green –downregulation). Difference are visualized at individual protein (left panel) and functional module (right panel) level, (B) Visualization of different functional types of cancer-associated mutations.

In another example, in order to illustrate the coverage of gene mutation frequencies over the DNA repair map, the information about mutations in breast cancer was obtained from the Catalogue Of Somatic Mutations In Cancer (COSMIC) database [32] and mapped the most frequently mutated oncogenes and tumor suppressor genes (TSG) as glyphs using NaviCell data visualization toolbox (Figure 7B). TSG mutations are more frequently found in modules of DNA

repair map than oncogenes. This might indicate that TSGs in those processes normally contribute to the restriction of uncontrolled divisions and unrepaired DNA, but are inactivated by mutations in cancer.

An additional way to visualize the omics data on top of the maps that we also developed, but do not detail in this chapter, is using Cytoscape plugin BiNoM were the map staining displaying module activity values and coloring individual protein nodes with corresponding transcriptome data is possible [24].

CONCLUSIONS AND PERSPECTIVES

Representation of relationships between cell molecules in the form of a biochemical process diagram depicts our current understanding of how cell activity is coordinated at the molecular level. The advantage of drawing biological processes in the form of interconnected network is not only to bring together components that participate in the described process, but also to allow capturing non-trivial interactions and regulatory circuits between those components. As the load of knowledge about biological mechanisms increasingly grows, organization, structuring and systematized representation of this data is essential for creating the global picture. Standardized representation of biological processes, intuitive maps navigation tools, community contribution to revising and updating the networks diagrams simplify construction of new networks and facilitate maintenance of existing signaling diagrams collections. These comprehensive signaling maps serve as a basis for modeling of signaling networks and efficient analysis and interpretation of high-throughput data [70].

In this chapter we have described the methodology developed in the group following our long-standing experience with comprehensive maps generation and manipulation. Using this approach we have created and currently maintain a pathway resource of Atlas of Cancer Signaling Network (ACSN) [9] and a collection of maps created in CellDesigner available at https://navicell.curie.fr/pages/maps.html. We have suggested a workflow for construction and annotation of signaling maps in CellDesigner, preparing the hierarchical modular structure of maps and also generation of different levels of the maps view, to allow semantic zooming-based exploration of maps in NaviCell. We have introduced NaviCell that is an environment for navigating large-scale maps of molecular interactions created in CellDesigner. NaviCell allows showing the content of the map in a convenient way, at several scales of complexity or abstraction; it provides an opportunity to comment on map content, facilitating curation and maintenance of the map. Finally we have shown how complex data can be visualized and interpreted in the context of the map.

Among many future challenges of the signaling network community are integration of similar efforts as improvement of network exchange formats and development of common network dynamic layouts. In addition, generation of comprehensive platforms for tools, data, and knowledge sharing in systems biology and biomedical research, similar to GARUDA initiative (http://www.garuda-alliance.org), will facilitate tools and resources compatibility improvement.

DOCUMENTATION

CellDesigner introduction and tutorial

http://celldesigner.org/documents.html

SBGN

http://www.sbgn.org/Main_Page

BiNoM manual

https://binom.curie.fr/docs/BiNoM_Manual_v2.pdf

Map creator guide

https://navicell.curie.fr/doc/NaviCellMapperAdminGuide.pdf

Map preparation in NaviCell format procedure

https://navicell.curie.fr/doc/NaviCellMapsPreparationProcedure.pdf

NaviCell Web Service guide

https://navicell.curie.fr/doc/ws/NaviCellWebServiceGuide.pdf

NaviCell Web Service introduction, tutorial and case studies

https://navicell.curie.fr/pages/nav_web_service.html

Interactive demo on data visualization using NaviCell

https://navicell.curie.fr/navicell/maps/cellcycle/master/index.php?demo=on

ACSN introduction, tutorial and case studies

https://acsn.curie.fr/documentation.html

DNA repair map from ACSN resource

https://acsn. curie.fr/navicell/maps/dnarepair/master/index.html

References

1. Kuperstein I, Grieco L, Cohen DPA, Thieffry D, Zinovyev A, et al. The shortest path is not the one you know: application of biological network resources in precision oncology research. Mutagenesis. 2015; 30: 191–204.
2. Cohen D, Kuperstein I, Barillot E, Zinovyev A, Calzone L. From a biological hypothesis to the construction of a mathematical model. Methods Mol Biol. 2013; 1021: 107–125.
3. Chowdhury S, Sarkar RR. Comparison of human cell signaling pathway databases--evolution, drawbacks and challenges. Database. 2015; 126.
4. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012; 40: 109–114.
5. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014; 42: 472–477.
6. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, et al. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res. 2005; 33: 284–288.

7. Paz A, Brownstein Z, Ber Y, Bialik S, David E, et al. SPIKE: a database of highly curated human signaling pathways. Nucleic Acids Res. 2011; 39: 793–799.

8. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res. 2013; 41: 793–800.

9. Kuperstein I, Bonnet E, Nguyen H-A, Cohen D, Viara E, et al. Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. Oncogenesis. 2015; 4: e160.

10. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, et al. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014; 42: 199–205.

11. Le Novère N, Hucka M, Mi H, Moodie S, Schreiber F. The Systems Biology Graphical Notation. Nat Biotechnol. 2009; 27: 735-741.

12. Czauderna T, Klukas C, Schreiber F. Editing, validating and translating of SBGN maps. Bioinformatics. 2010; 26: 2340-2341.

13. Hu Z, Mellor J, Wu J, DeLisi C. VisANT: an online visualization and analysis tool for biological interaction data. BMC Bioinformatics. 2004; 5: 17.

14. Kitano H, Funahashi A, Matsuoka Y, Oda K. Using process diagrams for the graphical representation of biological networks. Nat Biotechnol. 2005; 23: 961-966.

15. Flórez LA, Lammers CR, Michna R, Stülke J. CellPublisher: a web platform for the intuitive visualization and sharing of metabolic, signalling and regulatory pathways. Bioinformatics. 2010; 26: 2997-2999.

16. Kono N, Arakawa K, Ogawa R, Kido N, Oshita K. Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API. PLoS One. 2009; 4: e7710.

17. Kuperstein I, Cohen DP, Pook S, Viara E, Calzone L. NaviCell: a web-based environment for navigation, curation and maintenance of large molecular interaction maps. BMC Syst Biol. 2013; 7: 100.

18. Bonnet E, Viara E, Kuperstein I, Calzone L, Cohen DP1. NaviCell Web Service for network-based data visualization. Nucleic Acids Res. 2015; 43: W560-565.

19. Calzone L, Gelay A, Zinovyev A, Radvanyi F, Barillot E. A comprehensive modular map of molecular interactions in RB/E2F pathway. Mol Syst Biol. 2008; 4: 173.

20. Chanrion M, Kuperstein I, Barrière C, El Marjou F, Cohen D. Concomitant Notch activation and p53 deletion trigger epithelial-to-mesenchymal transition and metastasis in mouse gut. Nat Commun. 2014; 5: 5005.]

21. Tian H, Gao Z, Li H, Zhang B, Wang G1. DNA damage response--a double-edged sword in cancer prevention and cancer therapy. Cancer Lett. 2015; 358: 8-16.

22. Wang H, Zhang X, Teng L, Legerski RJ. DNA damage checkpoint recovery and cancer development. Exp Cell Res. 2015; 334: 350-3588.

23. Bonnet E, Calzone L, Rovera D, Stoll G, Barillot E, et al. BiNoM 2.0, a Cytoscape plugin for accessing and analyzing pathways using standard systems biology formats. BMC Syst Biol. 2013; 7: 18.

24. Bonnet E, Calzone L, Rovera D, Stoll G, Barillot E. BiNoM 2.0, a Cytoscape plugin for accessing and analyzing pathways using standard systems biology formats. BMC Syst Biol. 2013; 7: 18.

25. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR. WikiPathways: building research communities on biological pathways. Nucleic Acids Res. 2012; 40: D1301-1307.

26. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 2008; 36: D344-350.

27. Li Q, Cheng T, Wang Y, Bryant SH. PubChem as a public resource for drug discovery. Drug Discov Today. 2010; 15: 1052-1057.

28. Kanehisa M. Molecular network analysis of diseases and drugs in KEGG. Methods Mol Biol. 2013; 939: 263-275.

29. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S. Human Protein Reference Database--2009 update. Nucleic Acids Res. 2009; 37: D767-772.

30. Matsuoka Y, Ghosh S, Kikuchi N, Kitano H. Payao: a community platform for SBML pathway model curation. Bioinformatics. 2010; 26: 1381-1383.

31. Bartek J, Lukas J. DNA damage checkpoints: from initiation to recovery or adaptation. Curr Opin Cell Biol. 2007; 19: 238-245.

32. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015; 43: D805-811.

33. MacLachlan TK, Dash BC, Dicker DT, El-Deiry WS. Repression of BRCA1 through a feedback loop involving p53. J Biol Chem. 2000; 275: 31869-31875.

34. Yang SZ, Lin FT, Lin WC. MCPH1/BRIT1 cooperates with E2F1 in the activation of checkpoint, DNA repair and apoptosis. EMBO Rep. 2008; 9: 907-915.

35. Chen CF, Li S, Chen Y, Chen PL, Sharp ZD, et al. The nuclear localization sequences of the BRCA1 protein interact with the importin-alpha subunit of the nuclear transport signal receptor. J Biol Chem. 1996; 271: 32863-32868.

36. Zhang J, Willers H, Feng Z, Ghosh JC, Kim S. Chk2 phosphorylation of BRCA1 regulates DNA double-strand break repair. Mol Cell Biol. 2004; 24: 708-718.

37. Xia Y, Pao GM, Chen HW, Verma IM, Hunter T. Enhancement of BRCA1 E3 ubiquitin ligase activity through direct interaction with the BARD1 protein. J Biol Chem. 2003; 278: 5255-5263.

38. Ghosh S, Matsuoka Y, Kitano H. Connecting the dots: role of standardization and technology sharing in biological simulation. Drug Discov Today. 2010; 15: 1024-1031.

39. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013; 41: D808-815.

40. Chatr-Aryamontri A, Breitkreutz BJ2, Oughtred R3, Boucher L2, Heinicke S3. The BioGRID interaction database: 2015 update. Nucleic Acids Res. 2015; 43: D470-478.

41. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2012; 40: D857-861.

42. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011; 39: D685-690.

43. Schacherer F, Choi C, Götze U, Krull M, Pistor S. The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. Bioinformatics. 2001; 17: 1053-1057.

44. Zinovyev A, Viara E, Calzone L, Barillot E. BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks. Bioinformatics. 2008; 24: 876-877.

45. Wu G, Dawson E, Duong A, Haw R, Stein L. ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. Version 2. 1000.

46. Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P. iPath2.0: interactive pathway explorer. Nucleic Acids Res. 2011; 39: W412-415.

47. Pavlopoulos GA, Hooper SD, Sifrim A, Schneider R, Aerts J. Medusa: A tool for exploring and clustering biological networks. BMC Res Notes. 2011; 4: 384.

48. Day RN, Davidson MW. Fluorescent proteins for FRET microscopy: monitoring protein interactions in living cells. Bioessays. 2012; 34: 341-350.

49. Piehler J. New methodologies for measuring protein interactions in vivo and in vitro. Curr Opin Struct Biol. 2005; 15: 4-14.

50. Sklar LA, Edwards BS, Graves SW, Nolan JP, Prossnitz ER. Flow cytometric analysis of ligand-receptor interactions and molecular assemblies. Annu Rev Biophys Biomol Struct. 2002; 31: 97-119.

51. Berggård T, Linse S, James P. Methods for the detection and analysis of protein-protein interactions. Proteomics. 2007; 7: 2833-2842.

52. Phizicky EM, Fields S. Protein-protein interactions: methods for detection and analysis. Microbiol Rev. 1995; 59: 94-123.

53. Jubb H, Higueruelo AP, Winter A, Blundell TL. Structural biology and drug discovery for protein-protein interactions. Trends Pharmacol Sci. 2012; 33: 241-248.

54. Sambrook J, Russell DW. Detection of Protein-Protein Interactions Using the GST Fusion Protein Pulldown Technique. CSH Protoc. 2006.

55. Wu Y, Li Q, Chen XZ. Detecting protein-protein interactions by Far western blotting. Nat Protoc. 2007; 2: 3278-3284.

56. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003; 422: 198-207.

57. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol. 1998; 280: 1-9.

58. Dey B, Thukral S, Krishnan S, Chakrobarty M, Gupta S. DNA-protein interactions: methods for detection and analysis. Mol Cell Biochem. 2012; 365: 279-299.

59. Stormo GD, Zhao Y. Determining the specificity of protein-DNA interactions. Nat Rev Genet. 2010; 11: 751-760.

60. Tompa M, Li N, Bailey TL, Church GM, De Moor B. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol. 2005; 23: 137-144.

61. Kuhn DE, Martin MM, Feldman DS, Terry AV Jr, Nuovo GJ. Experimental validation of miRNA targets. Methods. 2008; 44: 47-54.

62. VanGuilder HD, Vrana KE, Freeman WM. Twenty-five years of quantitative PCR for gene expression analysis. Biotechniques. 2008; 44: 619-626.

63. Alam J, Cook JL. Reporter genes: application to the study of mammalian gene transcription. Anal Biochem. 1990; 188: 245-254.

64. Prediger EA. Detection and quantitation of mRNAs using ribonuclease protection assays. Methods Mol Biol. 2001; 160: 495-505.

65. Kurien BT, Scofield RH. Western blotting: an introduction. Methods Mol Biol. 2015; 1312: 17-30.

66. Bonner WA, Hulett HR, Sweet RG, Herzenberg LA. Fluorescence activated cell sorting. Rev Sci Instrum. 1972; 43: 404-409.

67. Manis JP. Knock out, knock in, knock down--genetically manipulated mice and the Nobel Prize. N Engl J Med. 2007; 357: 2426-2429.

68. Tiscornia G, Singer O, Ikawa M, Verma IM. A general method for gene knockdown in mice by using lentiviral vectors expressing small interfering RNA. Proc Natl Acad Sci U S A. 2003; 100: 1844-1848.

69. Shoemaker BA, Panchenko AR. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. PLoS Comput Biol. 2007; 3: 0337-0344.

70. Barillot E, Calzone L, Hupe P, Vert J-P, Zinovyev A. Computational Systems Biology of Cancer. CRC Press. 2012; 461-432.

Austin Publishing Group

Neuro-Ligands Optimization Using Molecular Modeling

Chaturvedi S[1] and Mishra Anil K[1]*

[1]Division of Cyclotron and Radiopharmaceutical Sciences, Institute of Nuclear Medicine and Allied Sciences, India

***Corresponding author:** Mishra Anil K, Division of Cyclotron and Radiopharmaceutical Sciences, Institute of Nuclear Medicine and Allied Sciences, India, Fax: +91.11.23919509; Email(s): akmishra63@gmail.com, akmishra@inmas.drdo.in

Published Date: December 01, 2016

INTRODUCTION

In an attempt to cut short the time, efforts and resources for the drug development, a whole new branch of Computer- Aided Drug Design **(CADD)** in pharmaceutical research has evolved. This branch **(CADD)** is based on the theoretical knowledge of over hundred years and is evolving at a considerable pace around the world to give a quick solution for an ideal drug candidate.

Computer Aided Drug Design

Computer aided drug design is a rational drug design approach that encompasses computational methods for studying virtual screening of compounds, physicochemical parameters evaluation and chemical interactions. Its roots can be traced back as early as 1953 when Monte Carlo and Metropolis algorithm were developed using advanced MANIAC computer systems [1]. In 90's, high throughput screening and combinatorial chemistry methods gained importance in drug discovery. The growing importance can be assessed by the fact that journals dedicated on only CADD are available. Success stories of CADD in clinics include Captopril- antihypertensive drug, Dorzolamide- for the treatment of glaucoma, Saquinavir- HIV-1 protease inhibitor, Zanamivir- neuraminidase inhibitor as few examples [2].

Initially, the work using CADD was based on structure-activity relationships which is dependent on ligand structure and hence popularly referred as Ligand based Method. Ligand based methods are based on the concept of "similarity" that is, molecules those bind to the same drug target will have similar structural and physicochemical parameters that can be correlated with the pharmacological effect. Further, new leads will be a result of extrapolation of this correlation. 2D methods such as fingerprint similarity searching algorithms, 3D QSAR- Quantitative structure activity relationship and pharmacophores modeling are the main methods [3,4]. However, with the access to protein structures, structure-based drug design came into existence. Established in 1971 with only seven structures, the Protein Data Bank [5] of the Research Collaboratory of Structural Bioinformatics **(RCSB)** is the repository available to access the crystallographic as well as NMR structures of proteins. The structures are stored in PDB format with a universal accession number. Thus, in the structure-based drug design, the potential ligands are mapped onto the 3D structure of the target and are evaluated as suitable drug candidates.

Computer Aided Drug Design for Neuroligands

The brain is the house to many neurotransmitters, neuro-receptors, and enzymes. Indeed, it is a complex interplay of balance between them that creates the delicate balance of neuro-behaviour.

CADD methods have been extensively used for drug discovery. Few success examples are listed above. CADD methods are also being used specifically for the development of neuro-ligands and have been widely used for pharmacophores generation, high throughput screening and drug design for various diseases namely Alzheimer's disease [6] and antidepressants [7].

Challenges for CADD for Neuroligands

What makes this topic of developing neuro-ligands through CADD interesting is, (a) fact that a large number of neurotransmitters receptors are GPCR membrane proteins and do not have the structural information in the form of PDB, (b) generation of unique receptor models using homology model with the help of the structural knowledge of a handful of receptors, and (c) interesting patterns of existence, as dimers and higher order oligomers, of neuro-receptors as majority of them are G-Protein coupled receptors.

G-protein Coupled Receptors

What are GPCR and why they belong to the intriguing family of proteins?

G-Protein coupled receptors are physiologically an important class of receptors. This superfamily or the largest family, to be more specific, of membrane integral proteins plays an important role in signal transduction pathways [8] and also modulates the responses of neurotransmitters in the brain. The family includes adrenergic, dopaminergic, serotonergic, and muscarinic receptors, opsins, and other related receptors. GPCR are also an attractive target for drug discovery with 40% drugs targeting GPCRs. Still, GPCR drug discovery is a challenge with an average value of only one new GPCR per year being drugged in the last decade [8]. Of the family, neuro-receptors present even a more challenging task.

The neuroreceptors identified till date belong to either the GPCR family or ligand-gated ion channels.

CADD for neuroreceptors has been applied for virtual screening predominantly using the ligand-based methods. For the structure-based methods, the impediment was the non-existence of GPCR receptors. Till late 2007, bovine rhodopsin was the "template of choice" and the only available template for the construction of structures based on homology model [9]. Only six receptors in various forms were crystallized till 2011 for the purpose of structure deposition as PDB [10]. The inability for crystallization has been attributed to the instability these receptors demonstrate after being devoid of their native cellular environment. Hence, various techniques for receptor stabilization have been used for the crystallization of nearly 21 new receptors [8]. Thus due to an upsurge in the number of PDB's submitted for GPCR, structure-based approach also has gained momentum. A comprehensive list of selected PDB's has been presented as Appendix 1 based on the compilation [8] and experience of the authors for easy reference.

Structure generation using PDB

For an effective structure generation for receptors whose PDB's are not available, homology modeling is performed. The basis of homology modeling is that one class of proteins is structurally similar. However, the challenges associated are many. Indeed, structure generation can be one of the trickiest and influencing component in structure based method. Few criteria for successful homology modelling are:

a) Selection of Template: A large number of templates are now available including the ligand-free and bound form. Hence, the template has to be selected that most mimics the end requirement. The choice of the template(s) for the homology model construction can have very many implications for example Cα positional variations of the order of 1- 3 Å (or more) in the active sites of enzymes which further may lead to substantial side chain displacements in the final model [11]. Outcomes of the blind GPCR modeling competitions using (D3R, A2AR, and CXCR4) reflect that similarity to a template and docking accuracy may not go hand-in-hand [12].

b) Sequence alignment [11]: The sequence alignments (single/ multiple) between the target and templates cannot be 100% because of gaps and insertions and there can be situations wherein a low sequence identity (< 40%) template has to be used. In cases where the sequence similarity might be of the order of 30% only, the reliability of the model needs to be validated through reference docking and validation.

c) Loop and side chain refinement [11]: Accurate prediction of residue side chain conformations which are physically relevant for ligand-receptor association and the positions of the residue side chain lining the active site of a homology constructed model can be difficult.

d) Optimization: Over optimization of a model might drive the model away from physical reality.

e) Experimental data compliance [11]: The model should be in line with the experimental data available viz, disulphide bridges, H-bonding between the residues, active site amino acids through site-directed mutagenesis, etc. A disparity can lead to erroneous results especially during virtual screening.

f) Coligands and role of water molecules: Mediating water molecules and counter ions inside the active site may mediate binding. However, the role is difficult to predict, except in the few conserved (tightly bound/structural) examples that are inferred from X-ray structures.

g) Multiple states existence: GPCR are known to exist in different functional states with preferential binding for agonists and antagonist in those states. An antagonist binding can be in a different plane of the membrane. This is exactly what has been reported for the difference in binding of antagonist ZM241385 with A2A versus β2-adrenergic receptors and rhodopsin. This binding is accompanied with differences in helical positions, extracellular loop organization and 'toggle switch' arrangement [13]. Hence, choice of a template should be done with care.

Every technique is beset with certain limitations, and this is true for CADD as well. While choosing a workflow purely based on ligand-based methods might appear to be simple and fast though these methods do not account for the protein structural framework. Choosing the structure based workflow may require homology models especially for neuro-receptors. The negative aspects of homology models have been listed above. Structure-based pharmacophores generation is thus, not as straightforward and requires that the ligand be docked to a large number of conformations as in case study example 1.3. After docking the scoring functions need to be properly utilized and optimized.

Few case studies of recent past are being discussed below to highlight the CADD for design and development of neuro-ligands.

CASE STUDY 1: DESIGN OF HIGH-AFFINITY LIGANDS THROUGH VIRTUAL SCREENING

The examples being discussed reflect the application of a combination of structure and ligand based methods. These methods can be used in any form of the arrangement (Figure 1):

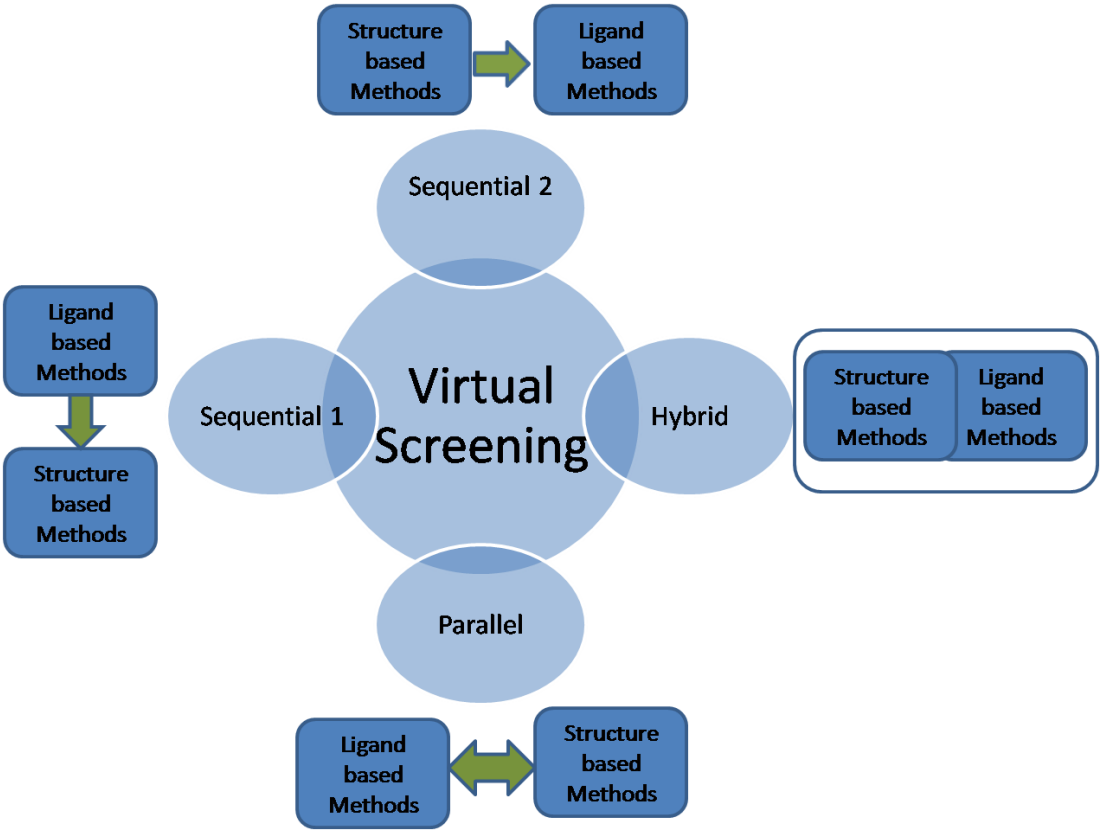


Figure 1: Arrangement of workflows for virtual screening.

Example 1.1: Sequential Manner of Structure Based – Ligand-Based Methods for Virtual Screening

Endocannabinoid receptors CB1 is an attractive drug target. Its agonist plays an important role in the treatment of various disorders. In the study published by Mella-Raipán et al. [14], both SAR and docking studies (Figure 2) were utilized to study the mode of interaction of benzimidazoles based ligands- 1 and 2-naphthyl, and 1 and 2-naphthoyl -2-pyridyl-benzimidazole derivatives. Initially, a library of most putative ligands based on variations in substituent was generated. The compounds conformers were generated and docked onto the receptor homology model. Docking simulation studies of the ligand with the receptor model helped in the prediction of the major interaction regions. After generating the optimized pose, 3D QSAR using CoMFA analysis was used to arrive at the high- affinity ligands where optimal substitution in the different regions of the derivatives were deduced using the energy contour maps.

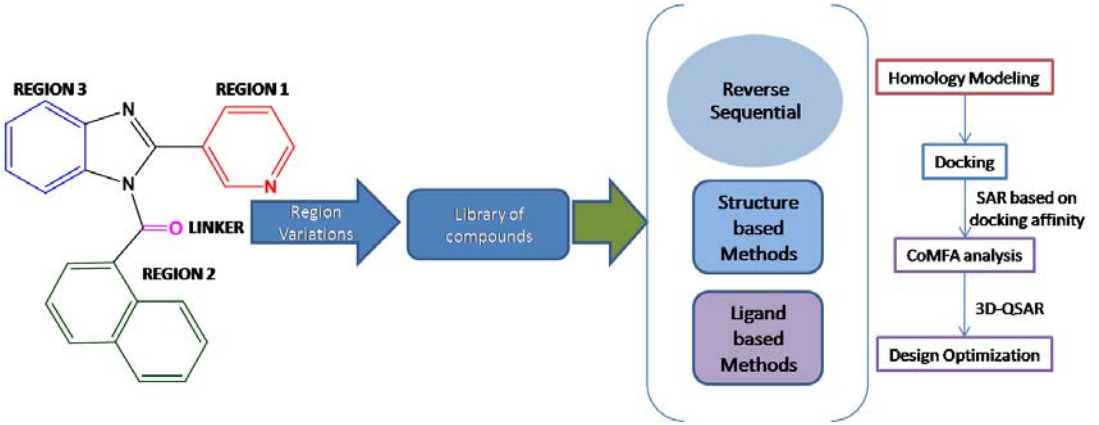


Figure 2: Schematic representation of Example 1.1 [14]

For the study following parameters were used for docking and QSAR studies (Table 1.1 and 1.2):

Table 1.1 : Workflow followed for Docking Studies.

Target:	**CB1R**
Template:	A2aR PDB ID: 3EML
Homology Model [15]:	Sequence alignement: 1. Clustal alignment 2. Manual modification -S-S- bridge between C257 and C264
Generated Models:	Number: 100 internal scoring function energy minimization protocol
Model validation	NIH SAVES server alpha-carbon root-mean-square deviation of 0.64 Å Ramachandran plot analysis- 97% of the residues in allowed regions
Docking	FRED v2.2.5 Scoring : PLP, Chemscore,Chemgauss3 scoring function Minimization of docked pose: CHARMM22 force-field in Discovery Studio v2.1

Table 1.2: Workflow followed for QSAR Studies.

Analysis	CoMFA(Comparative Molecular Field Analysis) using SYBYL-X 1.2
	PLS analysis between the CoMFA descriptors (independent variables) and the affinity values (dependent variables)
cross-validation analysis	LOO method (and SAMPLS), which deduces the square of the cross-validation coefficient (*q2*) and the optimal number of components *N*

Example 1.2: Hybrid Structure Based – Ligand– Based Methods

In the study of Xu et al. [16], agonistic ligands were predicted using Molecular Docking and Molecular Dynamics simulation. In short, the homology models of 5HT1AR was built using β2AR (PDB: 3SN6) and the fifty models generated were shortlisted depending on the lowest DOPE (Discrete Optimized Protein Energy) score. The PDB: SN6 represents the active state ternary complex composed of agonist- occupied monomeric β(2) AR and nucleotide-free Gs heterotrimer

with a resolution of 3.2Å. Dynamic pharmacophores modelling was carried to screen the potential agonistic ligands. Initially the compounds (taken from the database) were screened based on Lipinski rule of five for drug likeliness, docked and the most reasonable complex was compared. Later, Molecular Dynamics simulation revealed not only the interactions of the agonist but also the possible mechanism of activation. Finally, through the screening ten new 5-HT_{1A}R agonists were successfully identified, of which three ligands revealed high potency of Ki values less than 100 nM.

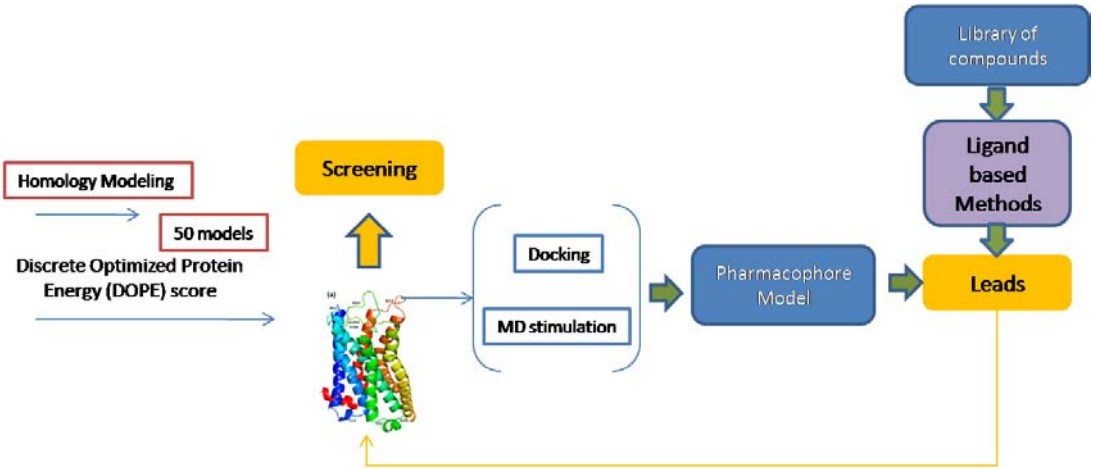


Figure 3: Schematic representation of Example 1.2.

Thus, the study reflects a combination of hybrid ligand based – structure- based combination (Figure 3). Initially, based on structure based method dynamic pharmacophores have been generated. Database compounds were screened parallel based on ligand- based and dynamic pharmacophores. The main highlights of the study are (Table 2.2):

Table 2.2: Workflow for Dynamic VS.

Target:	5HT1A
Template:	3SN6
Homology Model:	(a) ClustalX 2.0 program (b) Discovery Studio 3.5 (c) Discrete Optimized Protein Energy
Minimization:	Prime module of Schrodinger and Loop refinement
Model validation:	PROCHECK Ramachandran plots ≈98% of the residues are in overall allowed regions for both templates VADAR: 3D4S based model was more compact than the 2RH1
Binding site Prediction and validation	binding pocket : defined to include all residues within 10.0 Å of Cγ carbon atom of conserved D3.32
Docking	GoldSuite 5.0 and GoldScore
Molecular Dynamics Simulation	GROMACS 4.5.1 package
Cluster Analysis.	GROMACS
PharmacophoreModel Generation	GRID 22 program
Dynamic Pharmacophore-Based Virtual Screening	Dynamic pharmacophore model as a 3D query followed by screening of databases to get the Hits

Example 1.3: Multistep Structure Based – Ligand– Based Methods

Another study [4,17] for high-throughput screening was reported. The objective was to screen ligands against SERT (serotonin transporter). The difference in approach here is (a) a multistep protocol based on both ligand based and structure-based screening for virtual screening was demonstrated (Figure 4) (b) additional filter viz Veber filters, ADMET parameters for initial screening were tested to screen compounds of the order of ~3.24 million (b) 2D fingerprint-based screening to generate two-dimensional (2D) pharmacophore-based and structural (hashed chemical) fingerprints was used (c) Finally the compounds were docked flexibly onto the homology model of SERT and 74 active SERT binders belonging to 16 structural classes were identified and validated experimentally.

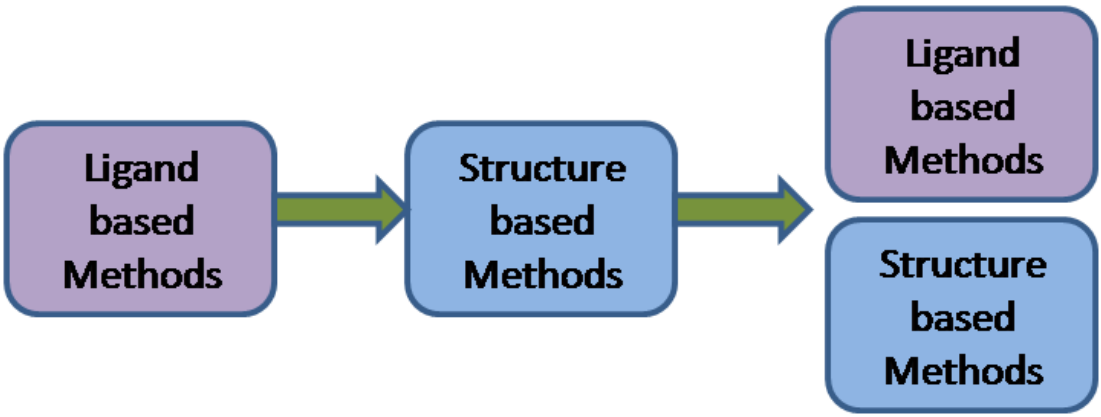


Figure 4: Schematic representation for example 1.3.

Table 1.3(a): Workflow for VS.

For Ligand based screening	
Databases	Asinex, Chem-Bridge, ChemDiv, Enamine, sand Life Chemicals, ZINC database
Preliminary Screening	Lipinski's "rule of 5" and Veber filters
Virtual Screening	Using JChem (a) 2D Fingerprint-Based Screening (b) Structural (hashed chemical) fingerprints
Secondary screening	pKa descriptor ADMET using the Schrödinger software module QikProp
3D Pharmacophore-Based Screening	HipHop algorithm

Table 1.3(b): Workflow for Docking.

For Structure based screening	
Target:	SERT
Template:	LeuT crystal structure PDB id 3F3A
Homology Model:	Internal Coordinate Mechanics **(ICM)** version 3.5 Build Model macro
Minimization:	side chain optimization- Montecarlo
Model validation:	ERRAT quality factor of 88.996
	Ramachandran plot-91.8 percent of the residues were in the core regions
	What Check report
Binding site Prediction and validation	(1) detection of the ligand binding pocket using the ICM Pocket Finder
	(2) biased-probability Monte Carlo **(BPMC)** sampling and minimization of the pocket side chains
Docking	(3) four-dimensional (4D) docking of fully flexible ligands into multiple pocket of 47 low-energy conformations of the ligand binding pockets
Scoring	Virtual ligand screening **(VLS)** scoring function of ICM

Table 1.3(c): Workflow for Hit to lead.

For H2L- second in silico screening	
Filters	Basic property and ADMET filters, 3D pharmacophore models, and the flexible docking procedure

CASE STUDY 2: DEMONSTRATION OF BINDING MODE USING DOCKING PARAMETERS- STRUCTURE BASED METHOD

CADD methods based on docking can be used to evaluate the effect of functionalization/ modifications on binding of ligands, for e.g., whether an agonist retains its agonistic mode or not. This aspect becomes important for the study of the "functional status" of the receptors. The functional imaging has gained impetus due to its implication in several neuropsychiatric diseases. The binding pattern can be exploited to investigate the binding index of the neurotransmitter or drugs to receptors, thereby giving an insight of the functional status. The basis of the functional status underlies in the fact governing the GPCR proteins according to which the high affinity and low affinity states display preferential binding with agonist and antagonist. Antagonists bind to the High-Affinity **(HA)** and Low-Affinity **(LA)** conformations of receptor with comparable affinity. In contrast, agonists bind preferentially to the HA state of the receptor, which is coupled to G-proteins and therefore agonists provide a measure of functional receptors. The paradigm shift from mere imaging of receptors to simultaneously image and quantify functional, demands the design and evaluation of agonist based ligands [18].

Example 2.1 : Evaluation of Binding Pattern Using Structure Based Method And Homology Model

In the study published by Chaturvedi et al., [19] docking parameters were evaluated to understand the effect of functionalization on the natural ligand of serotonin. The ligand was to be developed an imaging agent for active 5HT1A receptors and involved the functionalization with the dithiocarbamate moiety. The study was carried using the Prime (homology modeling) and Glide (docking) modules of Schrodinger. The highlights of the study (refer Table 2.1also) were as follows:

(a) Homology model of $5HT_{1A}$ using two templates – Two models were generated using human β2-adrenergic receptor in the absence (PDB $2RH_{1A}$) and in the presence (PDB 3D4S) of

cholesterol (Figure 5). The choice of the crystal structures was driven by (a) resolution wherein the 2RH1 has the highest resolution of 2.4 Å, (b) template similarity for modeling of aminergic GPCRs (c) mimic for real situation wherein the receptors are found bound to cholesterol units as in PDB 3D4S. The 3D4S model was more compact than the 2RH1 model.

(b) Docking of the ligands- Based on the binding pattern as exemplified by amino acids involved with serotonin- the natural ligand and the modified ligand, it was inferred whether the modified ligand will exhibit the agonistic mode or the antagonistic mode of binding. Thus, the study is based on Ligand site comparison.

(c) Further, in this study, the effect of cholesterol in the CCM (Cholesterol Consensus motif) of the receptor was also studied. The ligand reflected enhanced binding in the presence of cholesterol that can be ascribed to increase in hydrophobic interactions.

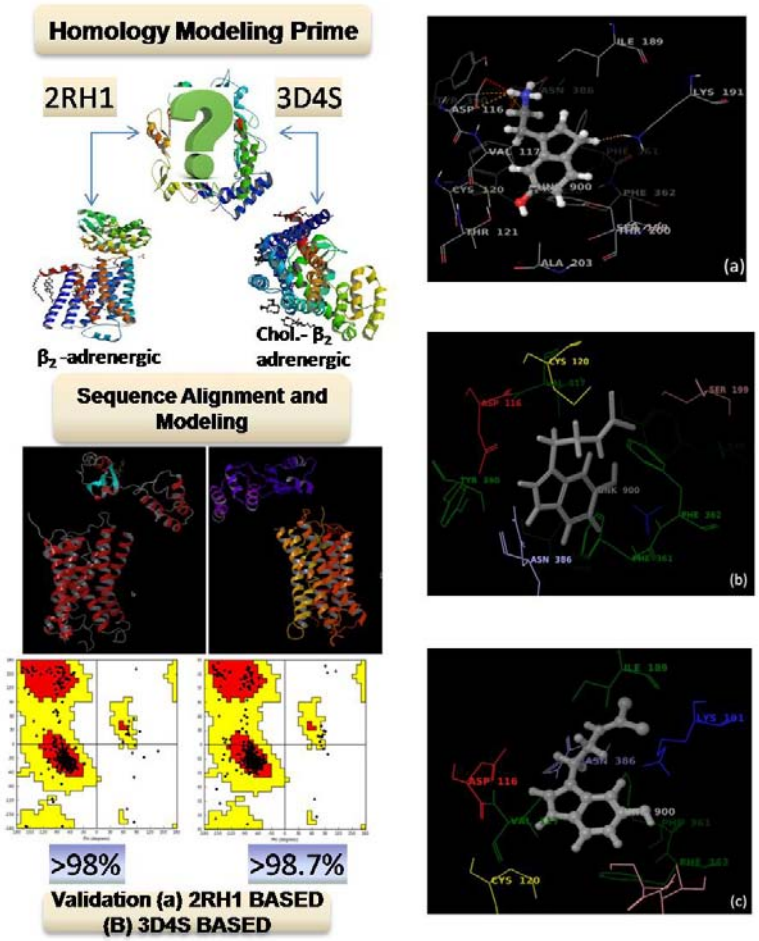


Figure 5: Comprehensive depiction for Example 2.1 [19] where (a) binding mode of SER (b) binding mode of SER-DTC on 2RH1 model (C) binding mode of SER-DTC on 3D4S model.

Table 2.1: Workflow for binding mode prediction.

Target:	5HT1A
Template:	2RH1A and 3D4S
Homology Model:	Clustal_W under the GPCR specific mode (a) highly conserved residues1 in each TM were anchored (b) gaps in the helices were manually removed
Minimization:	Prime module of Schrodinger and Loop refinement
Model validation:	PROCHECK Ramachandran plots ≈98% of the residues are in overall allowed regions for both templates VADAR: 3D4S based model was more compact than the 2RH1
Binding site Prediction and validation	Site map and Site-directed mutagenesis data
Docking and Scoring	Glide module of Schrodinger in extra precision XP mode: G-score and DOCK score

CASE STUDY 3: EFFECT OF MULTIVALENT INTERACTIONS ON RECEPTOR BINDING- STRUCTURE BASED METHODS

Polyvalency in biological systems is an important concept. Multivalent ligands bearing multiple pharmacophores are emerging as important therapeutic ligands because of their enhanced binding affinity [20]. Thus, study of the interaction on binding of a 'bivalent ligand' wherein two or more pharmacophores are linked by a functional spacer and is capable of interaction with two neighbouring receptors or at two sites of the same receptor becomes imperative (Figure 6).

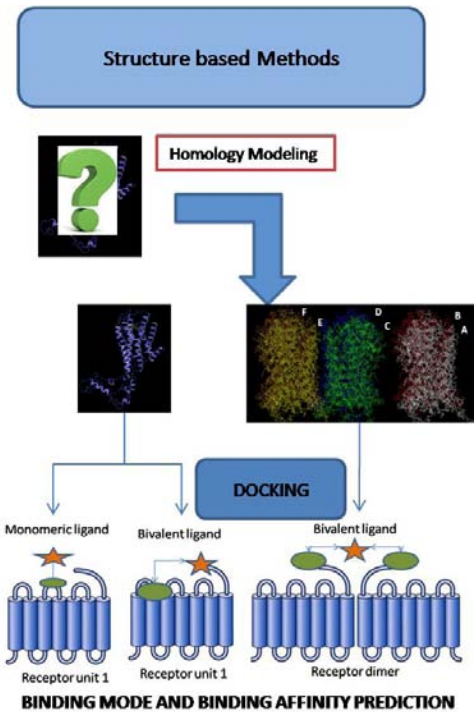


Figure 6: Schematic representation of the workflow to assess the binding pattern for bivalent ligand.

Example 3.1: Dimeric Ligand with Two Sites on the Monomeric Receptor.

The study reported by Sethi et al. [21], reflects the selectivity of the bivalent ligand for dopamine D2 receptor. As reported by the authors, dopamine D2 receptor (D2R) model was built making use of its high sequence similarity with D3 receptor. The PDB 3PBL is a lysozyme-chimeric protein (hD3-lysozyme chimera) modified for crystallisation and purification. Using the modelling studies, the authors could establish enhanced binding through multivalent interactions and calculate the binding affinity for the ligand.

Example 3.2: Predicting the Enhanced Binding of Dimeric Ligand on Monomeric/ Higher Order Receptor

Bivalent ligands are known to exhibit favourable thermodynamics than monovalent ligands.

In the study of Hazari et al. [22], monomeric, homodimeric, and multimeric 5-HT1A receptor models were built and screened for the binding pattern of a bivalent ligand. The main highlights are as follows:

(a) Homology models were built using the template β2-adrenergic receptor, 2RH1. Schrodinger software was extensively used for the study with Prime module for model generation, Glide and Induced fit for docking studies.

(b) Sequence alignment was performed using Clustal W under the GPCR- specific mode.

(c) Secondary structure prediction, supported with PSIPRED, was performed, and this is best known for optimal predictions.

(d) For the higher order structures PDB 1N3M was used. 1N3M is a theoretical model. It was used as a reference because it gave extended coordinates for tridimensional oligomeric model. The generated models were re-validated using standard tools.

(e) The models were validated using Ramachandran Plot and PROCHECK server. The authors inferred that high stabilization of the bivalent ligand due to the π-π interactions and hydrogen bonding resulted in high docking score of the ligand as compared to the monovalent ligand. Also, the modifications did not alter the antagonistic binding pattern of the ligand.

CASE STUDY 4: ENZYMES AS TARGETS FOR NEURO-INHIBITORS

Brain also houses enzymes which are involved in the metabolism of various neurotransmitters and the enzymes have been implicated in various disorders. Acetylcholinesterase inhibitors are widely used for the treatment of Alzheimer's disease and are the only FDA-approved AD therapies. These inhibitors slow the turnover of the neurotransmitter acetylcholine in the synapse. Non-competitive inhibitors may produce slow reversible ChE inhibition and the long term effects of slowly reversible, or irreversible, inhibitors on the overall cholinergic function are difficult to predict. Thus, design of inhibitor against the enzyme reflects another dimension of CADD for neuroligands [23].

The study which follows is based on energy calculations using molecular mechanics and quantum mechanical tools.

Example 4.1: Design of Reversible Inhibitors and Regenerators

Organophosphates are reversible inhibitors of acetylcholineesterase are used in the treatment of AD. However, organophosphates can cause interruption of AChE-mediated mechanism because the inhibited enzyme undergoes phosphorylation and regenerates very slowly leading to tremors, coma, and ultimately death. Design of reversible AChE inhibitors is thus important for the regeneration of irreversibly inhibited acetylcholineesterase from organophosphates. Hence with an aim to probe potential molecules as anticholinesterase inhibitors and as reactivators, computationally structure-based approach has been exploited in the work of Chadha et al. [24], for designing novel 2-amino-3-pyridoxime-dipeptides conjugates. The authors combined MD simulations with flexible ligand docking approach to determine binding specificity of 2-amino-3-pyridoxime dipeptides towards AChE (PDB 2WHP). The highlights of this study are as follows:

(a) MM-GBSA (Molecular mechanics +generalised Born surface area) provides approximate free energies of binding and were correlated with the docking score. The docking results depicted complementary multivalent interactions along with good binding affinity as predicted from MM-GBSA analysis.

(b) In order to gain insight on the mechanism, MD simulations under explicit solvent systems with NPT and NVT ensemble were carried to uncover the dynamic behavior of 2-amino-3-pyridoxime-(Arg-Asn) and expose its mobile nature and interactions. What was inferred from these studies were ability to form strong long range order contacts towards active site residues of the dioxime peptide, its approach towards inhibited serine residue and the hydrogen bonding contacts.

(c) For complete potential surface profile, 2-amino-3-pyridoxime induced reactivation pathway of sarin-serine adduct was investigated by the DFT approach. The authors concluded from DFT's transition state search and reaction scan, that oxime-Arg-Asn is able to reactivate phosphorylated serine residue along the barrierless pathway in gas and solvent phase model.

CONCLUSION

In this section, various applications of the CADD have been discussed specifically using reports on the development of neuroligands. The four sections are dedicated for three technique packages (a) virtual screening using ligand and structure based methods (b) homology modelling and docking studies and (c) molecular mechanics and quantum mechanics application for design of ligands.

The future of CADD in design of neuroligands is promising and is being increasingly applied to study different aspects- leads screening, binding modes, mechanistic aspects of receptor activation/ deactivation on binding with agonists/ antagonist respectively, prediction of transition

states and possible mechanism of inhibition/ reactivation on binding of inhibitors/ reactivators respectively with the enzymes.

Appendix 1: Table of comprehensive list of PDB [8], [25].

2007	b2, inactive	T4L fusion	carazolol	2.40	2RH1	Referenced in [8]
2007	b2, inactive	Fab complex	carazolol	3.40	2R4R	Referenced in [8]
2008	b2, inactive	T4L fusion	timolol	2.80	3D4S	Referenced in [8]
2010	b2, inactive	T4L fusion	ICI 118551	2.84	3NY8	Referenced in [8]
2010	b2, inactive	T4L fusion	Benzofuran derivative	2.84	3NY9	Referenced in [8]
2010	b2, inactive	T4L fusion	Alprenolol	3.16	3NYA	Referenced in [8]
2011	b2, active	Nanobody stabilised	BI-167107	3.50	3P0G	Referenced in [8]
2008	b1, inactive	StaR	cyanopindolol	2.70	2VT4	Referenced in [8]
2011	b1, inactive	StaR	salbutamol	3.05	2Y04	Referenced in [8]
2011	b1, inactive	StaR	dobutamine	2.50, 2.60	2Y00, 2Y01	Referenced in [8]
2011	b1, inactive	StaR	carmoterol	2.60	2Y02	Referenced in [8]
2011	b1, inactive	StaR	isoprenaline	2.85	2Y03	Referenced in [8]
2011	b1, inactive	StaR	carazolol	3.00	2YCW	Referenced in [8]
2011	b1, inactive	StaR	iodocyanopindolol	3.65	2YCZ	Referenced in [8]
2012	b1, inactive	StaR	bucindolol	3.20	4AMI	Referenced in [8]
2012	b1, inactive	StaR	carvedilol	2.30	4AMJ	Referenced in [8]
2013	b1, inactive	StaR	4-(piperazin-1-yl)-1H-indole	2.80	3ZPQ	Referenced in [8]
2013	b1, inactive	StaR	4-methyl-2-(piperazin-1-yl) quinoline	2.70	3ZPR	Referenced in [8]
2008	A2A, inactive	T4L fusion	ZM241385	2.60	3EML	Referenced in [8]
2011	A2A, active	T4L fusion	UK-432097	2.71	3QAK	Referenced in [8]
2011	A2A, inactive	StaR	caffeine	3.60	3RFM	Referenced in [8]
2011	A2A, inactive	StaR	XAC	3.31	3REY	Referenced in [8]
2011	A2A, inactive	StaR	ZM241385	3.30	3PWH	Referenced in [8]
2011	A2A, active	StaR	NECA	2.60	2YDV	Referenced in [8]
2011	A2A, active	StaR	adenosine	3.00	2YDO	Referenced in [8]
2012	A2A, inactive	StaR	1,2,4-triazine derivative	3.27	3UZA	Referenced in [8]
2012	A2A, inactive	StaR	1,2,4-triazine derivative	3.34	3UZC	Referenced in [8]
2012	A2A	Fab complex	ZM241385	3.10	3VGA	Referenced in [8]
2012	A2A	Fab complex	ZM241385	2.70	3VG9	Referenced in [8]
2012	A2A	inactive BRIL fusion	ZM241385	1.8	4EIY	Referenced in [8]
2012	S1P1	inactive T4L fusion	ML056	2.80	3V2Y	Referenced in [8]
2010	CXCR4	inactive T4L fusion	IT1t	2.50	3ODU	Referenced in [8]
2010	CXCR4	inactive T4L fusion	CVX15	2.90	3OE0	Referenced in [8]
2010	D3	inactive T4L fusion	eticlopride	2.89	3PBL	Referenced in [8]
2011	H1	inactive T4L fusion	doxepin	3.10	3RZE	Referenced in [8]
2011	b2, active	inactive T4L fusion	agonist complex	3.5	3PDS	[26]

2015	A2A active-like	thermostabilised	agonist CGS21680	2.6	4UG2	[27]
2014	class C mGluR5	--	mavoglurant	2.6	4OO9	[28]
2000	Rhodopsin, inactive	--	11-cis-retinal	2.8	1F88	[29]
2004	Rhodopsin, inactive	--	11-cis-retinal	2.2	1U19	[30]
2011	b2-Gs complex	complex	3.2	3SN6		[31]
Oligomeric structures						
2006	Rhodopsin	theoretical model	--	--	1N3M	[32]
2013	β1- basal state	thermostabilizing mutations	ligand-free	3.50	4GPO	[33]
2014	class C mGluR1	--	FITM	2.8	4OR2	[34]

References

1. http://home.comcast.net/~schlecht/Historical_Overview_of_Molecular_Modeling.pdf
2. Talele TT, Khedkar SA, Rigby AC. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. Curr Top Med Chem. 2010; 10: 127-141.
3. Acharya C, Coop A, Polli JE, Mackerell AD. Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. Curr Comput Aided Drug Des. 2011; 7: 10-22.
4. Gabrielsen M, Kurczab R, Siwek A, Wolak M, Ravna AW. Identification of novel serotonin transporter compounds by virtual screening. J Chem Inf Model. 2014; 54: 933-943.
5. Berman HM. The Protein Data Bank: a historical perspective. Acta Crystallogr A. 2008; 64: 88-95.
6. Domínguez JL, Fernández-Nieto F, Castro M, Catto M, Paleo MR. Computer-aided structure-based design of multitarget leads for Alzheimer's disease. J Chem Inf Model. 2015; 55: 135-148.
7. Immadisetty K, Geffert LM, Surratt CK, Madura JD. New design strategies for antidepressant drugs. Expert Opin Drug Discov. 2013; 8: 1399-1414.
8. Andrews SP, Brown GA, Christopher JA. Structure-based and fragment-based GPCR drug discovery. ChemMedChem. 2014; 9: 256-275.
9. McRobb FM, Capuano B, Crosby IT, Chalmers DK, Yuriev E. Homology modeling and docking evaluation of aminergic G protein-coupled receptors. J Chem Inf Model. 2010; 50: 626-637.
10. Congreve M, Langmead CJ, Mason JS, Marshall FH. Progress in structure based drug design for G protein-coupled receptors. J Med Chem. 2011; 54: 4283-4311.
11. Lounnas V, Ritschel T, Kelder J, McGuire R, Bywater RP. Current progress in Structure-Based Rational Drug Design marks a new mindset in drug discovery. Comput Struct Biotechnol J. 2013; 5: e201302011.
12. Beuming T, Sherman W. Current assessment of docking into GPCR crystal structures and homology models: successes, challenges, and guidelines. J. Chem. Inf. Model. 2012; 52: 3263-3277.
13. Maria Hodges, A pocket guide to GPCRs, doi: 10.1038/fa_psisgkb. 2008; 16
14. Mella-Raipán JA, Lagos CF, Recabarren-Gajardo G, Espinosa-Bustos C, Romero-Parra J, et al. Design, synthesis, binding and docking-based 3D-QSAR studies of 2-pyridylbenzimidazoles-a new family of high affinity CB cannabinoid ligands. Molecules. 2013; 18: 3972-4001.
15. Gonzalez A, Duran LS, Araya-Secchi R, Garate JA, Pessoa-Mahana CD, et al. Computational modeling study of functional microdomains in cannabinoid receptor type 1. Bioorg Med Chem. 2008; 16: 4378-4389.
16. Xu L, Zhou S, Yu K, Gao B, Jiang H, Zhen X, et al. Molecular modeling of the 3D structure of 5-HT (1A) R: discovery of novel 5-HT (1A) R agonists via dynamic pharmacophore-based virtual screening. J Chem Inf Model. 2013; 53: 3202-3211.
17. Gabrielsen M, Kurczab R, Ravna AW, Kufareva I, Abagyan R. Molecular mechanism of serotonin transporter inhibition elucidated by a new flexible docking protocol. Eur J Med Chem. 2012; 47: 24-37.
18. Lemoine L, Verdurand M, Vacher B, Blanc E, Le Bars D. [18F] F15599, a novel 5-HT1A receptor agonist, as a radioligand for PET neuroimaging. Eur J Nucl Med Mol Imaging. 2010; 37: 594-605.

19. Chaturvedi S, Kaul A, Yadav N, Singh B., and Mishra Anil K. Synthesis, docking and preliminary *in vivo* evaluation of serotonin dithiocarbamate as novel SPECT neuroimaging agent. Med Chem Comm 2013: 4: 1006-1014.

20. Lane JR, Sexton PM, Christopoulos A. Bridging the gap: bitopic ligands of G-protein-coupled receptors. Trends Pharmacol Sci. 2013; 34: 59-66.

21. Sethi SK, Varshney R, Rangaswamy S, Chadha N, Hazari, PP, Kaul A, et al. Design, synthesis and preliminary evaluation of a novel SPECT DTPA-bis-triazaspirodecanone conjugate for D 2 receptor imaging. RSC Advances, 2014; 4: 50153-50162.

22. Hazari PP, Schulz J, Vimont D, Chadha N, Allard M. A new SiF-Dipropargyl glycerol scaffold as a versatile prosthetic group to design dimeric radioligands: synthesis of the [(18) F] BMPPSiF tracer to image serotonin receptors. ChemMedChem. 2014; 9: 337-349.

23. Francis PT, Palmer AM, Snape M, Wilcock GK. The cholinergic hypothesis of Alzheimer's disease: a review of progress. J Neurol Neurosurg Psychiatry. 1999; 66: 137-147.

24. Chadha N, Tiwari AK, Kumar V, Lal S, Milton MD, et al. Oxime-dipeptides as anticholinesterase, reactivator of phosphonylated-serine of AChE catalytic triad: probing the mechanistic insight by MM-GBSA, dynamics simulations and DFT analysis. J Biomol Struct Dyn. 2015; 33: 978-990.

25. http://www.rcsb.org/

26. Rosenbaum DM, Zhang C, Lyons JA, Holl R, Aragao D. Structure and function of an irreversible agonist-Î² (2) adrenoceptor complex. Nature. 2011; 469: 236-240.

27. Lebon G, Edwards PC, Leslie AG, Tate CG. Molecular Determinants of CGS21680 Binding to the Human Adenosine A2A Receptor. Mol Pharmacol. 2015; 87: 907-915.

28. Doré AS, Okrasa K, Patel JC, Serrano-Vega M, Bennett K. Structure of class C GPCR metabotropic glutamate receptor 5 transmembrane domain. Nature. 2014; 511: 557-562.

29. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, et al. Crystal structure of rhodopsin: A G protein-coupled receptor. Science. 2000; 289: 739-745.

30. Okada T, Sugihara M, Bondar AN, Elstner M, Entel P. The retinal conformation and its environment in rhodopsin in light of a new 2.2 A crystal structure. J Mol Biol. 2004; 342: 571-583.

31. Rasmussen SG, DeVree BT, Zou Y, Kruse AC, Chung KY. Crystal structure of the Î²2 adrenergic receptor-Gs protein complex. Nature. 2011; 477: 549-555.

32. Fotiadis D, Jastrzebska B, Philippsen A, Müller DJ, Palczewski K. Structure of the rhodopsin dimer: a working model for G-protein-coupled receptors. Curr Opin Struct Biol. 2006; 16: 252-259.

33. Huang J, Chen S, Zhang JJ, Huang XY. Crystal structure of oligomeric Î²1-adrenergic G protein-coupled receptors in ligand-free basal state. Nat Struct Mol Biol. 2013; 20: 419-425.

34. Wu H, Wang C, Gregory KJ, Han GW, Cho HP. Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator. Science. 2014; 344: 58-64.

Austin Publishing Group

Rational Drug Discovery: Virtual Screening of DEN-2 Non-Competitive Inhibitors

Heh CH[1], Othman R[1,2], Yusof R[3] and Rahman NA[4]*

Drug Design and Development Research Group (DDDRG)

[1]Department of Pharmacy, Faculty of Medicine, University of Malaya, Malaysia

[2]Center for Natural Product Research & Drug Discovery (CENAR), University of Malaya, Malaysia

[3]Department of Molecular Medicine, Faculty of Medicine, University of Malaya, Malaysia

[4]Department of Chemistry, Faculty of Science, University of Malaya, Malaysia

***Corresponding author:** Noorsaadah Abd. Rahman, Department of Chemistry, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia, Email: noorsaadah@um.edu.my

Published Date: December 01, 2016

BRIEF INTRODUCTION

Dengue, including dengue fever, dengue haemorrhagic fever and dengue shock syndrome is among the major causes of morbidity and mortality, especially in children in many endemic Asian and South American countries [1,2]. Recent study estimated 390 million cases of dengue infection worldwide annually [3]. WHO stated that 40% of the world's population, which are in the tropical or sub-tropical regions of the world, is at risk from dengue infections [4]. Two predominant arthropod vectors, *Aedes aegypti* and *Aedes albopictus*, are implicated in the disease transmission [5-8]. However, to date, there is no licensed vaccine or anti-viral drug available in the market to protect against dengue diseases [9].

Boesenbergia rotunda (L.) Mansf. (synonym of *Boesenbergia pandurata*), known as fingerroot, Chinese ginger (China and Southeast Asia) or "temu kunci" (Malaysia and Indonesia) [10], is a common spice and herb belonging to the Zingiberaceae (ginger) family. Its extract has been reported to contain essential oils [11] and various small compounds such as boesenbergin, cardamonin, pinostrobin, pinocembrin, alpinetin, panduratin A, 4-hydroxypanduratin A, 5,7-dimethoxyflavone and 1,8-cineole [12-14]. It was previously reported that cardamonin (a chalcone) and pinostrobin (a flavanone) showed non-competitive inhibition towards DEN-2 NS2B-NS3 proteolytic activities, while panduratin A and 4-hydroxypanduratin A (both cyclohexenyl chalcone derivatives) showed competitive inhibition activities [14].

In this study, the models of DEN-2 NS2B-NS3 were studied computationally (*in silico*) using cardamonin, R-pinostrobin and S-pinostrobin to verify the suitability of the models as target receptors for non-competitive inhibition studies. When this study was conducted, there was only one DEN-2 NS2B-NS3 protease crystal (PDB id: 2FOM) [15] available in the Protein Data Bank (PDB). On the other hand, other NS2B-NS3 protease crystals from West Nile Virus (WNV) (PDB id: 2FP7; 2GGV; 2IJO; 3E90) [15-17], DEN-1 (PDB id: 3L6P; 3LKW) [18] and DEN-3 (PDB id: 3U1I; 3U1J) [19] have been reported. A few homology modelling studies of DEN-2 NS2B-NS3 have been performed using Hepatitis C Virus (HCV) NS3-NS4A (PDB id: 1JXP) [20,21], a mixture of NS2B from 2FP7, NS3 from 2FOM, 2FP7 and whole 2IJO [22] as templates. These studies focused on the active site (for competitive inhibition) of the protease. A suitable model for non-competitive inhibition (other than the active site) has yet to be explored.

Virtual screening of a series of small compounds from the ZINC database [23] with backbone structures similar to chalcone, flavanone and flavone were then performed towards the suitable DEN-2 NS2B-NS3 model in an attempt to discover potential non-competitive inhibitors. The selected compounds were then submitted to DEN-2 NS2B-NS3 protease cleavage inhibition assay to validate their activities in vitro. A novel anti-dengue candidate was then obtained from the *in silico* and *in vitro* results.

CASE STUDY DESCRIPTION

Homology Model Building

In this study, nine models of DEN-2 NS2B-NS3 protease, namely 2FOM and eight homology models (DH-1 to DH-8) generated using 2FP7, 2GGV, 2IJO, 3E90, 3L6P, 3LKW, 3U1I and 3U1J as the templates, were evaluated in order to obtain a suitable model for non-competitive inhibition study.

The crystal structures of the templates were obtained from the PDB. The homology models were built based on the amino acid sequence of PDB id: 2FOM.After removing water molecules and the substrates (if present) from the templates, homology models of the DEN-2 NS2B-NS3 protease were generated using Modeller 9.11 software [24]. Amino acid sequences alignment was

performed using Clustal X 2.0 software [25,26]. The sequence alignment showed that there is more than 50% identity between the crystal structures. Hence, the crystal structures are suitable to be used as template for DEN-2 protease homology model building. This is because a protein model that shares more than 30% sequence identity with another protein is indicative of an accurate structure for homology modelling [27]. Homology modelling was carried out by referring to the Modeller online manual [28], and run by using the command line:

mod9. 11 model-default.py

Ten homology models were generated using each of the templates above and were analysed using Ramachandran plots generated by Procheck software, to check the stereochemical quality of the protein structure [29]; and Verify3D software, to determine the compatibility of the 3D atomic models with their own 1D amino acid sequence [30]. These software's are available in their online server versions at the Structural Analysis and Verification Server of UCLA (University of California, Los Angeles; http://nihserver.mbi.ucla.edu/SAVES/). Several homology models were obtained for each template, and a model that produced the best scores in structural analyses, namely DH-1 to DH-8, were then selected for subsequent docking (blind) studies.

Docking of Standard Compounds

Blind docking allowed the ligands to be docked freely to the whole structure of the macromolecule. In this study, blind docking of the ligands into the DEN-2 apo protease (2FOM), and the homology models obtained, was performed using AutoDock 4.2 software. Chlorine atoms, water and glycerol molecules were removed from the 3D crystal structure of DEN-2 NS2B-NS3 apo protease (2FOM) [15] was retrieved from the PDB, and. AutoDock Tools 1.5.4 software was then used to add all hydrogen atoms, merging nonpolar hydrogen atoms, checking and repairing missing atoms, adding Gasteiger charges, checking and fixing total charges on residues, and assigning atom types to the protein structure. A grid box of the protein structure was then generated using AutoGrid 4 software with default atom types (carbon, hydrogen, oxygen and nitrogen), grid spacing of 0.41 *Å,* dimension of 126 x 126 x 126 points along the x, y and z axes, and centered on the protein, covering the whole protein for the blind docking.

As for the homology models, DH-1 to DH-8, docking parameters were set following those described above for 2FOM using the AutoDockTools 1.5.4 software.

Cardamonin, R-pinostrobin and S-pinostrobin were used as standard ligands. Structures (3D) of these ligands were constructed using Hyperchem Pro 8.0 software. The energies of all the ligands were minimized using Hyperchem Pro 8.0 software, employing the steepest descent and conjugate gradient methods (termination conditions set to a maximum of 500 cycles or rms gradient of 0.1 kcal/*Å* mol) [31]. The minimized structures were subsequently prepared with detected root of torsion and number of torsions for flexible-ligand docking using AutodockTools 1.5.4 software and saved as "ligand's name".pdbqt (e.g. cardamonin.pdbqt).

Parameters for blind docking of flexible ligands to DEN-2 protease were set to a population size of 150 individuals, and 10,000,000 number of energy evaluations for 100 runs to produce 100 distinct conformations using the Lamarckian genetic algorithm search function [32]. The resulting 100 distinct conformations were set to be clustered in the same group with RMSD of not more than 0.5 *Å*, for the ease of analysis. The flexible-ligand docking (blind) for each of the ligand was performed by applying all the parameters stored in the docking parameter file, namely "ligand's name".dpf.

AutoGrid 4 and AutoDock 4.2 softwares were installed in a workstation running on Ubuntu 10.04 Linux operating system. Grid map generation was run following instructions in the "AutoDock Version 4.2" user's manual [33] using the command line:

autogrid4 -p protein.gpf -l protein.glg

and docking job was run using the command line:

autodock4 -p ligand.dpf -l ligand.dlg

After the docking jobs were completed, the compounds were ranked based on the lowest estimated mean free energy of docking (ΔG_{dock}) coupled with the largest NumCl (number of conformations in a cluster).The number of distinct conformations that were grouped into the same cluster based on RMSD. ΔG_{dock} was calculated using Autodock 4.2 software, while the estimated inhibition constant (K_i) was calculated using the formula [32]:

$$K_{i\,dock} = e^{\Delta Gdock/RT}$$

where R is the gas constant, 1.987 cal K^{-1} mol^{-1}, and T is the reaction or body temperature, 310.15 K (37 ^{o}C).

(NumCl) was used as a measure of the probability of a particular conformer to interact with the macromolecule target, where the higher NumCl number is proportionate to increased probability of interaction. All of the docked conformers were then analysed for binding interactions using Ligplot 4.5.3 software. The hydrogen bonding distance was set to a range of 2.7 to 3.35 *Å*, and the hydrophobic interaction distance was set to a range of 2.9 to 3.9 *Å* [34]. For selection of conformation with the best binding affinity, the conformation having the largest NumCl and exhibiting interaction with Lys74 from NS3, was selected from the docked conformational cluster. In cases where there were two or more clusters with similar NumCl (difference in NumCl ≤ 10), the cluster that showed lower ΔG_{dock} was chosen. Further interaction analyses using Discovery Studio Visualizer 3.1 (Accelrys Software Inc.) were performed for better insight.

The most suitable DEN-2 NS2B-NS3 protease model for non-competitive inhibition study was then identified following the completion of analyses of results.

Virtual Screening

Autodock 4.2 software was used for virtual screening. The docking input files for the suitable model of DEN-2 NS2B-NS3 protease for non-competitive inhibition were prepared following the method described previously. The docking parameters of the compounds involved were modified to a population size of 150 individuals, and 1,750,000 number of energy evaluations for 20 runs using the Lamarckian genetic algorithm search function. Twenty distinct conformations that were produced were further clustered into the same group (NumCl) with RMSD of not more than 2.0 *Å*. The docking parameters were modified to reduce the duration for running the large number of docking calculations in the virtual screening process. R-pinostrobin was used as the standard since pinostrobin's reported K_{iexp} value (345 ± 70 μM) was smaller than cardamonin's (377 ± 77 μM) [14] and R-pinostrobin produced lower ΔG_{dock} than S-pinostrobin. Redocking of R-pinostrobin was done using the same docking parameters.

A series of small compounds with structures having more than 50% similarity to chalcone (3,458), flavanone (4,886) and flavone (4,997) (Figure 1) were downloaded from the ZINC database [23]. Raccoon 1.0 software [35] was used for the preparation of all the input files for the virtual screening.

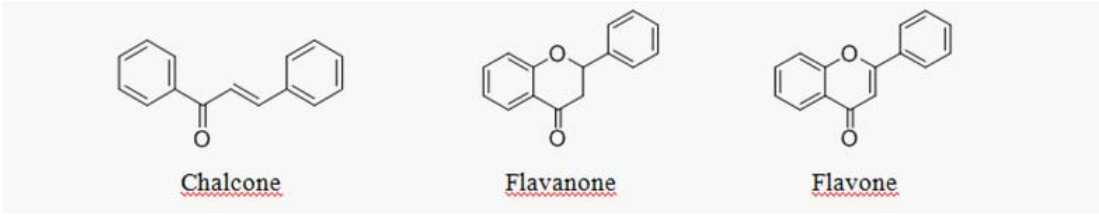


Figure 1: Structures of the chalcone, flavanone and flavone.

Virtual screening was run by using a script file generated by Raccoon 1.0, which sequentially and automatically runs the docking of all the compounds DEN-2 NS2B-NS3 protease using AutoDock 4.2. After completion, the compounds were ranked based on the lowest estimated binding energy with largest NumCl for ease of analysis.

Analysis of *In Silico* Result

Bash 4.1 software [36] was used to program automated sequential data submission, extraction and identification for high-throughput analysis. Compounds with ΔG_{dock} lower than the ΔG_{dock} for both of the standards (cardamonin and pinostrobin) and with NumCl more than 10, were further subjected to interaction analysis using Ligplot 4.5.3 software using the same parameters as described previously. Data about the group, name, NumCl, ΔG_{dock}, K_{idock} value, and interaction properties (hydrogen bonding and hydrophobic interactions) for each compound were extracted. Further interaction analyses using Discovery Studio Visualizer 3.1 were also performed.

Further selection of the compounds was then carried out to identify potential non-competitive inhibitors, based on the lowest ΔG_{dock} and interaction with Lys74 from NS3 [31]. The selected

compounds were then traced from the ZINC database for their availability for purchase. The purchased compounds were then subjected to DEN-2 protease inhibition assay for activity verification.

Verification of *In Silico* Result

From the virtual screening of 13 341 small compounds, 4 were identified to fulfill all the criteria of having ΔG_{dock} lower than that of R-pinostrobin (standard), NumCl more than 10, interacting with the suggested residue in the allosteric binding site (Lys74 from NS3), and are available for purchase (Figure 1). The *in vitro* DEN-2 protease inhibition assay for verification of *in silico* result for this study was performed following methods reported in previous studies [37,38]. Three out of the 4 tested compounds (compounds 1,2 and 4; Figure 2) showed significant better non-competitive inhibition activity when compared to the standard with compound 1 producing the most potent effect (Table 1).

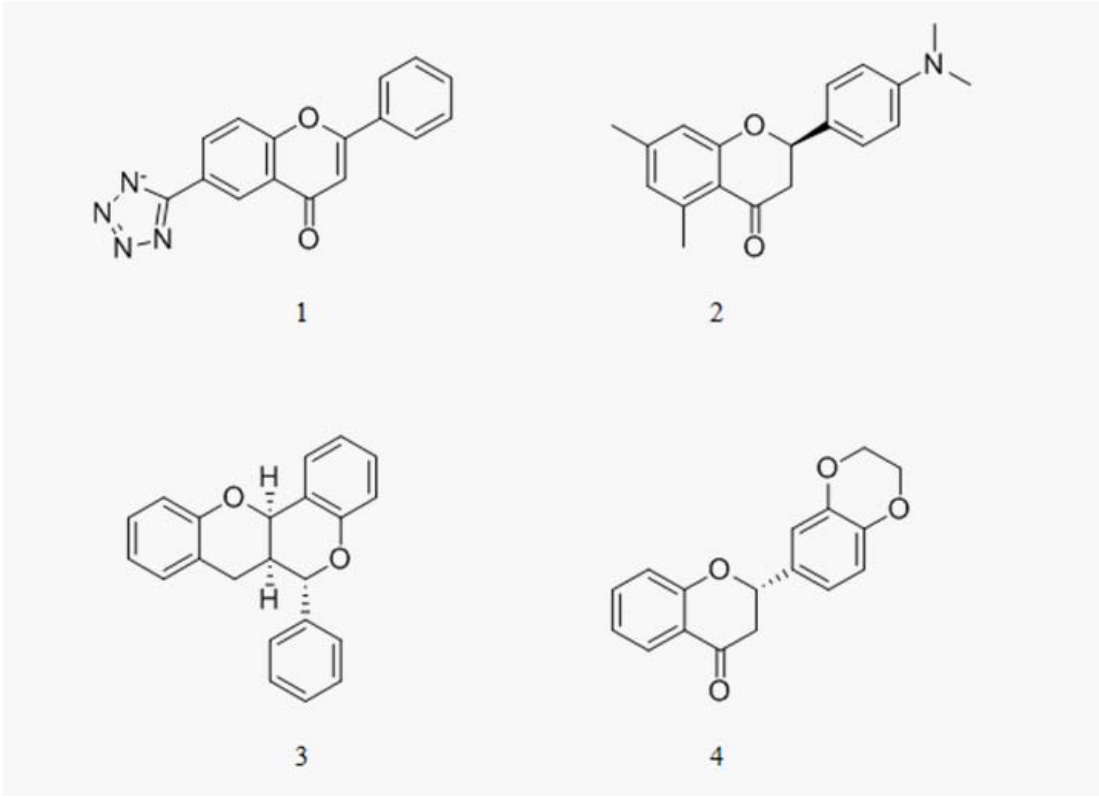


Figure 2: Structures of the small compounds identified from virtual screening against the DH-1 homology model, and were purchased due to their availability.

Table 1: NumCl, ΔG_{dock} and K_{idock} values of the best binding conformations of the small compounds from virtual screening towards DEN-2 NS2B-NS3 protease (homology model DH-1) compared to the K_{iexp} values obtained from protease bioassay in this study.

Compound	Compound Identity	NumCl	ΔG_{dock} (kcal mol^{-1})	$K_{i\ dock}$ (μM)	$K_{i\ exp}$ (μM) in this study
1	2-phenyl-6-(1H-1,2,3,4-tetraazol-5-yl)-4H-chromen-4-one	12/20	-6.17	45	69±9*
2	2-[4-(dimethylamino)phenyl]-5,7-dimethyl-3,4-dihydro-2H-1-benzopyran-4-one	13/20	-5.77	86	121±14*
3	6-phenyl-6a,12a-dihydro-6H,7H-chromeno[4,3-b]chromene	15/20	-5.33	175	510±120
4	2-(2,3-dihydro-1,4-benzodioxin-6-yl)-3,4-dihydro-2H-1-benzopyran-4-one	11/20	-5.29	187	186±38*
Standard	***R*-pinostrobin**	**6/20**	**-4.89**	**358**	**415± 85@**

NumCl = the number of conformations with RMSD < 2.0.

ΔG_{dock} = free energy of binding estimated from AutoDock 4.2 software.

K_{idock} = inhibition constant derived from ΔG_{dock}.

*indicates significant different (p value < 0.05) of unpaired t-tests for K_{iexp} values of compounds 1 - 4 compared with K_{iexp} value of standard pinostrobin.

@value is indicated for pinostrobin (not stereospecific)

CONCLUSION

Virtual screening of potential non-competitive inhibitors for DEN-2 NS2B-NS3 protease was performed yielding several compounds with higher binding affinities than the standard ligand used in this study. The results from *in vitro* inhibition assays supported the *in silico* results obtained. Compound 1 was found to be the best non-competitive inhibitor of DEN-2. This study also proposes that for non-competitive inhibition studies on DEN-2 NS2B-NS3 protease, an appropriate model should exhibit conformation of the allosteric binding site that resembles the homology model DH-1 (2FP7 as template). In conclusion, the rational discovery method described here has potential for use in the discovery of lead compounds for the treatment for dengue, as well as other disease targets.

References

1. Gubler DJ. Dengue and dengue hemorrhagic fever. Clin Microbiol Rev. 1998; 11: 480-496.
2. Guha-Sapir D, Schimmer B. Dengue fever: new paradigms for a changing epidemiology. Emerg Themes Epidemiol. 2005; 2: 1.
3. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW. The global distribution and burden of dengue. Nature. 2013; 496: 504-507.
4. WHO. Dengue and severe dengue fact sheet N°117. World Health Organization. 2015.
5. Simmons JS, St. Johns JH, Reynolds FHK. Experimental studies of dengue. Philipp J Sci. 1931; 44: 1-251.
6. Hammon WM, Rudnick A, Sather GE. Viruses associated with epidemic hemorrhagic fevers of the Philippines and Thailand. Science. 1960; 131: 1102-1103.
7. Gould DJ, Yuill TM, Moussa MA, Simasathien P, Rutledge LC. An insular outbreak of dengue hemorrhagic fever. 3. Identification of vectors and observations on vector ecology. Am J Trop Med Hyg. 1968; 17: 609-618.

8. Rosen L. The global importance of dengue infection and disease. Paper presented at the Proceedings of the International Conference on DHF, Kuala Lampur, Malaysia. University of Malaysia. 1983.

9. WHO. Questions and Answers on Dengue Vaccines: Phase III study of CYD-TDV in Latin America. World Health Organization. 2014.

10. Porcher MH. Multilingual Multiscript Plant Name Database: Sorting Boesenbergia Names. 2003.

11. Ultee AJ. The ethereal oil of Gastrochilus Panduratum. Ridl Verslag Akad Wetensc happen Amsterdam. 1957; 36: 1262-1264.

12. Jaipetch T, Kanghae S, Pancharoen O, Patrick V, Reutrakul V, et al. Constituents of Boesenbergia pandurata (syn. Kaempferia pandurata): Isolation, Crystal Structure and Synthesis of (±)-Boesenbergin A. Aust J Chem. 1982; 35: 351-361.

13. Pancharoen O, Picker K, Reutrakul V, Taylor W, Tuntiwachwuttikul P. Constituents of the Zingiberaceae. X. Diastereomers of [7-Hydroxy-5-Methoxy-2-Methyl-2-(4'-Methylpent-3'-Enyl)-2H-Chromen-8-yl] [3"-Methyl-2'-(3"'-Methylbut-2"'-Enyl]-6"-Phenylcyclohex-3"-Enyl]M Ethanone (Panduratin B), a Constituent of the Red Rhizomes of a Variety of Boesenbergia pandurata. Aust J Chem. 1987; 40: 455-459.

14. Kiat TS, Pippen R, Yusof R, Ibrahim H, Khalid N, Rahman NA. Inhibitory activity of cyclohexenyl chalcone derivatives and flavonoids of finger root, *Boesenbergia rotunda* (L.), towards dengue-2 virus NS3 protease. Bioorg Med Chem Lett. 2006; 16: 3337-3340.

15. Erbel P, Schiering N, D'Arcy A, Renatus M, Kroemer M. Structural basis for the activation of flaviviral NS3 proteases from dengue and West Nile virus. Nat Struct Mol Biol. 2006; 13: 372-373.

16. Aleshin AE, Shiryaev SA, Strongin AY, Liddington RC. Structural evidence for regulation and specificity of flaviviral proteases and evolution of the Flaviviridae fold. Protein Sci. 2007; 16: 795-806.

17. Robin G, Chappell K, Stoermer MJ, Hu SH, Young PR. Structure of West Nile virus NS3 protease: ligand stabilization of the catalytic conformation. J Mol Biol. 2009; 385: 1568-1577.

18. Chandramouli S, Joseph JS, Daudenarde S, Gatchalian J, Cornillez-Ty C. Serotype-specific structural differences in the protease-cofactor complexes of the dengue virus family. J Virol. 2010; 84: 3059-3067.

19. Noble CG, Seh CC, Chao AT, Shi PY. Ligand-bound structures of the dengue virus protease reveal the active conformation. J Virol. 2012; 86: 438-446.

20. Yan Y, Li Y, Munshi S, Sardana V, Cole JL. Complex of NS3 protease and NS4A peptide of BK strain hepatitis C virus: a 2.2 A resolution structure in a hexagonal crystal form. Protein Sci. 1998; 7: 837-847.

21. Lee YK, Rozana O, Habibah AW, Rohana Y, Noorsaadah AR. A Revisit into the DEN2 NS2B/NS3 Virus Protease Homology Model: Structural Verification and Comparison with Crystal Structure of HCV NS3/4A and DEN2 NS3. Malays J Sci. 2006; 25: 15-22.

22. Wichapong K, Pianwanit S, Sippl W, Kokpol S. Homology modeling and molecular dynamics simulations of Dengue virus NS2B/NS3 protease: insight into molecular interaction. J Mol Recognit. 2010; 23: 283-300.

23. Irwin JJ, Shoichet BK. ZINC-a free database of commercially available compounds for virtual screening. J Chem Inf Model. 2005; 45: 177-182.

24. Sali A, Blundell TL. Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol. 1993; 234: 779-815.

25. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 1997; 25: 4876-4882.

26. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA. Clustal W and Clustal X version 2.0. Bioinformatics. 2007; 23: 2947-2948.

27. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, et al. Comparative Protein Structure Modeling Using Modeller. Curr Protoc Bioinformatics. 2006; 5: Unit-5.6.

28. Webb B, Madhusudhan MS, Shen MY, Marti-Renom MA, Eswar N, Alber F, et al. MODELLER A Program for Protein Structure Modeling Release 9.15, r10497. 2015.

29. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr. 1993; 26: 283-291.

30. Lüthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature. 1992; 356: 83-85.

31. Othman R, Kiat TS, Khalid N, Yusof R, Newhouse EI, Newhouse JS, et al. Docking of noncompetitive inhibitors into dengue virus type 2 protease: understanding the interactions with allosteric binding sites. J Chem Inf Model. 2008; 48: 1582-1591.

32. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comp Chem. 1998; 19: 1639-1662.

33. Morris GM, Goodsell DS, Pique ME, Lindstrom WL, Huey R, et al. AutoDock Version 4.2. Automated Docking of Flexible Ligands to Flexible Receptors. User Guide, the Scripps Research Institute, La Jolla, CA. 2010.

34. Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng. 1995; 8: 127-134.

35. Forli S. Raccoon. AutoDock VS: an automated tool for preparing AutoDock virtual screenings (Version 1.0). 2010.

36. Fox B. Bash, the GNU Bourne-Again Shell (Version 4.1). 1989.

37. Yusof R, Clum S, Wetzel M, Murthy HM, Padmanabhan R. Purified NS2B/NS3 serine protease of dengue virus type 2 exhibits cofactor NS2B dependence for cleavage of substrates with dibasic amino acids in vitro. J Biol Chem. 2000; 275: 9963-9969.

38. Tomlinson SM, Watowich SJ. Substrate inhibition kinetic model for West Nile virus NS2B-NS3 protease. Biochemistry. 2008; 47: 11763-11770.

Austin Publishing Group

Comparative Evaluation of Docking Programs: A Case Study with Small Peptidic Ligands

Kumar M[1], Tiwari P[1] and Kaur P[1]*

[1]Department of Biophysics, All India Institute of Medical Sciences, India

***Corresponding author:** Punit Kaur, Department of Biophysics, All India Institute of Medical Sciences, New Delhi 110 029, India, Email: punitkaur1@hotmail.com

Published Date: December 19, 2016

INTRODUCTION

Design and discovery of novel drug lead molecules is of critical importance in human health care. The detailed and in-depth information about how drug compounds interact with the protein targets is essential for the development of newer candidate drug molecules. Computational approaches have emerged to be extremely relevant in rational drug design and have been embraced for developing and determining the molecule which would be most suitable for entering the drug development pipeline. This involves the *in silico* modeling of the designed lead compounds into the active site of the target protein for their best fit considering both steric aspects and functional group interactions so as to predict its activity and binding affinity towards the target protein. This approach is referred to as protein-ligand or molecular docking. Molecular docking is thus today, a well developed computational method to predict the most

probable binding mode/pose (position, conformation & orientation) of a small molecule (ligand) in the binding pocket or active site of a macromolecule (protein receptor) based on their spatial arrangement and chemical complementarity in order to maximize the interactions between them. This *in silico* technique thus can act as a three-dimensional **(3D)** filter to separate the probable hits (compounds that bind to the target and have the desired effect) and non-hits from large chemical compound libraries based on their shape and size and distribution of charges on surface of ligands complementary to a defined binding pocket. This computational weeding of ligands is known as Virtual Screening **(VS)** in contrast to *in vitro* screening of ligands referred to as High Throughput Screening **(HTS) [**1]. This method facilitates (i) enrichment of the chemical library by pin pointing newer druggable compounds and (ii) scaling down the ligand compound dataset for *in vitro* testing (chemical screening) resulting in cost reduction. Molecular docking today has gradually become an indispensible technique in the drug discovery pipeline for the identification of potential lead compounds due to its rapidity and cost-effectiveness **[**2].

Therapeutic agents/drugs are generally organic compounds that interfere with the biological activity of the target protein or nucleic acid. They include a rigid scaffold and contribute to the development of resistance [3] due to mutations in the target or efflux/influx genes especially in rapidly growing cells like cancer cells or in microbes. Multi-drug resistance is a cause of concern around the world and has necessitated the requirement for newer potent drugs. Consequently, the pharmaceutical industry is continuously on the lookout for novel mutation resilient drugs. Peptides form ideal candidates in this quest [4] for they are less toxic in comparison to organic compounds as they mimic the natural components in the human body. They are comparatively more target specific as they have higher affinity for a protein target leading to improved potency and efficacy. Moreover, peptides possess greater conformational flexibility than the comparatively rigid core of presently available drug leads and hence expected to be less prone to resistance. Peptides are also able to generate tremendous diversity due to possibility of twenty different constituent amino acids. Limitation of role of peptides as drug lead compounds is their lower stability at low pH of gut. However, modified peptides can provide stable lead compounds. Hence, specificity, conformational flexibility and presence of a well-developed elimination system in the host indicate that small peptides can form an ideal class of potential lead compounds for the development of potent drugs.

In vitro screening is an expensive exercise as it would require the synthesis of a large number of peptides of variable lengths and composition. Hence, molecular docking based *in silico* screening of peptide libraries would tremendously reduce the cost and filter out unlikely hits. Various molecular docking programs have been developed, optimized and evaluated mainly for the screening of chemical libraries of organic compounds. The reports on comparative evaluation of docking programs with regard to organic ligands [5-11] indicate that their predictive power in combination with various scoring functions varies from one to another with respect to the class of receptor proteins as well as flexibility of ligands. Therefore, a careful selection of the docking

program is an important criterion before initiating the computational analysis for peptide ligands. To date, a comparative analysis and evaluation of docking programs with regards to peptide ligands is not available. The objective of this case study is to assess the predictive power of three commonly used docking programs for peptide ligand docking to facilitate the choice of a relatively more reliable program. These programs include the open source (AutoDock) [12] and two softwares on the commercial platform (GOLD [13] and Glide [14]).

MOLECULAR DOCKING

Molecular docking approach concerns the (i) sampling of conformational, rotational and translational space of the ligand in the binding pocket of the protein and subsequently (Figure 1a and 1b) (ii) ranking the potential solutions generated in the form of docked poses (positional conformers) based on their appropriate placement in the active site of the protein [15]. A molecular docking program comprises two components; a search algorithm to generate the conformations and subsequently dock them in the binding pocket and a scoring function to rank those docked conformations. Search algorithms basically follow either a systematic or a random approach to produce diverse sterically feasible conformations of the ligands in the binding pocket by varying the flexible/rotatable bonds. This generates a large number of possible conformations for a single ligand. All these probable conformations are subsequently docked into the binding pocket of the protein to arrive at the best fit conformation which will yield the highest affinity and activity towards the target protein. Systematic search algorithms look for all the conformations of the ligand based on torsional degree of freedom in a stepwise manner according to defined parameters. This approach includes an incremental construction based search algorithm (implemented in the program DOCK [16,17] and Glide [14]) and molecular dynamics simulation based search algorithm (CDocker [18]). Random search algorithms explore conformations by varying the flexible torsions randomly. Genetic algorithm (AutoDock [12], GOLD [19]) and Monte Carlo simulation based algorithm (LigandFit20) are based on the random search approach. Out of the two, systematic search is comparatively more exhaustive and deterministic as it explores all feasible potential conformations in a stepwise manner and hence has a lesser probability of missing a true solution or conformation. As a result this approach is computationally extensive and time consuming. In contrast, the random search methods are rapid but non-deterministic and may overlook a correct potential conformation. Most docking programs utilize a combination of both these strategies, eg the programs DOCK and Glide initially involve an exhaustive incremental search followed by refinement with Monte Carlo simulations.

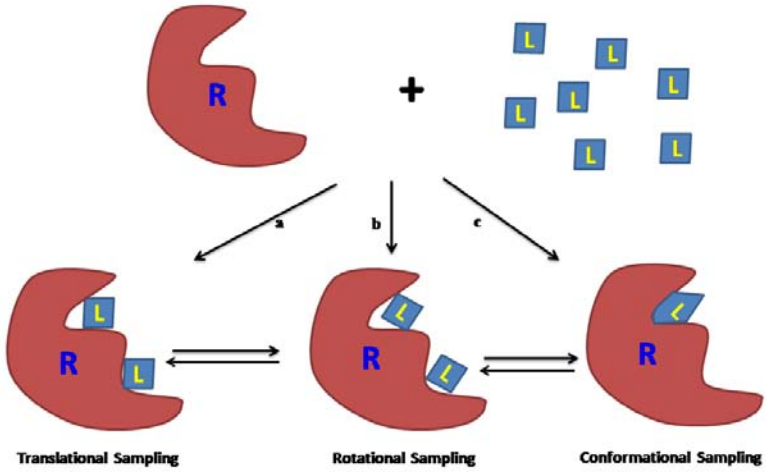


Figure 1a: Cartoon representation of molecular docking of receptor (R) and ligand (L). (a) Ligand translates in the receptor binding pocket looking for complementary region and (b) reorients in the suitable binding pockets by rotation in all three dimensions and (c) simultaneously conformation of ligand changes to maximize the interactions between them.

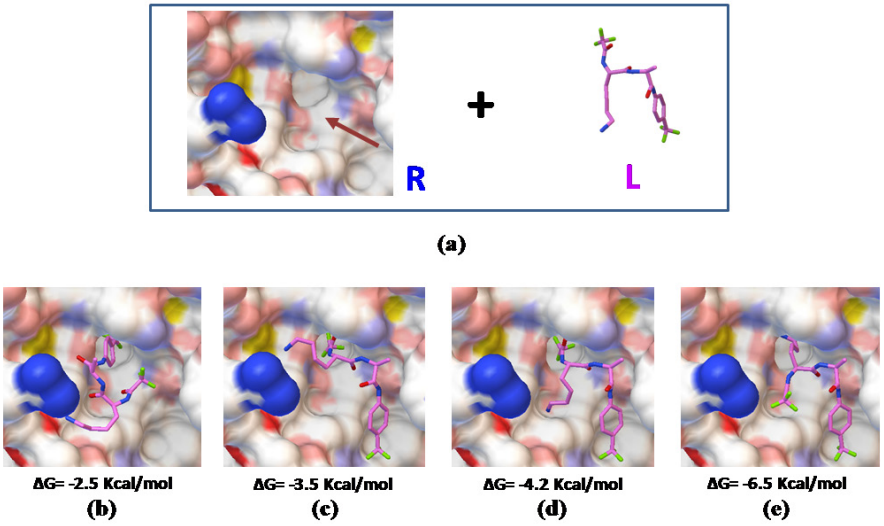


Figure 1b: Representation of (a) molecular docking of ligand (L) in receptor (R). The probable conformations adopted by the ligand on translation, rotation and subsequently docking in the binding site pocket are shown in (b), (c), (d) and (e). The docked conformation (e) has the best free energy of binding and is the most probable mode of binding.

Molecular Docking utilize scoring functions to rank the generated docked conformations. Scoring functions can be classified into three categories namely force field based, empirical and knowledge based scoring functions. Force field based scoring functions (DockScore [16], GoldScore

[19], CDocker Energy [18], G-Score [21], D-Score [21]) estimate the actual interactions between receptor and ligands based on force field parameters and thus rank these docked poses without any bias. However, the score obtained can vary from force field to force field due to differences in parameters and/or method employed to estimate the ligand-protein interactions. The major disadvantage of this scoring function is that it is solely dependent on enthalpic contribution of binding and does not include the entropic and solvation/desolvation contributions which play a major role in the binding process. As a result, it is unable to estimate the free energy of binding ($\Delta G = \Delta H - T\Delta S$) or binding constants ($\Delta G = -nRTlnKd$). Empirical scoring functions (LigScore [22], LUDI score [23,24], ChemScore [25], F-Score [26], X-Score [27]) rank the solutions according to the estimated free energy (ΔG) of binding between the receptor and docked ligand pose based on empirically derived energy function. This takes into account enthalpic as well as entropy and solvation/desolvation contributions on binding and hence provides a quantitative estimate of binding strength. During this form of scoring process, actual enthalpic contributions (like ionic interactions, hydrogen bond, aromatic interactions lipophilic interactions,) as well as entropic contributions (depending on number of rotatable torsions) are calculated and then each of them is multiplied by respective regression coefficient. Similarly a regression co-efficient is also calculated for solvation effect. The regression coefficient for each of the contributing factor in ligand binding is derived from the dataset of protein-ligand complexes to correlate/fit the predicted binding constant with their experimental values. The accuracy of predicted binding constant in empirical scoring is dependent on the experimental dataset used for regression analysis to derive the energy function and can therefore, be biased. Consequently, empirical scoring function may fail to predict accurately the binding constant for ligands whose similar ligands are not present in dataset used to calculate the regression co-efficient. Knowledge-based scoring functions (PMF [28,29], ASP [30], DrugScore [31], SMoG [32]) calculate the statistically weighted atom pair potentials between ligand and receptor protein derived from knowledge base of 3D structures of receptor-ligand complexes (PDB [33], CSD [34]). Since these are derived from actual experimental structures they include the entire repertoire of contributing factors to ligand binding unlike force field based scoring function. Neither are these empirically derived from a selected dataset but from part of an entire database and when compared to empirical scoring functions are less likely to be biased. Many commonly used docking programs like AutoDock and Glide use semi-empirical scoring functions that are hybrid of empirical and force field based scoring functions. Scoring functions encompasses both merits and limitations and might possibly complement each other. Therefore, a consensus of all scoring functions could prove to be a better strategy to rank the docked poses. In a majority of cases docking programs generate the correct solution but subsequently fail to rank them correctly partly due to imperfect scoring functions. The primary role of scoring functions is to rank the generated poses to identify the best pose amongst them based on their highest score. All three kinds of scoring functions are capable of performing this action. Their secondary role is to estimate the quantitative strength of binding (binding affinity) and only empirical (including semi-empirical) class of scoring function is able to calculate this reliably due to inclusion of entropic and solvation parameters along with enthalpy.

The predictive power of the binding mode of ligand depends on the ability of docking program to generate and replicate the native crystal orientation and geometry of ligand as observed in the protein 3D structure. Similarly predictive power of docking programs to calculate the binding affinity of ligand towards the protein after protein-ligand docking depends on its potential to correctly reproduce the experimental activity data. These values are dependent on the robustness of search algorithms as well as scoring functions. Knowledge based and force field based scoring functions provide an idea about the comparative binding strength (qualitative estimate to differentiate a ligand with higher affinity to the protein from another with low affinity) amongst different ligands docked in the same binding pocket but are unable to assess the binding constant as they do not calculate the free energy of binding. On the other hand, the empirical scoring functions have the ability to predict the binding constant (quantitative estimate like 9.4 kcal/mol). However, these may fall short in cases where ligand is dissimilar from ligands in dataset used to derive the regression coefficient. A more accurate way to calculate the free energy of ligand binding in receptor complex are molecular dynamics based approaches of Free Energy Perturbation **(FEP)** and Thermodynamic Integration **(TI)**35,36. These are based on statistical thermodynamic ensembles which take appropriate account of solvation due to presence of explicit water molecules. The accuracy of calculated free energy of binding is dependent on the versatility (parameterization with large variety of small molecules, ability to handle the metal ions and effect of polarization) of the force field and conformational sampling ability of molecular dynamics simulation. The force fields employed in the docking programs are well parameterized for proteins but do not contain accurate parameters for millions of non-peptidic drug-like compounds. The major drawback of this method is that it is computationally too expensive (more than 1000-times in comparison to empirical scoring function) to be used for virtual screening of chemical library of millions of compounds.

METHODOLOGY

The first report for prediction of ligand binding in the active site pocket of receptor through computational modeling approach utilized the DOCK program [37]. Since then several docking approaches both open source and commercial eg DOCK, AutoDock, GOLD, Glide, FlexX [26], CDocker [18], LigandFit [20], Hex [38] have been developed. The various programs are being continuously improved to better their performance and accuracy. Peptide-based drugs are being explored as possible therapeutic agents, and presently a comparative study on docking programs is unavailable. Therefore, an analysis of the capability of three molecular docking programs to correctly dock and assess the peptide at the active site of the protein was undertaken. The strategy comprises re-docking the peptide into the protein binding pocket and evaluating the capability of the docking program to correctly predict the orientation and peptide–protein interactions when compared to that in the already determined experimental peptide-protein 3D structure available in the PDB [33]. The docking programs selected for this study are GOLD, AutoDOCK and GLIDE as they have different class of docking algorithms.

Programs

GOLD (Genetic Optimisation for Ligand Docking) is a commercial but comparatively low cost docking program considered to be the gold standard among docking programs. Its conformational search approach is based on Genetic algorithm similar to AutoDock. Genetic algorithm is an evolutionary strategy to explore the conformational flexibility of the ligand while simultaneously sampling the binding pocket. It uses empirical scoring function to rank the generated docked poses.

AutoDock 4 is an open source docking program freely available to download. It is based on evolutionary algorithm (Lamarckian Genetic algorithm, a reverse variant of Genetic algorithm) for conformational search approach combined with semi-empirical scoring function which calculates free energy of binding between ligand and receptor based on precalculated grids. It also provides Monte Carlo simulated annealing option for conformational sampling of ligand. It is one of the more popular open source docking programs.

Glide is also a commercial docking program available from Schrodinger Incorporation. It uses systematic search algorithm incremental construction for conformational sampling of ligand with partial flexibility of receptor and utilizes a semi-empirical scoring function to calculate the free energy of binding for the ranking of generated docking conformations. Induced fit Glide docking also provides limited flexibility of receptor in loop region when used in combination of protein modeling program PRIME.

Dataset

Peptide Binding Protein Database (PepBind) [39] is a derived database that contains structures of protein-peptide complex available in PDB. At least 8 representative structures each of peptides ranging from dipeptide to nonapeptide (including modified peptides) that are non-covalently complexed with proteins belonging to different protein families were included in this study (Table 1). These peptides contain 9 to 41 rotatable torsion angles whereas the limitation of docking of programs to handle the conformational sampling in generally restricted to 20. Different protein families were chosen as these would have active sites with variable characteristics in order to sample a larger set of shape and size of binding pockets for docking of peptide ligands. This acquired dataset was utilized to evaluate the comparative ability of the three docking programs (AutoDock, GOLD and Glide) to reproduce the native geometry of the peptide as observed in the crystal structure.

Table 1: Comparison of docking of peptides with the experimental crystal structure.

Serial no.	PDB ID	Protein Family	Root mean squared deviation#		
			Glide	Auto Dock	GOLD
Dipeptide					
1	1A16	Creatinase/ N-terminal domain	0.87[1]	0.73[1]	4.72[1]
2	3TMN	Thermolysin-like	0.69[1]	3.00[6]	7.67[5]
3	4D2C	POT family	0.86[1]	1.20[1]	1.19[8]
4	1HSB	Class I Histocompatibility antigen	1.15[1]	2.48[6]	7.8[1]
5	1ELE	Pancreatic elastase	2.61[1]	0.90[1]	0.80[1]
6	1ELD	Pancreatic elastase	0.82[1]	1.20[1]	0.98[10]
7	1TMN	Thermolysin	0.41[1]	2.40[1]	6.66[1]
8	2EST	Elatase	0.57[1]	1.47[1]	1.28[10]
Tripeptide					
9	1FN8	Eukaryotic proteases	0.61[2]	1.81[5]	1.15[4]
10	1HSB	Class I Histocompatibility antigen	0.45[2]	0.25[1]	0.70[1]
11	2B6N	Subtilase	0.71[1]	3.72[2]	8.70[3]
12	1A30	Retroviral protease	1.71[1]	2.20[6]	2.61[2]
13	1BS6	Peptide deformylase	1.13[1]	0.87[1]	1.84[5]
14	1OOK	Eukaryotic proteases	1.22[1]	0.79[4]	0.80[1]
15	2H9T	Thrombin	0.80[1]	0.58[4]	0.73[1]
16	1XVM	Trypsin	0.55[3]	1.25[1]	0.65[10]
17	1ZY1	Polypeptide deformylase	0.81[1]	0.55[1]	0.73[2]
18	2V3X	Aminopeptidase P	1.26[2]	0.83[1]	1.19[8]
Tetrapeptide					
19	2I3H	Inhibitor of Apoptosis domain	0.53[1]	5.03[3]	0.42[10]
20	1TW6	Inhibitor of Apoptosis domain	0.55[2]	0.89[2]	0.48[1]
21	1W9E	PDZ domain	4.70[2]	1.27[1]	7.37[1]
22	2O1N	Phospholipase A2	3.96[1]	2.84[1]	4.58[7]
23	2NPH	Retroviral aspartyl protease	4.50[1]	3.55[10]	2.11[7]
24	1DKY	Hsp70 protein	1.97[7]	1.97[10]	1.96[10]
25	3BRH	Protein-tyrosine phosphatase	2.26[2]	3.37[1]	8.28[1]
26	1UOO	Prolyloligopeptidase	1.02[1]	1.70[2]	1.35[3]
27	2Y1L	Ankyrin repeats	1.89[7]	1.34[1]	0.93[7]
28	1FCH	Tetratricopeptide repeat	0.82[6]	3.64[8]	1.23[3]
Pentapeptide					
29	1EVH	RanBP1 domain	2.04[1]	0.48[1]	0.74[8]
30	1BE9	PDZ domain	0.79[1]	0.94[2]	1.39[1]
31	1BHX	Eukaryotic proteases	1.80[1]	1.72[1]	8.88[1]
32	1NVR	Protein kinase	2.79[1]	0.77[1]	4.44[8]

33	2B1N	Papain family cysteine protease	0.99[2]	1.82[1]	4.27[6]
34	2PCU	Pancreatic carboxypeptidases	6.06[10]	4.01[9]	1.29[9]
35	2Z3N	Leucyl/phenylalanyl-tRNA protein transferase	6.60[1]	1.08[1]	0.63[1]
36	2QL5	Caspase domain	6.59[4]	1.68[1]	5.55[4]
37	1OKV	Protein kinase	3.95[9]	1.73[1]	7.47[2]
38	3CBL	Tyrosine protein kinase	1.05[1]	3.26[1]	9.70[6]
39	2HPL	PUB domain	1.39[1]	0.76[1]	0.94[1]
Hexapeptide					
40	3D9T	Inhibitor of Apoptosis domain	1.72[2]	3.03[1]	0.47[1]
41	2ZGH	Trypsin	1.33[1]	0.61[2]	1.68[1]
42	1TP5	PDZ domain	1.76[3]	1.33[1]	7.66[1]
43	1E8N	Prolyloligopeptidase	6.29[3]	0.45[1]	5.21[9]
44	2MIP	Retroviral protease	8.25[4]	1.66[4]	1.34[1]
45	3DRI	Bacterial extracellular solute-binding	1.19[2]	1.22[6]	3.18[1]
46	1OL1	Protein kinase	1.46[1]	2.53[1]	4.59[5]
47	1AWU	Cyclophilinpeptidylprolylisomerase	2.23[3]	1.13[1]	1.80[2]
48	1JW6	Legume lectin	8.11[7]	3.69[4]	10.19[4]
49	1KL3	Avidin family	7.14[2]	1.42[3]	5.48[6]
50	1JK4	Neurophysin II	3.53[1]	1.82[1]	1.53[8]
51	2DS8	ClpX chaperone zinc binding domain	2.91[1]	1.14[7]	1.36[1]
52	2FOP	MATH domain	1.40[1]	1.22[7]	1.82[1]
53	3DDA	Clostridial neurotoxin zinc protease	5.71[1]	1.05[2]	4.03[1]
54	1AWR	Cyclophilinpeptidylprolylisomerase	2.78[1]	1.45[5]	3.00[1]
Heptapeptide					
55	2ZGJ	Trypsin	1.69[3]	0.88[6]	1.67[1]
56	1WBP	Protein kinase	8.22[9]	1.59[2]	7.71[6]
57	1E4X	Anti-TGF-α antibody Fab fragment	3.56[1]	0.60[5]	6.00[1]
58	1CZY	MATH domain	1.20[1]	1.21[4]	1.91[1]
59	3HBV	Secreted protease C	3.29[7]	0.65[1]	1.99[1]
60	1P7W	Subtilase	4.13[10]	1.35[1]	1.65[1]
61	3CVQ	Tetratricopeptide repeat	2.37[2]	1.95[1]	8.14[9]
62	2PV1	FKBP immunophilin/prolineisomerase	3.01[1]	1.15[1]	4.15[1]
63	1DKX	Heat shock protein 70kD (HSP70)	4.25[9]	1.30[3]	7.21[9]
64	1P7V	Subtilase	8.05[9]	1.46[1]	1.93[2]
Octapeptide					
65	2PW1	2F5 Fab fragment heavy chain	7.33[5]	1.99[2]	7.87[5]
66	1ELW	Tetratricopeptide repeat	1.86[2]	0.98[1]	3.29[2]
67	1D01	MATH domain	0.93[1]	0.85[1]	6.31[3]
68	2DP4	Subtilase	5.66[1]	2.68[2]	9.02[7]
69	3BOO	Clostridial neurotoxin	7.19[5]	2.07[1]	6.13[8]

70	1IID	N-myristoyltransferase	3.29[1]	1.45[1]	3.14[5]
71	1ZT1	Immunoglobulin C1-set domain	0.60[1]	1.06[1]	3.87[9]
72	2GTZ	Immunoglobulin C1-set domain	1.56[1]	0.85[1]	0.70[1]
73	2HD4	Subtilase	6.12[2]	3.07[1]	4.47[6]
74	3JQ5	Phospholipase A2	8.26[2]	2.52[1]	6.71[10]
Nonapeptide					
75	3DRG	Bacterial extracellular proteins	2.19[7]	3.32[6]	3.29[6]
76	1H24	Protein kinase	6.00[1]	3.64[8]	7.91[8]
77	1MFG	PDZ domain	2.21[1]	0.96[4]	4.88[8]
78	1SB1	Eukaryotic proteases	7.45[5]	1.76[2]	8.31[3]
79	1F7A	Retroviral protease	1.75[5]	1.57[5]	1.45[1]
80	3RM1	EF-hand domain pair	6.88[8]	1.26[1]	8.95[6]
81	1U00	Heat shock protein 70kD	3.11[9]	0.73[1]	5.96[1]
82	3BXN	Immunoglobulin C1-set domain	0.90[1]	0.54[3]	4.15[2]
83	2H6P	Immunoglobulin C1-set domain	4.99[3]	0.78[1]	7.05[2]
84	1AO7	Immunoglobulin C1-set domain	12.03[1]	0.68[1]	20.71[10]
Number (Percentage) of Successfully Docked Peptides in the 84 Protein families *			**42(50%)**	**63(75%)**	**37(44%)**

Preparation and Docking Process

Peptide/modified peptides were extracted from their respective protein complexes taken from PDB and used as ligands. Water molecules and other co-crystallized salts/ions or molecules were removed from the complex and the resulting cleaned proteins were treated as receptor. Force field parameters (AutoDock 4 for AutoDock; SYBYL for GOLD and OPLS-2005 for Glide) and hydrogen atoms were added on both receptor proteins and ligands. This is a routine process in molecular modeling and simulation process including docking because calculations of all the energetics are based on these parameters. Hydrogen atoms were added because PDB files lack hydrogen atoms and this is required for hydrogen bond estimation in protein-ligand interactions. This was followed by a short energy minimization of receptor proteins and peptide ligands to remove any prevailing steric clashes. The binding site for docking studies was defined by generating the grid box around the ligands. Receptor proteins were kept rigid during docking process while ligands were taken to be flexible. This was done to ensure that unrestricted conformational sampling of ligand was performed by docking programs due to the presence of rotatable torsions. Since the peptide bond in the ligands is a partial double bond, this was treated as non-rotatable by applying constraints/restraints during docking process. As a result, the total number of rotatable bonds in docking set of peptidic ligands decreased to 2 to 24 (from 9 to 41) had were within the permissible limits of the docking programs. The default parameters in built in the programs were used for both docking and scoring the ligand. GlideScore was used as scoring function for Glide; ΔG was calculated for AutoDock and GoldScore for GOLD. All the docking studies were performed employing potential grid for calculations of non-covalent interactions between receptor and ligands.

RESULTS & DISCUSSION

Performance of evaluated docking programs has been analyzed according to their ability to reproduce the crystal geometry of peptide ligands in combination with their default scoring functions. Performance efficiency of docking programs varied depending on the size of peptides (ranging from di to nonapeptide). The smaller the peptide length, the greater was the accuracy of docking. The number of rotatable bonds increases with the size of peptide ligands and affect the docking precision since ligands with higher number of rotatable bonds have greater conformational flexibility/sampling space. Most docking programs have the ability to handle a limited number of rotatable bonds. GOLD and Glide can accommodate upto 20 rotatable bonds competently while it is reported that AutoDock works most efficiently upto 10 rotatable bonds. External constraints on peptide backbone in ligands also limit its torsional degree of freedom during conformational sampling. The number of rotatable bonds in the ligand dataset in this study was between 2 to 24 depending on the peptide length.

Top ten docked poses (potential solutions) of each peptide ligand for their respective protein were generated and subsequently ranked by all the three docking programs. The best pose among the ten was obtained by superimposing each of the ten docked poses onto the bound peptide conformation in the protein complex structure determined experimentally. Accuracy of docking result was determined by calculating the root mean squared deviation (r.m.s.d) of docked conformations with respect to crystal conformation of each ligand (Figure 2 & Table 1). The closer the predicted pose was to the bound peptide ligand in the crystal structure, lesser was the deviation obtained in the calculated positional r.m.s.d. The docked pose (both conformational and positional) with r.m.s.d of less than 2.0 $\AA$ with respect to crystal conformation was considered a successful docking since resolution of crystal structures present in the dataset was about 2 $\AA$.

Neither of the three program predicted all the poses correctly for all the peptides used in the study. AutoDock was able to reproduce successful docking mode in 75% of cases, followed by Glide (50%) and GOLD (44%) (Table 1 and Figure 3a). Glide and GOLD indicated a higher rate of successful docking for smaller peptides as compared to larger peptides. GOLD was able to give only 10% success rate for docking of octa or nonapeptide ligands. AutoDock outperformed Glide and GOLD remarkably for peptide ligands with 5 to 9 amino acid residues demonstrating that AutoDOCK has the capability to dock rotatable bonds of more than 10. Besides these comparative observations between the three docking programs, there were two other noteworthy observations. First, it has been observed that there were cases where Glide and AutoDock failed to provide the correct binding mode of ligands, but GOLD produced the correct pose in spite of an overall poor performance among these three docking programs. Similarly, though AutoDock outperformed Glide, there were cases where AutoDock failed but Glide provided a successful docking Table 1. This indicates that no single docking program is perfect for the docking of peptide ligands and the programs have to be chosen judiciously. Moreover, the various programs can complement each

other when used in combination. The second significant observation was that though docking programs produced the correct docked solution (binding mode), but the scoring functions were unable (almost 50% of the time) to rank the potential solutions suitably (Table 1 and Figure 3b). This clearly indicates that scoring functions are lacking in comparison to docking function. This is probably because generating possible conformations and sampling the potential binding space are less complex than calculating the binding interactions between the protein and ligand. Chemistry of binding of ligands not only depends on the enthalpic contribution but also on entropic factor that is difficult to evaluate. However, identification of the best solution among generated poses can be improved by consensus scoring wherein different scoring functions can be used to score one particular docking. Rescoring the generated docked solutions/pose with several other scoring functions of all three classes and subsequently arriving at the consensus based solution might increase the reliability of obtaining the most accurate solution closest to that derived experimentally. This study clearly indicates that docking and scoring unless used judiciously can point towards an erroneous solution as the better result. Therefore, validation of the docking protocol with a known experimental receptor and ligand structure is essential before initiating virtual screening of chemical libraries. Validation will assist in the recognition of a more appropriate docking program and scoring function to execute the docking studies with unknown ligands.

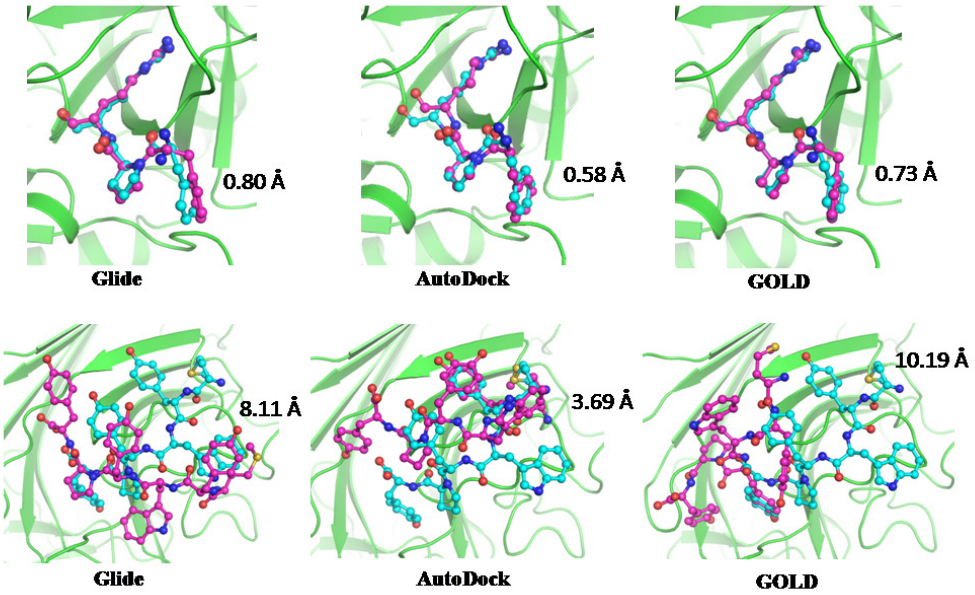


Figure 2: Predicted Best (upper panel) and worst (lower panel) docking result of the three programs along with calculated r.m.s deviation between docked conformation (ball and stick in magenta) and crystal conformation of ligands (ball and stick in cyan). Crystal structure used for docking validation is respective PDB: 2H9T (for upper panel) and PDB: 1JW6 (for lower panel).

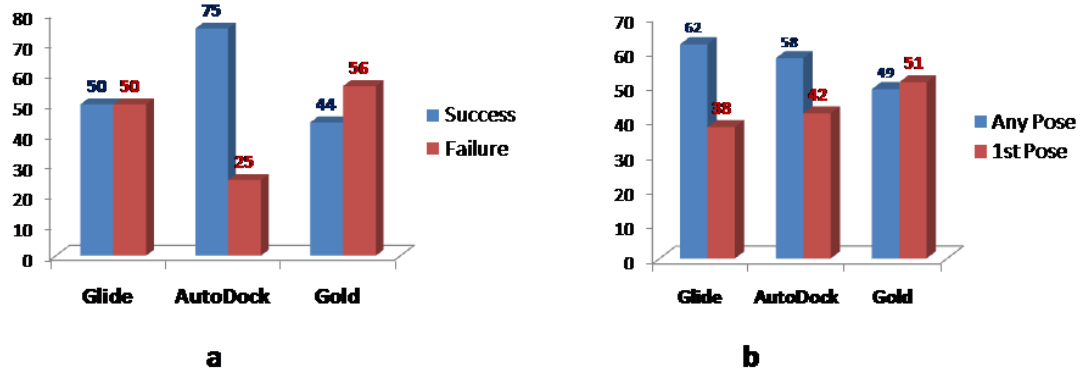


Figure 3: Bar diagram representation of docking experiments with 84 peptide complex. (a) represents overall success (%) of docking programs individually and (b) represents correct ranking (%) among successful docking only.

CONCLUSION

Molecular docking is a handy computational method to predict the binding mode as well as binding strength of a ligand with its receptor when their 3D structures are available which if discreetly employed can result in significant savings of time and cost. This study on the analyses of the predictive power of binding mode for peptidic ligands in three docking programs reiterates earlier studies carried out with small organic ligands except for the significantly higher success rate obtained with AutoDock as compared to Glide and GOLD. This might be due to the rigid peptide bond present in peptide ligands that allows lesser conformational sampling and the ability of AutoDock to correctly dock the peptide into its respective protein molecule and score it efficiently. Ranking of docked poses by scoring function was observed to be inaccurate in about 50% cases even though docking was successful wherein the docked pose matching the experimental structure was ranked much lower. This clearly points out that though the correct pose was present among the ten poses generated by the docking programs but scoring function did not recognize it. Therefore, users must be skeptical about the best ranked docked pose indicated by scoring function. Re-ranking of the docked poses generated by docking program with several scoring functions of different classes followed by identification of the best pose from consensus scores could improve the success rate. This study also reveals that no single docking program is ideal and they complement each other due to incorporation and dependency of different docking and scoring approaches. It is also extremely relevant to validate and verify the docking programs to suitably reproduce the known experimental data before proceeding with docking experiments. This would improve the reliability of the predictive power of docking programs and ultimately result in higher reliability of the retrieved ligands from screened chemical libraries. However, notwithstanding the inherent limitations of the molecular docking programs, these programs are extremely useful in the drug discovery pipeline as they not only screen out probable lead molecules from databases but simultaneously contribute information about the protein-ligand interactions and their binding mode essential for de novo drug design and discovery.

ACKNOWLEDGEMENT

This study was performed with hardware and software supported by funds from Indian Council of Medical Research, New Delhi in the form of 'Bio-Medical Informatics Centre'.

References

1. Bajorath J. Integration of virtual and high-throughput screening. Nat Rev Drug Discov. 2002; 1: 882-894.
2. McInnes C. Virtual screening strategies in drug discovery. Curr Opin Chem Biol. 2007; 11: 494-502.
3. Blair JM, Webber MA, Baylay AJ, Ogbolu DO, Piddock LJ. Molecular mechanisms of antibiotic resistance. Nat Rev Microbiol. 2015; 13: 42-51.
4. Otvos L. Peptide-based drug design: here and now. Methods Mol Biol. 2008; 494: 1-8.
5. Bissantz C, Folkers G, Rognan D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem. 2000; 43: 4759-4767.
6. Bursulaya BD1, Totrov M, Abagyan R, Brooks CL 3rd. Comparative study of several algorithms for flexible ligand docking. J Comput Aided Mol Des. 2003; 17: 755-763.
7. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, et al. A critical assessment of docking programs and scoring functions. J Med Chem. 2006; 49: 5912-5931.
8. Zhou Z, Felts AK, Friesner RA, Levy RM. Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets. J Chem Inf Model. 2007; 47: 1599-608.
9. Cheng T, Li X, Li Y, Liu Z, Wang R. Comparative assessment of scoring functions on a diverse test set. J Chem Inf Model. 2009; 49: 1079-1093.
10. Plewczynski D, ÅaÅºniewski M, Augustyniak R, Ginalski K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. J Comput Chem. 2011; 32: 742-755.
11. Xu W1, Lucke AJ, Fairlie DP. Comparing sixteen scoring functions for predicting biological activities of ligands for protein targets. J Mol Graph Model. 2015; 57: 76-88.
12. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009; 30: 2785-2791.
13. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol. 1997; 267: 727-748.
14. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, et al. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem. 2004; 47: 1739-1749.
15. Brooijmans N, Kuntz ID. Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct. 2003; 32: 335-373.
16. Ewing TJ, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des. 2001; 15: 411-428.
17. Lang PT, Brozell SR, Mukherjee S, Pettersen ET, Meng EC, et al. DOCK 6: Combining Techniques to Model RNA-Small Molecule Complexes. RNA. 2009; 15: 1219-1230.
18. Wu G, Robertson DH, Brooks CL 3rd, Vieth M. Detailed analysis of grid-based molecular docking: A case study of CDOCKER-A CHARMm-based MD docking algorithm. J Comput Chem. 2003; 24: 1549-1562.
19. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins. 2003; 52: 609-623.
20. Venkatachalam CM, Jiang X, Oldfield T, Waldman M. Ligand Fit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model. 2003; 21: 289-307.
21. Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. Proteins. 1999; 37: 228-241.
22. Krammer A, Kirchhoff PD, Jiang X, Venkatachalam CM, Waldman M. LigScore: a novel scoring function for predicting binding affinities. J Mol Graph Model. 2005; 23: 395-407.
23. Böhm HJ. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. J Comput-Aided Mol Des. 1994; 8: 243-256.

24. Böhm HJ. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. J Comput Aided Mol Des. 1998; 12: 309-323.

25. Eldridge MDI, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des. 1997; 11: 425-445.

26. Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. J Mol Biol. 1996; 261: 470-489.

27. Wang R, Lai L, Wang S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J Comput Aided Mol Des. 2002; 16: 11-26.

28. Muegge I, Martin YC. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. J Med Chem. 1999; 42: 791-804.

29. Muegge I. PMF scoring revisited. J Med Chem. 2006; 49: 5895-5902.

30. Mooij WT, Verdonk ML. General and targeted statistical potentials for protein-ligand interactions. Proteins. 2005; 61: 272-287.

31. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol. 2000; 295: 337-356.

32. DeWitte RS, Shakhnovich EI. SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. J Am Chem Soc. 1996; 118: 11733-11744.

33. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28: 235-242.

34. Allen FH1. The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Crystallogr B. 2002; 58: 380-388.

35. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res. 2000; 33: 889-897.

36. van Lipzig MM, ter Laak AM, Jongejan A, Vermeulen NP, Wamelink M, et al. Prediction of ligand binding affinity and orientation of xenoestrogens to the estrogen receptor by molecular dynamics simulations and the linear interaction energy method. J Med Chem. 2004; 47: 1018-1030.

37. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. J Mol Biol. 1982; 161: 269-288.

38. Macindoe G, Mavridis L, Venkatraman V, Devignes MD, Ritchie DW. HexServer: an FFT-based protein docking server powered by graphics processors. Nucleic Acids Res. 2010; 38: W445-449.

39. Das AA, Sharma OP, Kumar MS, Krishna R, Mathur PP. PepBind: a comprehensive database and computational tool for analysis of protein-peptide interactions. Genomics Proteomics Bioinformatics. 2013; 11: 241-246.

Austin Publishing Group

Protein Interaction Study of Novel Mutants of Human Hsp70 and Ad5 Motif (PNLVP)

Elengoe A[1] and Hamdan S[1]*

[1]Department of Biosciences and Health Sciences, Faculty of Bioscience and Medical Engineering, Universiti Teknologi Malaysia, Malaysia

***Corresponding author:** Hamdan S, Department of Biosciences and Health Sciences, Faculty of Bioscience and Medical Engineering, Universiti Teknologi Malaysia, Malaysia, Tel: +6075558444; Fax: +6075558515; Email: saleh65@utm.my

Published Date: December 01, 2016

ABSTRACT

In this study, we will explore the protein interaction between Nucleotide Binding Domain **(NBD)** of human heat shock 70 kDa protein (Hsp70) and E1A 32 kDa motif (PNLVP) of human adenovirus serotype 5 (Ad5) in the induction of viral replication. This protein interaction may enhance tumor cell death rate in cancer treatment. Unfortunately, the specific protein interaction between NBD and PNLVP motif is still unknown. To investigate this protein interaction, you will need to construct three dimensional structures of NBD mutants (K71L and T204V) and study its physiochemical characterization using ESBRI, Cys_Recand SOPMA (Self-Optimized Prediction Method from Alignment) servers. After that, you will determine its stabilities by potential energy analysis after run the 50 ns Molecular Dynamics **(MD)** simulation. Then, the stable structure of NBD will be docked with the PNLVP motifusing Autodock version 4.2 and performed for 50 ns MD simulation. Finally, hydrogen bonds, Secondary Structures and Surface Accessible Solvent Area **(SASA)** analyses will be carried out to determine the most stable and best binding affinity with PNLVP motif among all the three protein-ligand complexes. Thus, the Hsp70 structure-based drug discovery may be potential as a cancer treatment.

Keywords: NBD; PNLVP motif; Molecular dynamics simulation; Docking.

INTRODUCTION

Worldwide, cancer is the leading cause of death [1]. However, the success rate of conventional methods such as surgery, chemotherapy and radiotherapy to treat breast cancer has not been very high. Furthermore, these treatments could cause damage to normal cells, DNA which leads to mutation, cardiomyopathy, liver failure and developing other types of cancer [2]. Adenoviral gene therapy is a new therapeutic approach [3] for cancer but recombinant adenovirus therapy alone failed to kill tumor cells completely. This is due to lack of expression of Coxsackie Adenovirus Receptor **(CAR)** and co-receptors (integrin $\alpha_v\beta_3$ and $\alpha_v\beta_5$ classes) in tumor cells which leads to poor infection of adenovirus. Thus, tumor cells hinder the replication of adenovirus. Several researchers indicated that hyperthermia might induce the viral replication in tumor cells [4]. For hyperthermia hypothesis, Heat Shock Protein **(Hsp)** is the main player. Based on Glotzer *et al.* (2000) study, it has been demonstrated that Hsp especially Hsp70 induces the replication of avian adenovirus CELO [5]. Moreover, Hsp70 induction plays a vital role in DNA replication of bacteriophage [6]. In this study, how the NBD and its mutants (K71L and T204V) interact with PNLVP motif and its stabilities were first time determined via Molecular Dynamics **(MD**) simulation and docking approaches. Thus, the model may be used for designing a more potent structure based drug to increase efficacy of adenovirus replication in tumor cells.

In Silico Investigations of Protein Interaction between NBD, K71L, T204V and PNLVP Motif

Firstly, you have to obtain the three dimensional structure of NBD from RCSB Protein Databank (PDB: 1HJO). Protein Databank is an information portal to 111749 biological macromolecular structures. Then, you will mutate the functional amino acid residues (K71L and T204V) using PyMol software [7-8]. PyMol is a molecular visualization system that can generate high-quality three dimensional images of small molecules and biological macromolecules, such as proteins. These residues will be chosen because they play an important role in catalytic activity and stabilization of structure. The conserved Lys71 is a catalytically important residue that affects ATP hydrolysis [9]. The proposed mechanism of ATP hydrolysis suggested that the role of Lys71 in accepting a proton from the hydroxide ion or water molecule involved is in-line with a nucleophilic attack [9-11]. The inorganic phosphate group **(Pi)** is coordinated by a salt bridge with Lys71, hydrogen bonds to Thr13 and Thr204 and interacts directly with a calcium ion. A water molecule mediates additional interactions with the protein's main chain at positions 202, 203 and 204. The Pi-binding site is on the protein face opposite the highly conserved Gly32 loop that has been implicated in the binding of nucleotide release factor **(GrpE)** to the ATPase domain of Hsp40 **(DNAK)** [12]. Therefore, there are potential channels for Pi exit to the protein surface. However, release of the inorganic phosphate group has been implicated in the conformational transition of Hsp70 molecular chaperone [13]. Phospho-threonine was postulated as an intermediate of ATP hydrolysis. In addition, ATPase activity of Hsp70 initiates viral DNA replication. This has

been demonstrated for bacterial DNAJ which stimulates ATPase activity of Hsp70 to start DNA replication of SV40 [14]. Next, you will determine the salt bridges in NBD protein, K71L and T204V mutants using ESBRI program [15]. ESBRI is software available as web tool. It analyses the salt bridges in a whole protein structure or in a single protein chain, among complexed chains and those between user-specified charged residues and the rest of protein, obtained from experimental data or modeling studies or molecular dynamics simulations. You will calculate the disulphide bonds using the Cys_Rec program [16]. The program performs prediction of SS-bonding states of cysteine and locating of disulphide bridges in proteins. Furthermore, you will predict the secondary structure features of NBD, K71L and T204V proteins using Self Optimized Prediction Method from Alignment **(SOPMA)** server [17]. SOPMA is an improvement of SOPM method. This method is based on the homologue method of Levin *et al.* It correctly predicts 69.5% of amino acids for a three-state description of the secondary structure (alpha-helix, beta-sheet and coil) in a whole database containing 126 chains of non-homologous (less than 25%identity) proteins. Validation of generated models was performed by GROMOS [18] and ANOLEA (Atomic Non-Local Environment Assessment) programs [19]. ANOLEA is a web server that performs energy calculations on a protein chain, evaluating the "Non- Local Environment" **(NLE)** of each heavy atom in the molecule. The energy of each pairwise interaction in this non-local environment is taken from a distance-dependent knowledge-based mean force potential that has been derived from a database of 147 non-redundant protein chains with a sequence identity below 25% and solved by X-ray crystallography with a resolution lower than 3 $\AA$.

After that, you will simulate NBD, K71L and T204V using the Gromacs package 4.6.3 [20], adopting the GROMOS 53a6 force field parameter to explore and compare the protein internal dynamics before docked with PNLVP motif [8].In the MD simulations, the proteins will be simulated to examine the structural stability at a temperature of 303 K (27˚C). The protein models will be solvated in a cubic box of explicit simple point charge (SPC) water molecules. One sodium ion will be added to neutralize the total charge of the system for the NBD and T204V proteins, while two sodium ions will be added for D364S mutant. The entire system for the NBD protein, K71L and T204V mutants will be minimized using 749, 1004 and 961 steps of steepest descent, respectively. Linear constraint **(LINCS)** algorithm will be carried out to constrain bond distances, enabling a 2 fs time step and the temperature will be controlled with a Berendsen thermostat with a relaxation time of 0.2 ps. Particle mesh Ewald **(PME)** summation will be used with a direct space cut-off of 1.4 $\AA$ to evaluate the electrostatic and the attractive parts of the Lennard-Jones energies and forces. The system simulated will be first equilibrated at a constant number of particles, volume, temperature and pressure for 50 ps using Periodic Boundary Conditions **(PBC).** All of the resulting trajectories will be analyzed using GROMACS utility. Potential energy analysis will be performed. Next, the binding sites of the protein will be identified using Q-SiteFinder [21-22]. Q-Site Finder is a method for prediction of ligand binding sites. To locate energetically favorable binding sites, it uses the interaction energy between the protein and a simple van der Waals

probe. Energetically favorable probe sites are clustered according to their spatial proximity. After that, the clusters are ranked based on the total of interaction energies for sites within each cluster.

To date, the three-dimensional model of E1A 32 kDa of human adenovirus C serotype 5 **(Ad5)** is not available in the protein database. The complete amino acid sequence of E1A 32 kDa will be retrieved from UniProt Knowledgebase (UniProtKB) (accession number: P03255).The UniProtKB is the central hub for the collection of functional information on proteins such as amino acid sequence, protein name or description, taxonomic data and citation information; with accurate, consistent and rich annotation. Basic Local Alignment Search Tool Protein **(BLASTP)** against the RCSB Protein Databank will be carried out to find a suitable template for homology modeling. Crystal structure of (PDB ID: 2 KJE) will be selected as a template based on maximum identity with high positives and lower gaps percentage. The percentage of query coverage, sequence identity, positive and gap between the template and target protein were 13%, 100%, 100% and 0% respectively. It was built using Easy Modeller 2.1 software [23], the Graphical User Interface **(GUI)** of Modeller 9.10 [24], with the 2KJE protein as a template. The three-dimensional model of the E1A 32 kDa motif (PNLVP) will be created using the built three-dimensional model of E1A 32 kDa as a template. The same homology modeling and 50 ns (50,000 ps) MD simulation approach will be performed before docking with the NBD protein, K71L and T204V mutants.

Then, NBD, K71L and T204V will be docked with PNLVP motif using Autodock Version 4.2 program [25, 8]. Autodock is an automated procedure for predicting the interaction of ligands with biomacromolecular targets. In the protein, non-polar hydrogen atoms will be merged with carbon atoms; and total Kollman and Gasteiger charge will be added to the protein. It will make sure that there are no unbound atoms in the protein. Kollman and Gasteiger partial charges will be also assigned to the ligand and all torsions will be allowed to rotate during docking. The NBD and ligand will be converted from the PDB format to the PDBQT format. A grid box will be used around the active site to cover the entire protein binding site and allow ligands to move freely; and affinity maps NBD, K71L and T204V (74 × 88 × 108, 70 × 60 × 70, 60 × 70 × 95 containing total grid points of 727,575, 307,501, and 411,445, respectively) will be calculated by AutoGrid. One hundred Lamarckian Genetic Algorithm **(LGA)** runs with default parameter settings will be performed. Docking will bereclustered for 0.5, 1.0 and 2.0 tolerances. The largest docked conformations will be clustered at RMS of 1.0 nm and played ranked according to the native Autodock scoring function. The best conformation with the lowest docked energy will be chosen from the docking search. The interactions of complex NBD protein-ligand conformations including hydrogen bonds and bond lengths will be analyzed. The same docking simulation approach will be performed with the single point mutants of NBD (K71L and T204V).

Then, fifty ns MD simulation of NBD, K71L and T204V-PNLVP motif complexes will be carried out to determine their stability throughout the simulation period. The docked complexes of the PNLVP motif with the NBD protein and mutants (K71L and T204V) will be used as a starting point for MD simulations. The GROMACS package 4.6.3 [20]; adopting the GROMOS53a6 force

field parameter, will beused to run MD simulations. The protein topology will be constructed by pdb2gmx with GROMOS53a6 force field. You will use a cubic box setting a minimal distance of 1.0 between the protein and edge of the box, which will then be solvated using periodic boundary conditions and the SPC (simple point charge) water model in this study. The ligand topology file will be generated using the PRODRG server to include the heteroatom due to limitations of GROMACS to parameterize the heteroatom group in the PDB file [26]. To make the system neutral, you will add one sodium ion around the molecule for the NBD and T204V proteins, whereas two sodium ions will be added for K71L protein. The entire system for the NBD protein, K71L and T204V mutants will be minimized using 993, 945 and 1023 steps of steepest descent, respectively. After energy minimization with particle-mesh Ewald algorithm at every step, the system will then be equilibrated at a constant temperature (303 K), volume, number of particles in system and pressure (1 bar) for 50 ps. Under constant volume equilibration, the temperature will be maintained by Berendsen weak coupling method. Moreover, under constant pressure equilibration, the temperature will be controlled by Berendsen weak coupling method and the pressure will be maintained by Parrinello-Rahman baro-stat method. After completion of the two equilibration phases, production of MD simulations will be conducted for 1 ns after taking away the position restraints. Finally, the equilibrated structures will be subjected to MD simulations for 50 ns (50,000 ps) with a LINCS algorithm 2 fs time step to constrain all the bonds. The non-bonded list will be generated using an atom-based cut-off of 10 $\AA$. The trajectory snapshots will be taken for structural analysis at every pico-second. The H-bonds, secondary structures and solvent accessible surface area between the protein and ligand in the docked complex during the MD simulation will be analyzed using Gromacs analysis tools.

In this case study, ESBRI results showed that the NBD, K71L and T204V consists of 14, 19 and 20 salt bridges respectively which were formed by arginine residues. An increase in the number of salt bridges contributes the stability of protein to be improved. Salt bridge plays vital roles in structure and function of protein. The disruption of a salt bridge reduces the protein stability [27]. It is also involved in allosteric regulation, recognition of molecular, oligomerization, flexibility, domain motions and thermo stability [28-29]. The presence of arginine in the protein model increases the thermo stability of a molecule by providing more electrostatic interactions through their guanidine group. The Cys_Rec analysis indicates the number of disulphide bonds which provide stability to the protein structures. Based on the Cys_Rec results, it exhibited that NBD, K71L and T204V contain three disulphide bonds. Both of the analyses revealed that T204V had the most stable structure among all the protein models.

The secondary structure indicates whether a given amino acid lies in a helix, strand or coil. Alpha-helices in proteins are generated by local hydrogen bonding between C=O and N-H groups that are close together in the polypeptide chain. The significance of the helix is to generate the dipole moments, which contribute to the binding of small charged molecules to proteins. The individual peptide dipoles in helices contribute to making a macrodipole, with the amino-

terminal end of the helix polarized positively and carboxy-terminal end polarized negatively. In helices, favorable electrostatic interactions are established between positively charged species and the end of the helix dipole, whereas negatively charged side chains and cations interact with the carboxy-terminal ends [30]. SOPMA view showed that presence of alpha-helix dominated among secondary structure elements followed by random coils, extended strand and beta turns at various positions in all the mutants of NBD. K71L consists of 44.21%, 18.68%, 8.16% and 28.95% for alpha-helix, extended strand, beta turn and random coil respectively. While, T204V consists of 44.47%, 18.95%, 6.84% and 29.74% for alpha-helix, extended strand, beta turns and random coil respectively. Furthermore, SOPMA analysis indicated that NBD, K71L and T204V consist of 13 α-helixes. NBD had the tenth helix as the longest α- helix whereas ninth and eleventh helices were the shortest α-helix. In addition, K71L and T204V had the eleventh helix as the longest α- helix whereas fifth, tenth helices and twelve and thirteen helices were the shortest α-helix for K71L and T204V respectively.

For ANOLEA [18] and GROMOS [19] analysis, the y-axis of the plot represents the energy for each amino acid of the protein chain. Negative energy values (in green) represent favorable energy environment whereas positive values (in red) unfavorable energy environment for a given amino acid. ANOLEA and GROMOS results demonstrated that most of the amino acids in favorable energy environment (green bars) (Figure 1). Therefore, all mutant models (K71L and T204V) were good and reasonable. T204V had the most favorable energy compared with the NBD and K71L.

In the current study, three 50 ns MD simulations were performed with NBD, K71L and T204V before docked with PNLVP motif. The potential energy analysis revealed that NBD, K71L and T204V had -1689267.125, -1687322.875 and -1689947.000 kJ/mol respectively. The potential energy of all protein models was low. This implied that the folding of all the protein models refined was stable. However, the degree of stability varies depending on the energy. The potential energy of T204V mutant was found to be the lowest (-1689947.000 kJ/mol) among all the protein models. This revealed that T204V is the most stable protein model.

The area (cubic *Å*) and volume (cubic *Å*) of predicted active sites for K71L and T204V were determined using Q-SiteFinder. The area for K71L and T204V were 291 and 445 cubic *Å* respectively. While the volume for K71L and T204V were 34383 and 34423 cubic *Å*. In addition, the NBD, K71L and T204V were successfully docked with PNLVP motif (Figure 2). The negative and lowest value of binding energy, ΔG_{bind} (-7.73 Kcal/mol) indicated strong bonds between T204V and the PNLVP motif, and demonstrated that the protein was in a most favorable conformation when compared with the NBD (-7.31 Kcal/mol) and K71L (-7.40 Kcal/mol).

In this study, three 50 ns MD simulations were performed with NBD, K71L and T204V-PNLVP motif complexes. The hydrogen bond analysis indicated that the NBD, K71L and T204V-PNLVP motif complexes shows five, six and six intermolecular hydrogen bonds respectively, which were

determined along the simulation period (Figure 3). At least 1-2 intermolecular interactions were kept for the entire 50 ns trajectories; which inferred the stability of the NBD, K71L and T204V-PNLVP motif complexes.

In addition, secondary structure also one of factor that influence the stability of the protein. All protein-ligand complexes consist of coil, turn, β-sheet, β-bridge, bend, A-helix and 3-helix. All the protein-ligand complexes remained relatively stable during the 50 ns MD simulations (Figure 4).

Solvent-accessible surface area (SASA) defined as the surface area of a biomolecule that is accessible to a solvent [31]. In 1973, Shrake and Rupley developed 'rolling ball' algorithm to calculate SASA [32]. Solvent accessibility was divided predominantly into buried and exposed region which describes the least accessibility and high accessibility of the amino acid residues to the solvent [33]. (Figure 5) shows the SASA against the simulation period for NBD, K71L and T204V-PNLVP motif complexes. The SASA maintained constant along the 50 ns MD simulation for all protein-ligand complexes. Increase or decrease in the SASA infers the changes in amino acid residues. Modification of SASA could affect the tertiary structure of protein.

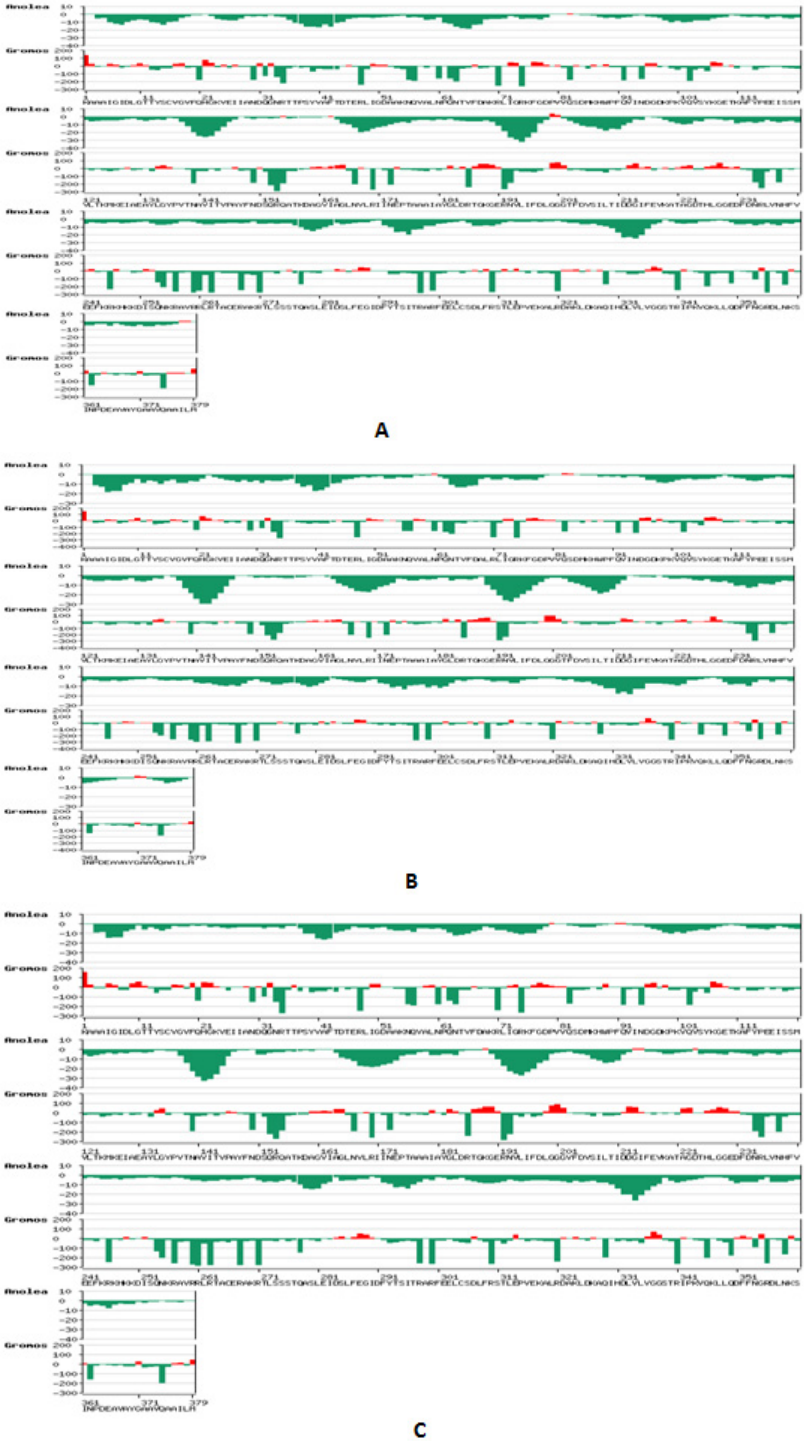


Figure 1: Evaluation of (**A**) NBD Protein; (**B**) K71L; and (**C**) T204V Protein Models Using ANOLEA and GROMOS Analysis.

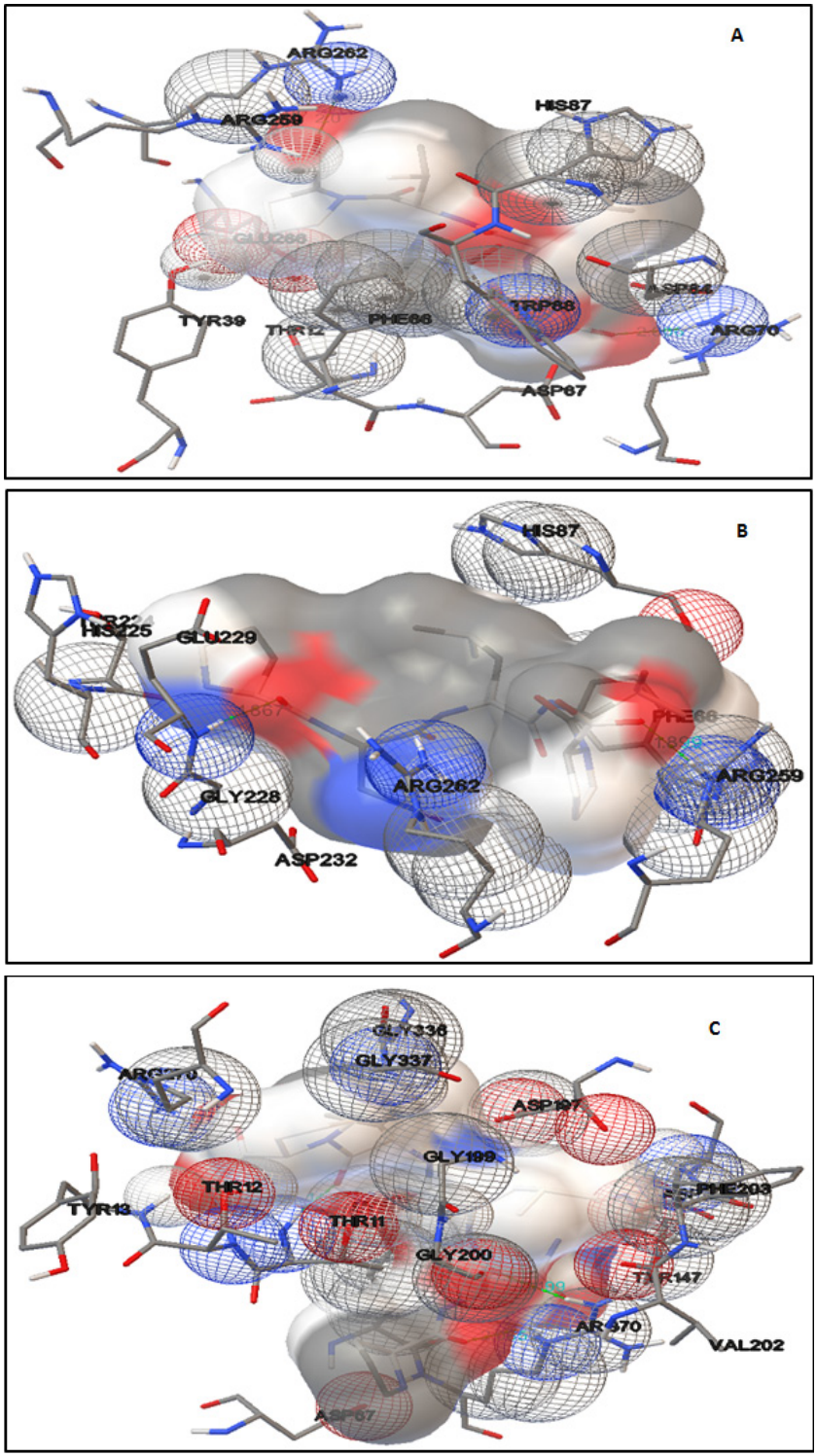


Figure 2: Docking of the (**A**) NBD Protein; (**B**) K71L; and (**C**) T204V with the PNLVP Motif.

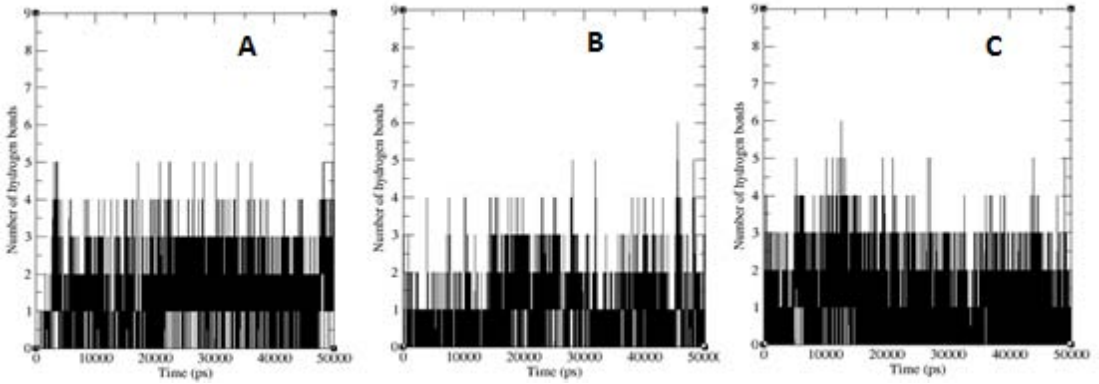


Figure 3: Number of Hydrogen Bonds for the (**A**) NBD; (**B**) K71L; and (**C**) T204V-PNLVP Motif Complex Structures.

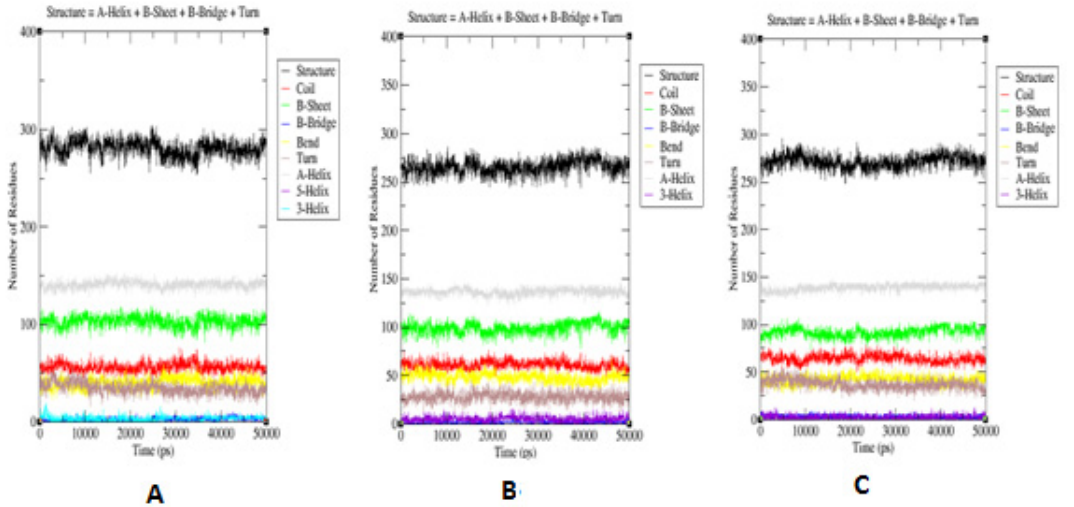


Figure 4: Secondary Structure Analysis for the (**A**) NBD; (**B**) K71L; and (**C**) T204V-PNLVP Motif Complex Models. The Structure Was Composed Of A-Helix, B-Sheet, B-Bridge And Turn.

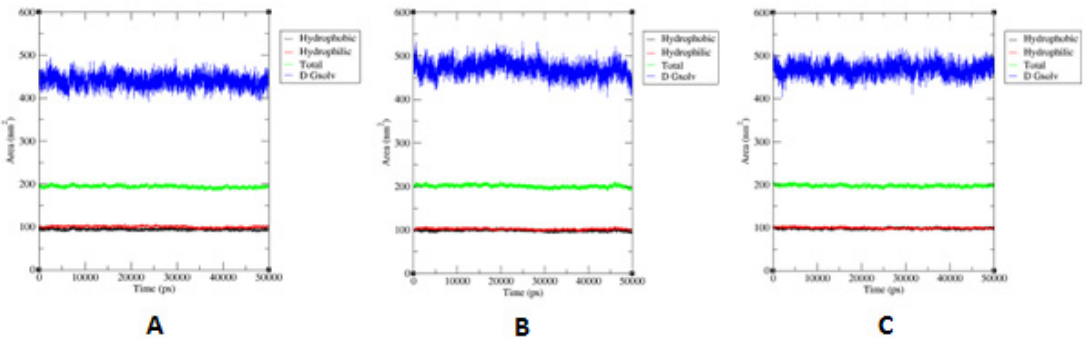


Figure 5: Solvent Accessible Surface Area (SASA) Analysis for the (**A**) NBD; (**B**) K71L; and (**C**) T204V-PNLVP Motif Complex Structures

CONCLUSION

The novel mutant (T204V) had a better interaction with PNLVP motif than NBD protein and K71L mutant. In addition, T204V-PNLVP motif complex had the most stable structure amongst all the protein-ligand models. Thus, further biochemical and *in vivo* investigation of *in silico* interpretations of this protein-ligand complex will be a new approach for designing Hsp70 structure based drug in cancer treatment.

ACKNOWLEDGMENT

This work was supported by a grant No.04H06 from Universiti Teknologi Malaysia.

References

1. Bresalier RS, Kopetz S, Brenner DE. Blood-based tests for colorectal cancer screening: do they threaten the survival of the FIT test? Dig Dis Sci. 2015; 60: 664-671.
2. Hawkins LK, Hermiston T. Gene Delivery from the E3 Region of Replicating Human Adenovirus: Evaluation of the E3B Region. Gene Ther. 2001; 8: 1142-1148.
3. Abaan D, Criss WE. Gene Therapy in Human Breast Cancer. Turk J Med Sci. 2002; 32: 283-291.
4. Thorne SH, Brooks G, Lee YL, Au T, Eng LF, et al. Effects of Febrile Temperature on Adenoviral Infection and Replication: Implications for Viral Therapy of Cancer. J Virol. 2005; 79: 581-591.
5. Glotzer JB, Saltik M, Chiocca S, Michou AI, Moseley P. Activation of heat-shock response by an adenovirus is essential for virus replication. Nature. 2000; 407: 207-211.
6. Wickner S, Skowyra D, Hoskins J, McKenney K. DnaJ, DnaK, and GrpE heat shock proteins are required in oriP1 DNA replication solely at the RepA monomerization step. Proc Natl Acad Sci U S A. 1992; 89: 10345-10349.
7. Delano WL. The PyMOL Molecular Graphics System.
8. Elengoe A, Naser MA, Hamdan S. Modeling and docking studies on novel mutants (K71L and T204V) of the ATPase domain of human heat shock 70 kDa proteins 1. Int J Mol Sci. 2014; 15: 6797-6814.
9. O'Brien MC, Flaherty KM, McKay DB. Lysine 71 of the chaperone protein Hsc70 is essential for ATP hydrolysis. J Biol Chem. 1996; 271: 15874-15878.
10. Flaherty KM, Wilbanks SM, DeLuca-Flaherty C, McKay DB. Structural basis of the 70 kilodalton heat shock cognate protein ATP hydrolytic activity. J. Biol. Chem. 1994; 269: 12899-12907.
11. McCarty JS, Walker GC. DnaK as a thermometer: threonine-199 is site of autophosphorylation and is critical for ATPase activity. Proc Natl Acad Sci USA. 1991; 88: 9513-9517.
12. Buchberger A, Schröder H, Büttner M, Valencia A, Bukau B, et al. A conserved loop in the ATPase domain of the DnaK chaperone is essential for stable binding of GrpE. Nat Struct Biol. 1994; 1: 95-101.
13. Wawrzynow A, Banecki B, Wall D, Liberek K, Georgopoulos, C, et al. ATP hydrolysis is required for the DnaJ-dependent activation of DnaK chaperone for binding to both native and denatured protein substrates. J. Biol. Chem. 1995; 270: 19307-19311.
14. Liu JS, Kuo SR, Makhov AM, Cyr DM, Griffith JD, et al. Human Hsp70 and Hsp40 chaperone proteins facilitate human papillomavirus-11 E1 protein binding to the origin and stimulate cell-free DNA replication. J Biol Chem. 1998; 273: 30704-30712.
15. Costantini S, Colonna G, Facchiano AM. ESBRI: a web server for evaluating salt bridges in proteins. Bioinformation. 2008; 3: 137-138.
16. Roy S, Maheshwari N, Chauhan R, Sen NK, Sharma A. Structure prediction and functional characterization of secondary metabolite proteins of Ocimum. Bioinformation. 2011; 6: 315-319.
17. Geourjon C, Deléage G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput Appl Biosci. 1995; 11: 681-684.
18. Van Gunsteren W. Biomolecular simulations: The GROMOS96 manual and user guide. 1996. VdFHochschulverlag ETHZ.
19. Melo F, Feytmans E. Assessing protein structures with non-local atomic interaction energy. J Mol Biol. 1998; 277: 1141-1152.

20. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE. GROMACS: fast, flexible, and free. J Comput Chem. 2005; 26: 1701-1718.

21. Laurie AT, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 2005; 21: 1908-1916.

22. Burgoyne NJ, Jackson RM. Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. Bioinformatics. 2006; 22: 1335-1342.

23. Kuntal BK, Aparoy P, Reddanna P. Easy Modeller: A graphical interface to MODELLER. BMC Res Notes. 2010; 3: 226.

24. Fiser A, Sali A. Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol. 2003; 374: 461-491.

25. Sanner MF. Python: a programming language for software integration and development. J Mol Graph Model. 1999; 17: 57-61.

26. Schüttelkopf AW, van Aalten DM. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. Acta Crystallogr D Biol Crystallogr. 2004; 60: 1355-1363.

27. Kumar S, Tsai CJ, Ma B, Nussinov R. Contribution of salt bridges toward protein thermostability. J Biomol Struct Dyn. 2000; 17: 79-85.

28. Kumar S, Nussinov R. Relationship between ion pair geometries and electrostatic strengths in proteins. Biophys J. 2002; 83: 1595-1612.

29. Kumar S, Nussinov R. Salt bridge stability in monomeric proteins. J Mol Biol. 1999; 293: 1241-1255.

30. Petkso GA, Ringe D. Protein Struct. & Funct. 2004. New Science Press Ltd.

31. Doss CG, Nagasundaram N. Investigating the structural impacts of I64T and P311S mutations in APE1-DNA complex: a molecular dynamics approach. PLoS One. 2012; 7: e31677.

32. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol. 1973; 79: 351-371.

33. Gilis D, Rooman M. Predicting Protein Stability Changes upon Mutation using Database-Derived Potentials: Solvent Accessibility Determines the Importance of Local versus Non-local Interactions along the Sequence. J Mol Biol.1997; 272: 276-290.

Austin Publishing Group

Structure Based Drug Design in Identification of Novel Androgen Receptor Antagonist

Divakar S[1], Hariharan S[2] and Ramanathan M[1]*

[1]Department of Pharmacology, PSG College of Pharmacy, India

[2]Department of Medicinal Chemistry, PSG College of Pharmacy, India

***Corresponding author:** Ramanathan M, Department of Pharmacology, PSG College of Pharmacy, Coimbatore, India, Tel: +918870009199; Email: muthiah.in@gmail.com

Published Date: December 01, 2016

ABSTRACT

Prostate cancer is the second most frequent and sixth leading cause of mortality among male population. Androgen receptor antagonist is one of the regimens for prostate cancer. The relapsed prostate cancer patients were found to express point mutations in the ligand binding domain of the androgen receptor. The mutations at amino acids, threonine 877 and tryptophan 741 were frequently expressed among the androgen independent prostate cancer patients. These point mutations were responsible for the development of resistance against androgen receptor antagonist. Identification of novel androgen receptor antagonist was hampered because of the expression of different types of point mutation and the lack of insight into the binding mode of the androgen receptor antagonist with the mutated receptor. In this study, we have discussed the method adopted by us in designing a novel androgen receptor antagonist that could tolerate the point mutations. We also discussed the role of the point mutations in the development of resistance against androgen receptor antagonist.

Keywords: Prostate cancer; Androgen receptor; Resistance; T877A; W741L

Abbreviations: Prostate Cancer **(PCa)**; Androgen Receptor **(AR)**; Castration Resistant Prostate Cancer **(CRPC)**; Androgen Independent Prostate Cancer **(AIPC)**; Heat Shock Protein **(HSP)**; Androgen Receptor Response Elements **(AREs)**; Steroid Receptor Co-activator **(SRC)**; Transcriptional Intermediary Factor **(TIF)**; cAMP Response Element Binding **(CREB)**; CREB Binding Protein **(CBP)**; Dihydrotestosterone **(DHT)**; Androgen Deprivation Therapy **(ADT)**; Combined Androgen Blockade **(CAB)**; Gonadotropin Releasing Hormone **(GnRH)**; Phosphatase and Tensin Homolog **(PTEN)**; E26 Transformation Specific **(ETS)**; Transmembrane Protease Serine 2 **(TMPRSS2)**; Mirror-Image Polydactyly Gene 1 Protein **(MIPOL)**; Threonine to Alanine **(T877A)**; Threonine to Serine **(T877S)**; Valine to Methionine **(V715M)**; Alanine to Threonine **(A721T)**; Asparagine to Aspartic Acid **(N756D)**; Histidine to Tyrosine **(H874Y)**; Tryptophan to Cysteine **(W741C)**; Tryptophan to Leucine **(W741L)**; Aspartic Acid to Glycine **(D879G)**; Protein Data Bank **(PDB)**; Hydrogen Bond **(H-Bond)**.

INTRODUCTION

Prostate Cancer

The growth, development, and homeostasis of the prostate gland are under the control of androgens. The growth of prostate gland occurs significantly during puberty and after that, the androgens continue to play an important role in its function. In some men, with increasing age, androgen dependant proliferation of the prostate gland resumes, resulting in benign prostatic hyperplasia or malignant prostate cancer [1]. Most of the Prostate Cancer **(PCa)** relapses within 18 to 24 months of hormonal therapy. The relapsed PCa would be either Castration Resistant Prostate Cancer **(CRPC)** or Androgen Independent Prostate Cancer **(AIPC)** [2].

The enzymes involved in the synthesis of Dihydrotestosterone **(DHT)** are over expressed in CRPC resulting in the relapse. The CRPC could be controlled by including the drugs which inhibit the enzymes involved in the synthesis of DHT [3]. Unlike CRPC, the AIPC doesn't depend on androgens for proliferation; instead it adapts various mechanisms to neutralize the hormonal therapy. Clinically, AIPC is defined as the ability of the PCa cells to grow in the castrated plasma levels of androgen. Point mutation in the ligand binding domain of the Androgen Receptor **(AR)** is one of the major reasons behind the androgen independent proliferation of the PCa cells [4,5].

Androgen Receptor

Human AR is a nuclear receptor encoded by a single gene located on human X-chromosome at Xq11-12 region that spans more than 90kb and has 8 exons [6]. AR is a 900-920 amino acid protein and its variations are due to the polymorphism in the length of polyglutamine (CAG repeats) and polyglycine (GGN repeats) tracts in the first exon [7]. Like other nuclear receptors, AR is also divided into 4 regions: variable N-terminal domain or Activation Function 1 **(AF-1)**, DNA binding domain, hinge region, and ligand binding domain or Activation Function 2 **(AF-2)**.

The ligand binding domain of AR is coded by exons 6, 7, 8, and c-terminal part of exon 5. The ligand binding domain of AR has 12α helixes and a β sheet that fold to form a hydrophobic ligand binding pocket to which the androgens bind [8]. Helix 4, 5, and 10 are the primary contact sites for androgens. The binding of androgens with AR majorly involves hydrophobic interactions with amino acids Val746, Met742, Gln711, Met745, Leu707, Leu704, and Trp741. The hydrogen bond interactions also play a critical role in the specific binding of androgens to AR [9].

Androgen Dependant Androgen Receptor Activation

Androgen is required for the maximum activation of the AR. The AR predominantly resides in cytoplasm, associated with Heat Shock Protein **(HSP)** and chaperons. The HSP also has an active role, HSP90 bind to AR and keeps it in an active conformation that is necessary for androgens to bind [10]. The nuclear localization signal of the AR resides within amino acids 742 to 817, which keeps the AR within the cytoplasm in the resting state. The nuclear localization signal is dominant over the nuclear export signal in the resting state [11].

The ligands bind to the ligand binding domain of the AR. The helix-12 of the ligand binding domain attains different conformation with agonist and antagonist. Androgen binding causes the helix-12 to lie over the ligand binding pocket, which will reveal the domain for intramolecular AF-1 and AF-2 interaction (N/C terminal interaction). The N/C terminal interaction is specific to AR. The AF-2 region of the other nuclear receptors will preferentially interact with leucine rich LxxLF motifs in the co-activators. The AF-2 region of the AR interacts with phenylalanine rich FxxLF motifs in the AF-1 region. The N/C terminal interaction will lead to cascade of events like phosphorylation, dimerization, nuclear translocation, binding to specific androgen receptor response elements, co-activator recruitment, and initiation of transcription [12,13]. Class 1 family of co-activators, Steroid Receptor Co-activator **(SRC)**-1, Transcriptional Intermediary Factor-2 **(TIF2)**, SRC-3, and class 2 family of co-activators, which are also known as transcriptional integrators, p300/CREB (cAMP Response Element Binding) binding protein, regulate the transcriptional activity of AR [14]. The co-activators decondense the chromatin and facilitate the binding of RNA polymerase for the initiation of transcriptional activity. The AR co-activators preferentially interact with the glutamine rich region (1053 - 1123) in the AF-1. Mutant AR that doesn't have this glutamine region is inactive [13]. Antagonist binding displaces the helix-12 away from the ligand binding pocket and unveils the binding surface for co-repressor NcoR/SMRT interaction, which leads to the inhibition of the AR mediated transcriptional activity [15].

Androgen Independent Androgen Receptor Activation

The aim of the treatment in metastatic PCa is to reduce the plasma levels of prostate specific antigen and androgen. This can be established by either Androgen Deprivation Therapy **(ADT)** or Combined Androgen Blockade Therapy **(CAB)**. ADT includes surgical or pharmacological castration. Surgical castration involves bilateral orchiectomy (removal of testes) and the standard castrate level of testosterone could be achieved within 12hr. Pharmacological castration can be

achieved by either Gonadotropin Releasing Hormone **(GnRH)** receptor agonist or GnRH receptor antagonist. There will be an initial rise in the testosterone concentration by using GnRH receptor agonist but after 2 to 4 weeks, castration levels of testosterone will be achieved. In CAB therapy, in addition to castration (surgical or pharmacological), AR antagonist or 5α reductase inhibitors are used to prevent AR activation by adrenal androgens [16,17].

The hormonal therapy has beneficial outcome for 18-24 months but the PCa relapses because of many other pathological conditions. Some of the major reasons are, over expression of enzymes involved in DHT synthesis and subsequent rise in intracraine androgen level [18], down regulation of enzymes involved in DHT catabolism [19], somatic point mutations in the ligand binding domain of AR [4,5], AR over expression [20], truncation of AR ligand binding domain, and constitutive activation of AF-1 region [21], Phosphatase and Tensin homolog **(PTEN)** loss [22], increased telomerase activity [23], mutations in p53 [24], down regulation of E-Cadherin and over expression of N-Cadherin which increases the heterotypic cell adhesion and metastasis [25], translocation and fusion of E26 Transformation Specific **(ETS)** genes with androgen responsive genes like Transmembrane Protease Serine 2 **(TMPRSS2)** or prostate specific Mirror-Image Polydactyly gene 1 protein **(MIPOL1)** gene [26,27].

The fusion between androgen responsive and ETS genes were identified in 40 to 70% of aggressive PCa cases [27]. The ETS family of transcriptional factors is oncogenic and controls the cell differentiation, cell division, metastasis, angiogenesis etc. The fusion of ETS genes with androgen responsive genes results in the over expression of ETS genes whenever AR is activated. Somatic point mutation in the ligand binding domain of AR was also frequently identified in drug resistance PCa patients [4,5]. The mechanism by which the AR mutates was unknown and many point mutations were identified among AIPC patients. The mutated AR does not depend on androgen for activation; instead it can be activated by other steroidal hormones like estradiol, progesterone, and AR antagonists. For the ETS to over-express, it requires AR activation, since it was fused with an AR responsive genes. The hormonal therapy decreases the plasma testosterone level but the mutated AR does not require androgen for activation. Consequently, the mutated AR could activate the expression of ETS genes without androgen. The AR mutation and ETS chromosomal rearrangement in conjunction could co-ordinate the AIPC progression and drug resistance for AR antagonist and cytotoxic drugs. A novel AR antagonist that resists the mutations and decreases the expression of ETS genes could well be the future for PCa treatment.

Mutations in the Ligand Binding Domain of Androgen Receptor

The somatic point mutations can be broadly classified into two types

Mutations expressed in flutamide treated patients

The mutations, T877A (Threonine to Alanine), T877S (Threonine to Serine), V715M (Valine to Methionine), A721T (Alanine to Threonine), N756D (Asparagine to Aspartic acid), and H874Y

(Histidine to Tyrosine) were identified in AIPC patients who underwent flutamide containing therapy [5,28,29]. These mutations cause resistance to flutamide treatment and these mutant ARs are activated by flutamide.

Mutations expressed in bicalutamide treated patients

The mutations, W741C (Tryptophan to Cysteine), W741L (Tryptophan to Leucine), and D879G (Aspartic acid to Glycine) were identified in AIPC patients who underwent bicalutamide containing therapy [28,30]. These mutations were responsible for bicalutamide resistance and these mutant ARs are activated by bicalutamide.

In general, these mutations broaden the ligand specificity of the receptor. For example, the T877 amino acid mutated AR was activated by non androgens like cyproterone acetate, flutamide, estrogens, glucocorticoids, progesterone etc., whereas, bicalutamide antagonizes the T877 mutant AR. Similarly the W741 amino acid mutated AR was activated by non androgens like bicalutamide, estrogens, glucocorticoids, progesterone etc., whereas, flutamide antagonizes the W741 mutant AR [30,31,32,33,34].

COMPUTATIONAL METHODS

Docking

The ligands for the docking studies were prepared using LigPrep, Schrödinger. The ligands were geometrically refined and assigned appropriate protonation state at pH 7.0 ± 2.0. The energy minimization was carried out by OPLS 2005 force field [35]. The proteins (PDB: 2AMA, 2AX6, 1Z95 & helix-12 truncated ARs) were prepared using protein preparation wizard in Maestro, Schrödinger [36]. The preprocessed protein was then used to generate the grid for docking. The grid was assigned by picking the ligand as the center of the grid and the grid box was generated by applying default parameters. The docking was carried out using GLIDE, Schrödinger. GLIDE XP (extra precision) method was followed for docking calculations [37].

Homology Modelling

The helix-12 truncated AR was generated using PRIME, Schrodinger [38]. The Prime suite was used for protein structure prediction, side chain optimization, loop prediction, active site refinement and energy minimization. The amino acid sequence of the ligand binding domain of AR was obtained from NCBI (Protein accession No: P10275.2).

The template protein was identified through blast search. The AR crystal structure, 2AMA was the template for wild type AR and 2OZ7 was the template for T877A mutated AR. The alignment between the target and template sequence was carried out by clustalW method. ClustalW could be used when there is a high sequence identity between the target and template. Prime allows us to build the homology model in two different types, either knowledge based or energy based. We used a knowledge based method were it uses the structure information from the template for gaps

and insertions. As default, the side chains were only optimized for the residues that are not from the template. The ligand and a water molecule (HOH-108) were retained in the final model. The homology model was then energy minimized (OPLS 2005) by VSGB solvation and we also refined the active site (5*Å* from the ligand) of the model.

CASE STUDY

Target Selection

It is impossible to consider all those mutations while designing an AR antagonist. The mutations in T877 amino acid occur frequently among the flutamide treated patients [34]. At present, flutamide was largely replaced with bicalutamide and there was no report of T877A/S mutation among bicalutamide treated patients. Mutations in W741 predominate among the bicalutamide treated patients [28,30]. The T877A mutated AR was expressed in LNCaP cell line. The LNCaP could also express the W741L/C mutation upon incubation with bicalutamide in an androgen depleted medium for 6-13 weeks [30]. So, the mutations are treatment specific and might keep on piling up depending on the antagonist. Targeting the wild type AR for drug designing is essential because the mutations occur during the course of treatment and as discussed above, the mutations are treatment specific. Understanding the mechanism by which these mutations convert an antagonist to agonist is essential because a novel AR antagonist should tolerate these types of mutations that might occur during the course of treatment.

The crystal structure of the wild type and some of the mutated ARs are available in Protein Data Bank **(PDB)**. Unfortunately, none of them are in antagonist bound conformation. Generally, the steroidal receptor antagonist has bulkier groups than the endogenous agonist. Antagonist form similar H-Bond interaction like agonist but due to their bulkier nature, they displace the helix-12 away from the ligand binding pocket [39,40]. The helix-12 of the ligand binding domain adapts closed and open conformation in response to agonist and antagonist respectively [41,42]. Consequently, we modelled an AR that lacks helix-12 for the structure based virtual screening. The truncated receptor has 671-881 amino acids instead of 671-919 amino acids.

Docking Studies

Docking studies were carried out to evaluate the role of the point mutation in converting an AR antagonist to agonist. We chose 4 different types of AR: wild type (PDB: 2AMA), T877A (PDB: 2AX6), W741L (PDB: 1Z95), and helix-12 truncated ARs (homology model). The ligands were DHT, testosterone, non steroidal AR antagonist flutamide, bicalutamide, and enzalutamide.

Androgen

The androgen, DHT, has binding affinity towards all the ARs but did not form H-Bond interaction with T877A mutated AR (Table1). The DHT and testosterone have similar type H-Bond interactions with wild type AR (Figure 1). The keto group of DHT and testosterone formed H-Bond interaction with Arg752, while the 17β hydroxyl group formed H-Bond interaction with Asn705

and Thr877. Testosterone did not bind with the T877A mutated AR and didn't form any H-Bond interaction with W741L mutated AR. The binding energy of the androgens was less in the mutated ARs than the wild type AR. The DHT has a dock score of -10.80 kcal/mol with wild type AR but the dock score had decreased with T877A (-5.84 kcal/mol) and W741L (-6.71 kcal/mol) mutated ARs. Similarly the testosterone has a dock score of -10.68 kcal/mol with wild type AR but the dock score had decreased with T877A (no binding) and W741L (-5.95 kcal/mol) mutated ARs. This indicates that the mutations might affect the binding of the androgens with AR. Ligand binding assay also proved that the mutations could decrease the binding affinity of androgens [43].

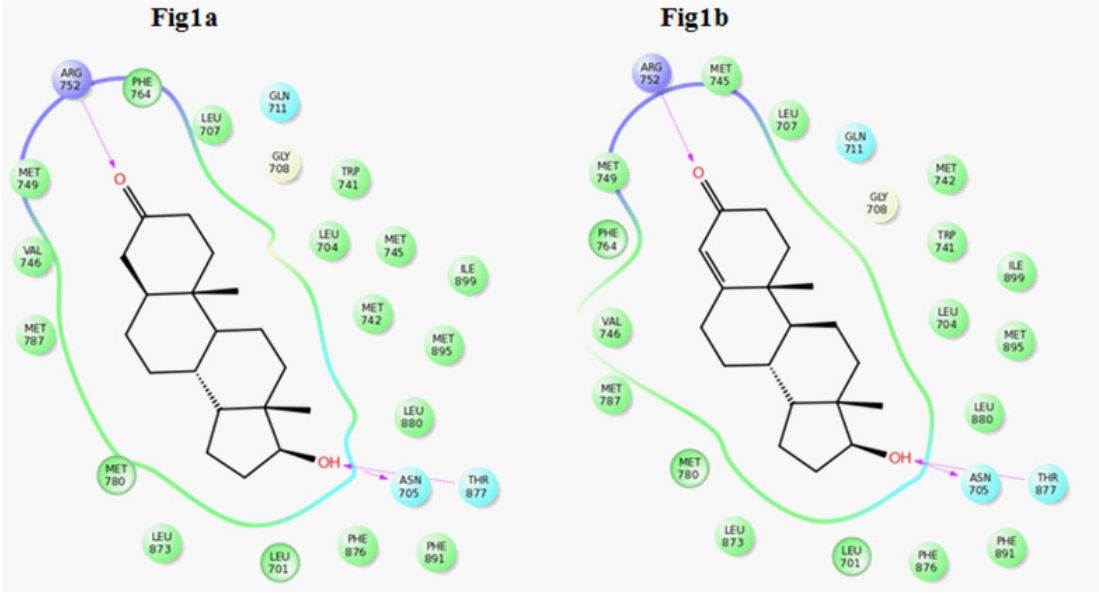


Figure 1: Binding interactions of androgens.

1a: DHT forms H-Bond interaction with Arg752, Asn705 and Thr877 with wild type AR. **1b**: Testosterone forms H-Bond interaction with Arg752, Asn705 and Thr877 with wild type AR.

Table 1: Docking score and H-Bond interaction of androgen receptor ligands.

Ligands	Wild type		T877A		W741L		Truncated	
	Dock Score[1]	H-Bond	Dock Score	H-Bond	Dock Score	H-Bond	Dock Score	H-Bond
DHT	-10.80	Arg752, Asn705, Thr877	-5.84	×	-6.71	Arg752, Gln711, Thr877, HOH108	-7.26	HOH108, Asn705
Testosterone	-10.68	Arg752, Asn705, Thr877	×	×	-5.95	×	-7.17	HOH108, Asn705
Flutamide	-7.76	Asn705, Thr877	-8.25	Arg752, Gln711, Asn705, Leu704, HOH108*, HOH213#	-7.22	Arg752, Gln711, Leu704, Asn705, HOH108	-6.22	Arg752, Gln711, Asn705, HOH108
Bicalutamide	-2.548	×	×	×	-10.27	Arg752, Gln711, Leu704, Asn705, HOH108	-9.22	Arg752, Gln711, Asn705
Enzalutamide	×	×	×	×	×	×	-5.87	Arg752, Gln711

[1]Dock score in kcal/mol.

*HOH108 could form bridged H-Bond interactions with Arg752, Gln711 and Met745.

#HOH213 could form bridged H-Bond interaction with Leu873.

Flutamide

Flutamide has binding affinity for all the AR types used in this study. Flutamide has the highest binding energy with T877A mutated AR (dock score = -8.25 kcal/mol). Flutamide has a dock score of -7.76 kcal/mol and -7.22 kcal/mol with wild and W741L mutated ARs, respectively. As discussed above, T877A mutation causes resistance specifically to flutamide. In wild type AR, the hydroxyl group of flutamide forms H-Bond interactions with Thr877 and Asn705 (Figure 2). In T877A mutated AR, the nitro group of flutamide formed H-Bond interactions with Gln711, Arg752, HOH108, and Met745. The hydroxyl group of flutamide, in the absence of Thr877 amino acid, formed H-Bond interaction only with Asn705. The keto group of the amide formed H-Bond interaction with Leu873 mediated through a water molecule (HOH213) and the amino group of the amide formed H-Bond interaction with Leu704. The exact mechanism by which the T877A mutation converts the flutamide from antagonist to agonist was unknown, but we could observe that the size of the ligand binding pocket in T877A was bigger than the wild type AR (Figure 2). The binding conformation of flutamide with wild and T877A mutated ARs was also different. Flutamide attained a bent conformation when bound to T877A mutated AR. It is possible that the T877A mutation could increase the space within the ligand binding pocket to accommodate for larger ligands without disturbing the closing of helix-12.

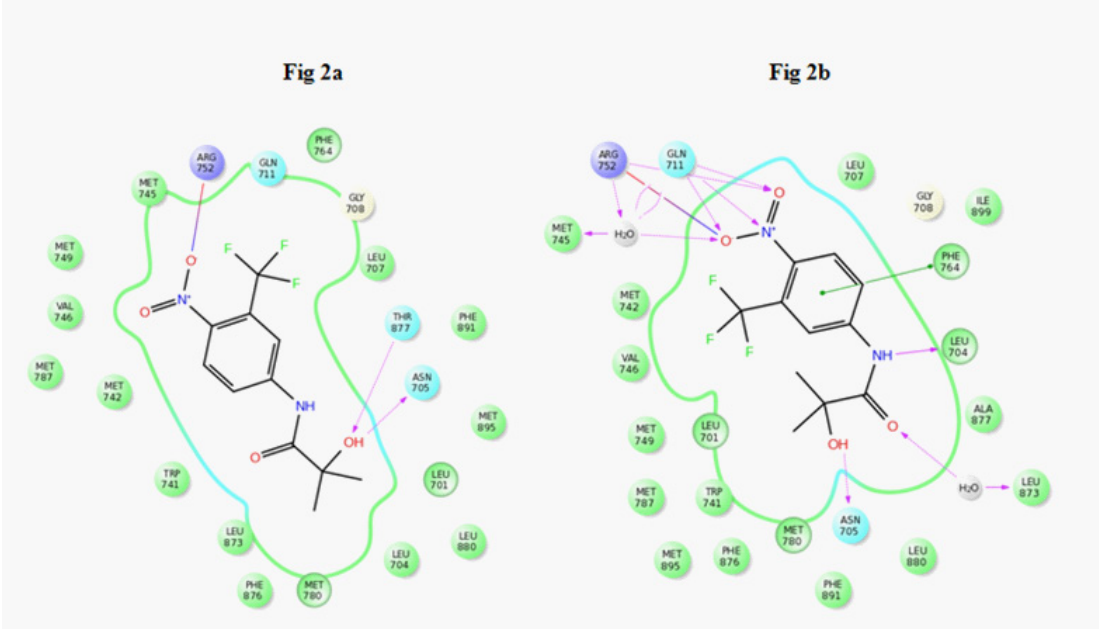


Figure 2: Binding interactions of flutamide.

2a: Flutamide forms H-Bond interaction with Thr877, Asn705 and forms ionic interaction with Arg752 in wild type AR. **2b**: Flutamide forms H-Bond interactions with Arg752 (ionic and H-Bond), Gln711, Leu704, Leu873 (bridged through HOH) and Met745 (bridged through HOH) in T877A mutant AR.

Bicalutamide

The AR antagonist, bicalutamide doesn't bind with T877A AR and has low binding energy with wild type AR (dock score = -2.54 kcal/mol). Bicalutamide is an antagonist to wild and T877A mutated ARs. This reveals that the bicalutamide was bulky enough so that it couldn't bind with AR which has helix-12 in agonist conformation. Bicalutamide has the highest binding energy with W741L mutated AR (dock score = -10.27 kcal/mol) and this mutation causes resistance specifically to bicalutamide. In W741L mutated AR, the nitrile group of bicalutamide formed H-Bond interactions with Arg752, Gln711, HOH108, and Met745. The amine and hydroxyl group formed H-Bond interaction with Leu704 and Asn705 respectively (Figure 3a). It is obvious that the mutated leucine amino acid (W741L) is much smaller in size than the tryptophan amino acid, so we tried to compare the binding mode of bicalutamide in presence of leucine (W741L AR) and tryptophan (wild type AR). The binding comparison was carried out by superimposing the bound conformation of bicalutamide in W741L and wild type (helix-12 truncated) ARs (Figure 3b). The truncated AR has wild type tryptophan amino acid. The B ring of bicalutamide attains a different conformation in the presence of tryptophan and leucine. In the presence of tryptophan, the B ring of bicalutamide was shifted nearer to helix-12 amino acid Ile 899, indicating a possible

steric clash with it. In the presence of leucine (mutated AR) the B ring of bicalutamide was well inside the ligand binding pocket and far from the helix-12. This indicates that the W741L mutation also keeps the bulky B ring of bicalutamide within the binding pocket so that it could switch the antagonist activity of bicalutamide into an agonist.

The unexplained mystery of the mutations T877A and W741L was that flutamide, which is comparatively a smaller ligand than bicalutamide, could antagonize W741L mutated AR but not the T877A AR. Similarly bicalutamide could antagonize T877A mutated AR but not the W741L mutated AR. The possible explanation to this is, the mutations might be specific to the binding conformation of their respective antagonist. For instance, the B ring of bicalutamide was involved in steric interference with helix-12. So, the W741L mutation could be deliberate to keep the B ring of the bicalutamide away from the helix-12. In case of flutamide, there could be some other group involved in the steric interference with helix-12 and hence flutamide works as an antagonist in W741L mutated AR.

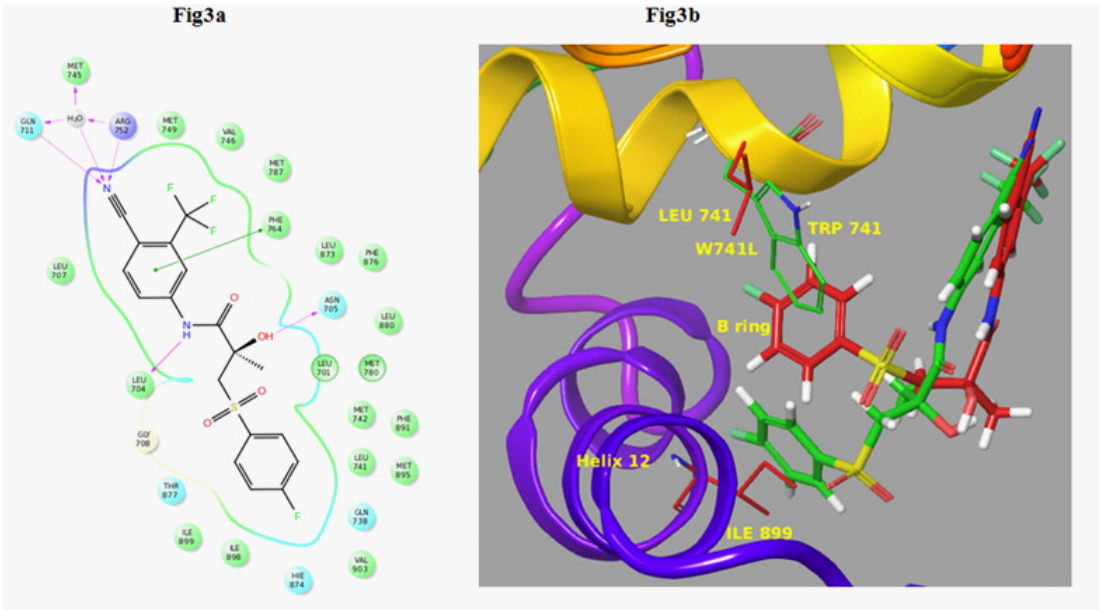


Figure 3: Binding mode of bicalutamide.

3a: Bicalutamide forms H-Bond interaction with Arg752, Gln711, Leu704, Met745 (bridged through HOH) and Asn705 in W741L mutated AR. **3b:** Figure represents the superimposition of the wild (green colored carbon) and W741L mutated (red colored carbon) ARs and the difference in binding mode of bicalutamide. Note the lack of steric clashes of bicalutamide with helix-12 in the mutated type, which in turn leads to resistance.

Enzalutamide

Enzalutamide is a novel AR antagonist which was recently approved and doesn't have the resistance with T877A and W741L mutated ARs. Enzalutamide doesn't have binding affinity with wild, T877A and W741L, which all have helix-12 in agonist conformation. The enzalutamide has binding affinity only with helix-12 truncated AR (dock score = -5.87 kcal/mol). This reveals that enzalutamide has enough bulk that prevents it from binding with ARs that has helix-12 in agonist conformation. The nitrile group of enzalutamide forms H-Bond interactions with Arg752 and Gln711 (Figure 4). This also proves that the truncated helix-12 model is a viable way to identify novel AR antagonists by structure based virtual screening method.

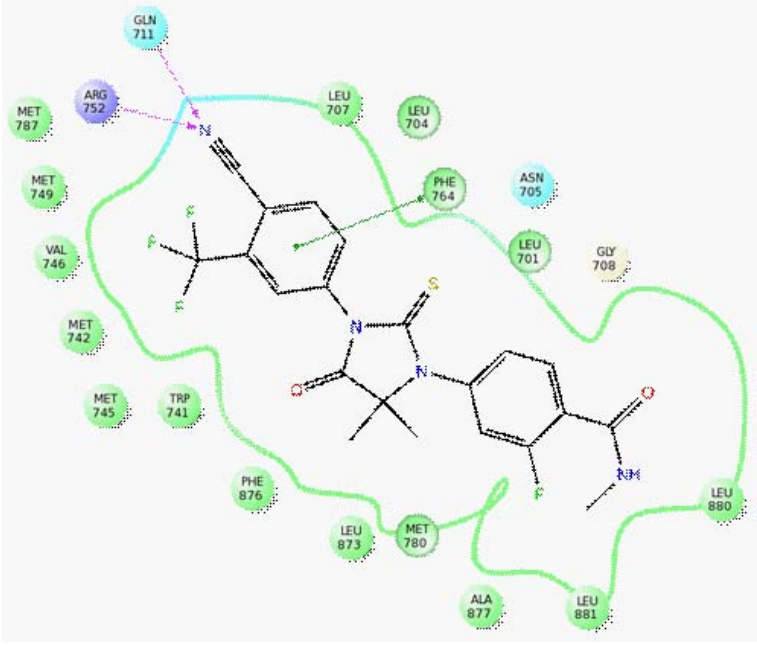


Figure 4: Binding interactions of enzalutamide.

Enzalutamide forms H-Bond interactions with Arg752 and Gln711 with helix-12 truncated AR.

Derivation of Structure Based Drug Design Method

The mutation increases the space within the ligand binding pocket so that it could accommodate for larger molecules without disturbing the helix-12 [44]. The 12th helix truncated AR model could then be used to screen chemical databases (ZINC, SPECS etc) for the identification of novel AR antagonist scaffold. The AR with helix-12 closed conformation could also be included in the docking process to eliminate the ligands that also have good binding score with helix-12 truncated AR. This step will eliminate the less bulk and highly flexible ligands so that the identified ligands might overcome the space increasing mutations that occur in the ligand binding pocket. This way, we can also differentiate between an AR agonist/partial agonist from pure antagonist.

The novel AR antagonist should also form essential H-Bond interactions with the AR for specific binding. In general, the AR ligands could form H-Bond interactions with amino acids Arg752, G1n711, Met745, Leu704, Asn705, and Thr877. The amino acids Arg752, Gln711, and Met745 are present deep inside the binding pocket. The amino acids Thr877 and Asn705 are present nearer to helix-12. An AR antagonist should preferentially form H-Bond with the deep residues, Arg752, Gln711, and Met745 either directly or mediated through a water molecule (bridged H-Bond). The H-Bond interactions with Asn705 and Thr877 might prevent the antagonist from abutting (steric clash) the helix-12 [45]. In the docking studies, flutamide and bicalutamide which exhibit resistance with the mutated AR had formed H-Bond interaction with Asn705 and Thr877 amino acids. Enzalutamide, which doesn't have resistance with Thr877 or W741L mutated ARs, formed H-Bond interaction with Arg752 and Gln711 and didn't form H-Bond interaction with either Asn705 or Thr877.

CONCLUSION

The point mutations in AR increase the binding affinity for the antagonist and decrease the binding affinity for agonist androgens. The mutation converts the antagonist into agonist, which promotes the growth of the cancer cells. Hence, it is essential to design an antagonist that will tolerate the point mutation that may occur during the course of the treatment. This is possible by designing AR antagonists that possess optimal bulk. The bulky groups should also attain optimal conformation in order to abut helix-12. The method described in this manuscript is capable of identifying new chemical entities that remain impervious to mutation in the AR ligand binding domain.

References

1. Watabe T, Lin M, Ide H, Donjacour AA, Cunha GR. Growth, regeneration, and tumorigenesis of the prostate activates the PSCA promoter. Proc Natl Acad Sci U S A. 2002; 99: 401-406.
2. Feldman BJ, Feldman D. The development of androgen-independent prostate cancer. Nat Rev Cancer. 2001; 1: 34-45.
3. Attard G, Reid AH, Yap TA, Raynaud F, Dowsett M, et al. Phase I clinical trial of a selective inhibitor of CYP17, abiraterone acetate, confirms that castration-resistant prostate cancer commonly remains hormone driven. J Clin Oncol. 2008; 26: 4563-4571.
4. Tilley WD, Buchanan G, Hickey TE, Bentel JM. Mutations in the androgen receptor gene are associated with progression of human prostate cancer to androgen independence. Clin Cancer Res. 1996; 2: 277-285.
5. Taplin ME, Bubley GJ, Shuster TD, Frantz ME, Spooner AE, Ogata GK, et al. Mutation of the androgen receptor gene in metastatic androgen independent prostate cancer. N Engl J Med. 1995; 332: 1393-1398.
6. Brown CJ, Goss SJ, Lubahn DB, Joseph DR, Wilson EM, French FS, et al. Androgen receptor locus on the human X chromosome: regional localization to Xq11-12 and description of a DNA polymorphism. Am J Hum Genet. 1989; 44: 264-269.
7. Zeegers MP, Kiemeney LA, Nieder AM, Ostrer H. How strong is the association between CAG and GGN repeat length polymorphisms in the androgen receptor gene and prostate cancer risk? Cancer Epidemiol Biomarkers Prev. 2004; 13: 1765-1771.
8. Matias PM, Donner P, Coelho R, Thomaz M, Peixoto C, et al. Structural evidence for ligand specificity in the binding domain of the human androgen receptor. Implications for pathogenic gene mutations. J Biol Chem. 2000; 275: 26164-26171.
9. Pereira-de-Jésus-Tran K, Côté PL, Cantin L, Blanchet J, Labrie F, et al. Comparison of crystal structures of human androgen receptor ligand binding domain complexed with various agonists reveals molecular determinants responsible for binding affinity. Protein Sci. 2006; 15: 987-999.
10. Fang Y, Fliss AE, Robins DM, Caplan AJ. Hsp90 regulates androgen receptor hormone binding affinity in vivo. J Biol Chem. 1996; 271: 28697-28702.

11. Saporita AJ, Zhang Q, Navai N, Dincer Z, Hahn J. Identification and characterization of a ligand-regulated nuclear export signal in androgen receptor. J Biol Chem. 2003; 278: 41998-42005.

12. He B, Kemppainen JA, Wilson EM. FXXLF and WXXLF sequences mediate the NH2-terminal interaction with the ligand binding domain of the androgen receptor. J Biol Chem. 2000; 275: 22986-22994.

13. Bevan CL, Hoare S, Claessens F, Heery DM, Parker MG. The AF1 and AF2 domains of the androgen receptor interact with distinct regions of SRC1. Mol Cell Biol. 1999; 19: 8383-8392.

14. Heinlein CA, Chang C. Androgen receptor (AR) coregulators: an overview. Endocr Rev. 2002; 23: 175-200.

15. Hodgson MC, Shen HC, Hollenberg AN, Balk SP. Structural basis for nuclear receptor corepressor recruitment by antagonist-liganded androgen receptor. Mol Cancer Ther. 2008; 7: 3187-3194.

16. Heidenreich A, Bellmunt J, Bolla M, Joniau S, Mason M, Matveev V, et al. EAU guidelines on prostate cancer. Part 1: screening, diagnosis, and treatment of clinically localised disease. Eur Urol. 2011; 59: 61-71.

17. Heidenreich A, Bastian PJ, Bellmunt J, Bolla M, Joniau S, et al. EAU guidelines on prostate cancer. Part II: Treatment of advanced, relapsing, and castration-resistant prostate cancer. Eur Urol. 2014; 65: 467-479.

18. Stanbrough M, Bubley GJ, Ross K, Golub TR, Rubin MA, Penning TM, et al. Increased expression of genes converting adrenal androgens to testosterone in androgen-independent prostate cancer. Cancer Res. 2006; 66: 2815-2825.

19. Ji Q, Chang L, Stanczyk FZ, Ookhtens M, Sherrod A, et al. Impaired dihydrotestosterone catabolism in human prostate cancer: critical role of AKR1C2 as a pre-receptor regulator of androgen receptor signaling. Cancer Res. 2007; 67: 1361-1369.

20. Wang LG, Johnson EM, Kinoshita Y, Babb JS, Buckley MT, et al. Androgen receptor overexpression in prostate cancer linked to Pur alpha loss from a novel repressor complex. Cancer Res. 2008; 68: 2678-2688.

21. Dehm SM, Schmidt LJ, Heemers HV, Vessella RL, Tindall DJ. Splicing of a novel androgen receptor exon generates a constitutively active androgen receptor that mediates prostate cancer therapy resistance. Cancer Res. 2008; 68: 5469-5477.

22. Sircar K, Yoshimoto M, Monzon FA, Koumakpayi IH, Katz RL, et al. PTEN genomic deletion is associated with p-Akt and AR signalling in poorer outcome, hormone refractory prostate cancer. J Pathol. 2009; 218: 505-513.

23. Zhang W, Kapusta LR, Slingerland JM, Klotz LH. Telomerase activity in prostate cancer, prostatic intraepithelial neoplasia, and benign prostatic epithelium. Cancer Res. 1998; 58: 619-621.

24. Nesslinger NJ, Shi XB, deVere-White RW. Androgen-independent growth of LNCaP prostate cancer cells is mediated by gain-of-function mutant p53. Cancer Res. 2003; 63: 2228-2233.

25. Jennbacken K, Tesan T, Wang W, Gustavsson H, Damber JE, et al. N-cadherin increases after androgen deprivation and is associated with metastasis in prostate cancer. Endocr Relat Cancer. 2010; 17: 469-479.

26. Hessels D, Smit FP, Verhaegh GW, Witjes JA, Cornel EB, et al. Detection of TMPRSS2-ERG fusion transcripts and prostate cancer antigen 3 in urinary sediments may improve diagnosis of prostate cancer. Clin Cancer Res. 2007; 13: 5103-5108.

27. Rahim S, Beauchamp EM, Kong Y, Brown ML, Toretsky JA, et al. YK-4-279 inhibits ERG and ETV1 mediated prostate cancer cell invasion. PLoS One. 2011; 6: e19343.

28. Taplin ME, Rajeshkumar B, Halabi S, Werner CP, Woda BA, et al. Cancer and Leukemia Group B Study 9663. Androgen receptor mutations in androgen-independent prostate cancer: Cancer and Leukemia Group B Study 9663. J Clin Oncol. 2003; 21: 2673-2678.

29. Balk SP. Androgen receptor as a target in androgen-independent prostate cancer. Urology. 2002; 60: 132-138.

30. Hara T, Miyazaki J, Araki H, Yamaoka M, Kanzaki N, et al. Novel mutations of androgen receptor: a possible mechanism of bicalutamide withdrawal syndrome. Cancer Res. 2003; 63: 149-153.

31. Tan J, Sharief Y, Hamil KG, Gregory CW, Zang DY, et al. Dehydroepiandrosterone activates mutant androgen receptors expressed in the androgen-dependent human prostate cancer xenograft CWR22 and LNCaP cells. Mol Endocrinol. 1997; 11: 450-459.

32. Urushibara M, Ishioka J, Hyochi N, Kihara K, Hara S. Effects of steroidal and non-steroidal antiandrogens on wild-type and mutant androgen receptors. Prostate. 2007; 67: 799-807.

33. Fenton MA, Shuster TD, Fertig AM. Functional characterization of mutant androgen receptors from androgen-independent prostate cancer. Clin Cancer Res. 1997; 3: 1383-1388.

34. Taplin ME, Bubley GJ, Ko YJ, Small EJ, Upton M. Selection for androgen receptor mutations in prostate cancers treated with androgen antagonist. Cancer Res. 1999; 59: 2511-2515.

35. LigPrep version 3.1, Schrödinger, LLC, New York, NY. 2014.

36. Maestro version 9.8, Schrödinger, LLC, New York, NY. 2014.

37. Glide, version 3.3, Schrödinger, LLC, New York, NY. 2014.

38. PRIME version 3.7, Schrödinger, LLC, New York, NY. 2014.

39. Bohl CE, Gao W, Miller DD, Bell CE, Dalton JT. Structural basis for antagonism and resistance of bicalutamide in prostate cancer. Proc Natl Acad Sci USA. 2005; 102: 6201-6206.

40. Bohl CE, Miller DD, Chen J, Bell CE, Dalton JT. Structural basis for accommodation of nonsteroidal ligands in the androgen receptor. J Biol Chem. 2005; 280: 37747-37754.

41. Shiau AK, Barstad D, Loria PM, Cheng L, Kushner PJ. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. Cell. 1998; 95: 927-937.

42. Kauppi B, Jakob C, Färnegårdh M, Yang J, Ahola H, Alarcon M, et al. The three-dimensional structures of antagonistic and agonistic forms of the glucocorticoid receptor ligand-binding domain: RU-486 induces a transconformation that leads to active antagonism. J Biol Chem. 2003; 278: 22748-22754.

43. Zhao XY, Boyle B, Krishnan AV, Navone NM, Peehl DM. Two mutations identified in the androgen receptor of the new human prostate cancer cell line MDA PCa 2a. J Urol. 1999; 162: 2192-2199.

44. Osguthorpe DJ, Hagler AT. Mechanism of androgen receptor antagonism by bicalutamide in the treatment of prostate cancer. Biochemistry. 2011; 50: 4105-4113.

45. Guo C, Pairish M, Linton A, Kephart S, Ornelas M. Design of oxobenzimidazoles and oxindoles as novel androgen receptor antagonists. Bioorg Med Chem Lett. 2012; 22: 2572-2578.

Austin Publishing Group

Hepatitis C Viral Polymerase Inhibition Using Directly Acting Antivirals: A Computational Approach

Elfiky AA*[1,2], Gawad WA[1] and Elshemey WM[1]

[1]Biophysics Department, Faculty of Science, CairoUniversity, Egypt

[2]Biochemistry and Structural Biology Department, Center of Molecular Protein Science CMPS, Lund University, Sweden

***Corresponding author:** Elfiky AA, Biochemistry and Structural Biology Department, Center of Molecular Protein Science CMPS, Lund University, Sweden, Tel: +201115121528; Fax: +46736478322; Email: abdo@sci.cu.edu.eg, abdo.mohamed@biochemistry.lu.se

Published Date: December 01, 2016

Abbreviations: NS5B-Non-Structural 5B protein; RdRp-RNA dependent RNA polymerase; SVR-Sustained Viral Response; DAA -Direct Acting Antivirals; PEG-Poly Ethylene Glycol; HCV-Hepatitis C Virus; NI-Nucleotide Inhibitor; NNI-Non-Nucleotide Inhibitor; HOMO-Highest Occupied Molecular Orbital; LUMO-Lowest Unoccupied Molecular Orbital; PHYRE-Protein Homology/anologY Recognition Engine; SCOP-Structural Classification of Proteins; SAVES-Structure Analysis and Verification Server; QSAR-Quantitative Structure-Activity Relationship.

INTRODUCTION

Hepatitis C Virus (HCV)

Hepatitis C virus, as it appears from its name, is a liver- affecting virus. HCV is a blood prone virus that was first discovered in 1989 by Choo and coworkers. The virus was termed Non-A Non-B hepatitis. HCV can develop liver cirrhosis and reduce the functionality of liver in patients. Developed hepatocellular carcinoma has been recorded in some HCV patients after long periods of

chronic hepatitis viral infection. Today more than 200 million people worldwide are infected with chronic liver diseases that lead to liver cirrhosis and development of hepatocellular carcinoma [1-4]. The worldwide HCV prevalence is around 3% of the population. Egypt has the highest ratio of chronic liver disease prevalence that affects about 14% of the population most of which (90%) belongs to the genotype 4a [2, 5-9].

HCV is a small virus from the Flaviviridae family. It consists of RNA as the genetic material enveloped by protein capsid [6]. HCV genome is approximately 9600 base pairs single stranded RNA that encodes a polyprotein consisting of about 3000 amino acid residues. The polyprotein then cleaved by both viral and host cell proteases to 10 proteins (Figure 1) some of which are part of the structure of the virus [core, E1, E2 and p7], which are called structural proteins. Others have specific functions in viral replication [NS2, NS3, NS4A, NS4B, NS5A and NS5B]. These are termed Non-Structural **(NS)** proteins [2,3,6,10].

HCV genome is characterized by high mutation rate. About six main genotypes are present to date (1, 2, 3, 4, 5 and 6). The nucleotide sequence differs by 31% to 34% among the different genotypes. Genotypes are further classified into subtypes; more than 100 subtypes are present to date (1a, 1b, 2a, 2b etc…). The sequence similarity among different HCV subtypes in each individual genotype is about 90%. HCV circulates in infected patient's blood in the form of a number of different but closely related variants called quasi-species [11].

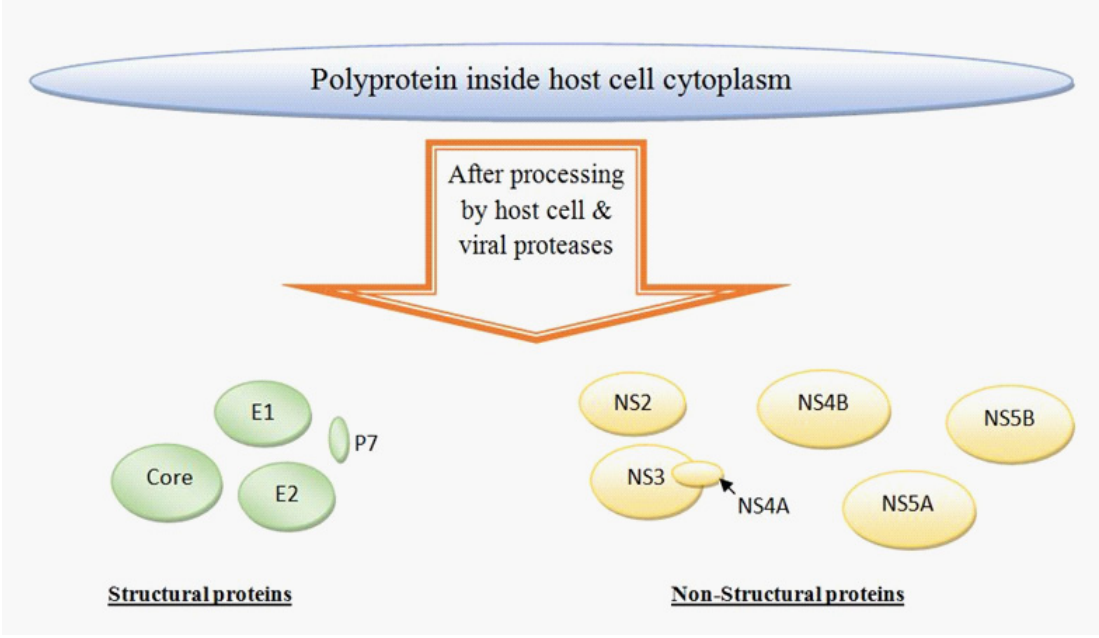


Figure 1: HCV polyprotein before and after cleavage by viral and host cell proteases.

HCV RNA-Dependent RNA Polymerase (RdRp)

HCV RNA dependent RNA polymerase **(RdRp)** is a part of the NS5B protein. It plays an important role in viral replication cycle. NS5B represents an excellent target for selective HCV inhibitors because its biochemical activity is limited to the RNA viruses with no effect on mammalian cells [12].

NS5B (Figure 2) is 68 KDa tail-anchored protein with an alpha helical trans-membrane domain consisting of 21C-terminal amino acids [13]. The domain architecture of NS5B RdRp is the same as other polymerases consisting of thumb, fingers and palm domains resembling the right hand [14,15]. The palm domain contains two consecutive metal binding aspartates that form the active site motif GDD (G88, D89 and D90) and carry out the nucleotidyl transfer reaction. Fingers and thumb domains regulate nucleic acid binding. Beside the active site, several other pockets that act as allosteric binding sites [12,15,16].

Due to the high mutation rate characterizing HCV genome, the production of efficient DAAs that inhibit NS5B RdRp protein remains challenging. To resolve this problem and improve the viral response a combination therapy was suggested by many authors with different drugs having different binding modes of action [4, 17].

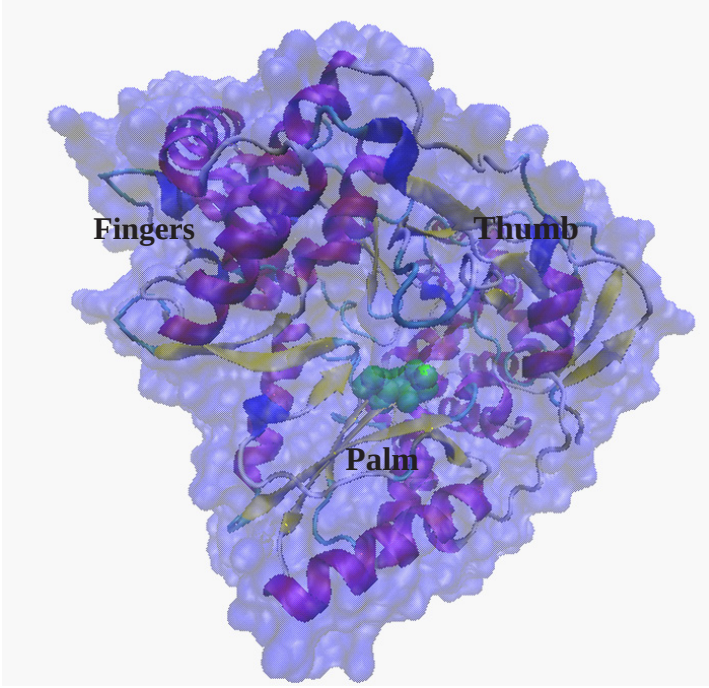


Figure 2: Structure of the protein NS5B polymerase from PDB file 4AEP downloaded from the Protein Data Bank. The active site motif GDD of the RdRp part is represented by green Van Der Waal **(VDW)** spheres. The figure was generated using Visulaising Molecular Dynamics **(VMD)** software.

HCV Treatment

Rational drug therapy

The rational drug therapy for HCV until five years ago was only the double therapy. It is a combination of PEG related interferon alpha and the wide range antiviral Ribavirin **(PEG-IFN/RBV)**. This regimen gave varying Sustained Virologic Response **(SVR)** rates that depend on the viral genotype. SVR is up to 80% in genotypes 2 and 3, 60-70% for genotype 4 but only 40-50% in genotype 1 [9,18]. Unfortunately, the combination therapy is expensive and not tolerated by some patients. Interferon develops diversities of side effects that may lead to stopping the medication in some cases [6,10,19].

Due to the above-mentioned reasons, researchers started to direct their attention to interferon-free regimens. They worked on drugs that directly act on specific proteins that are important in viral replication. These types of drugs are called Direct Acting Antivirals **(DAAs)**.

Direct Acting Antivirals (DAAs)

The use of Direct Acting Antiviral **(DAA)** drugs that act on specific viral and/or host cell proteins gives good results in many cases. In the year, 2011, the FDA approved two drugs for the treatment of HCV genotype 1 in combination with interferon alpha and Ribavirin. The two approved drugs (Telaprevir and Boceprevir) are DAAs that target NS3 serine protease domain of the NS3 protein of HCV [20]. Sofosbuvir is a nucleotide NS5B polymerase inhibitor that was approved by FDA in December 2013 as a free drug or in combination with interferon against genotype 1.

For almost all therapies that were developed to act on the viral proteins, drug resistance occurred due to the high mutation rate induced by the nature of HCV (a single stranded RNA virus). One can mix a cocktail of DAAs to overcome resistance, putting into consideration the toxicity of the mixed drugs [21].

There are two types of DAAs against HCV NS5B polymerase. The first type is called Nucleotide Inhibitors **(NIs)** in which the nucleotide-like analogue is introduced into HCV NS5B polymerase active site to stop the polymerization process. NIs are classified into two subtypes; sugar modified nucleotide analogues and nitrogenous base modified analogous. Some drugs that act on NS5B polymerase are not related to the structure of nucleotides. These drugs called Non-Nucleotide Inhibitors **(NNIs)**. This is the second type of DAAs against HCV NS5B polymerase.

NIs are successful candidates in the treatment of HIV and herpes viruses. NIs compete with the nucleotides (natural substrate: Adenine, Guanine, Cytosine and Uracil) on HCV polymerase active site. They are, generally, prodrugs that are activated by phosphorylation inside the host cell. Once an NI becomes attached to the polymerase active site it stops the polymerization process, hence they are termed chain terminator inhibitors. They can also interfere with the cellular proliferative machinery [12, 22]. IDX-184, R7128 and Sofosbuvir (Figure 3) are examples of NIs against HCV.

These drugs are now either under clinical trials phases II (IDX-184 and R7128) or already approved (Sofosbuvir), [4,12,21,23-25]. These DAAs give good results in terms of increasing the SVR rate when administered in combination with rational regimen of double therapy [10, 26].

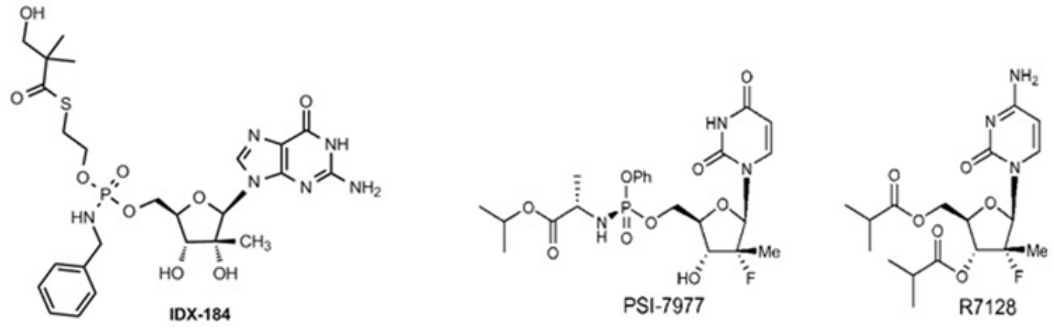


Figure 3: The structures of some nucleotide inhibitors, Sofosbuvir (PSI-7977), IDX-184 and R7128.

Molecular Modeling

Molecular modeling can be simply considered as a range of computerized techniques. These are based on the basic laws of physics and experimental data which can be used either to analyze molecules (number and types of atoms, bond, bond lengths, angles and dihedral angles) or molecular systems (nucleophilicity, electrophilicity and electrostatic potentials). Moreover,

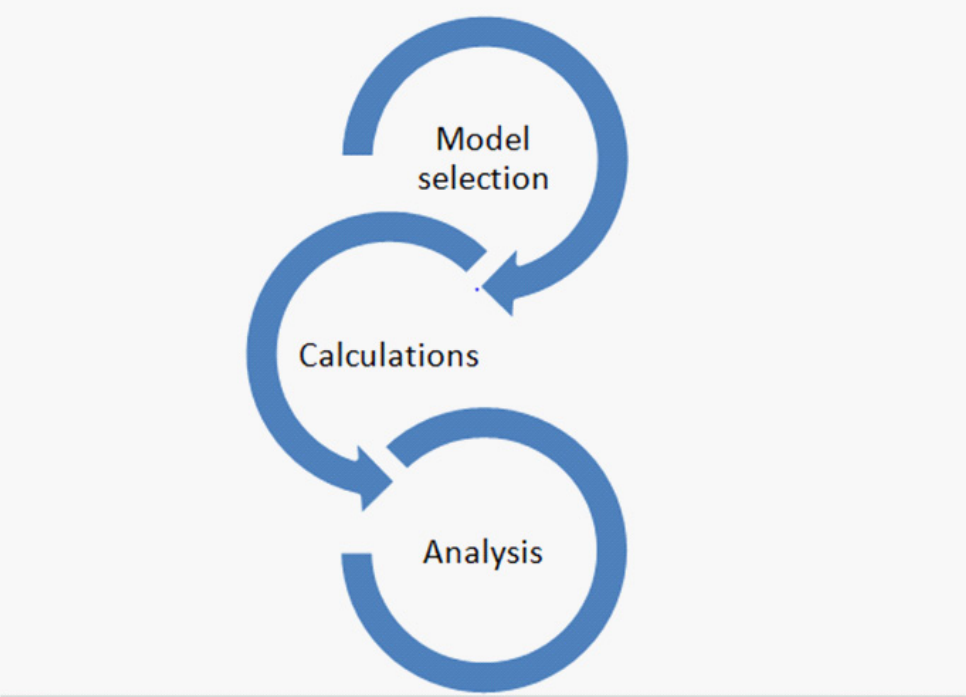
it can predict molecular and biological properties which are useful in the understanding of structure-activity relationship and in rational drug design [27].

As concluded in the scheme below, molecular modeling consists of three stages: The model selection, with which the calculations will be carried out. This step is governed by the complexity of the system and the computational time requirement. One may make calculations using Molecular Mechanics **(MM)**, Quantum Mechanics **(QM)** or hybrid MM/QM. If working on small molecules or peptides, one may use QM or semi-empirical QM [28]. If working on large molecules (like proteins or DNA), one should use an MM model. If working on the active site of large molecules, one may use the hybrid MM/QM model.

The selection of the calculation type is the second stage. Different calculations are possible depending on the goal of the experiment. Among the very large number of available calculations are; geometry optimization, vibrational spectra calculation, NMR spectra calculation and single point energy calculations.

The final and most important stage is the analysis of results. This stage will provide answers about the problem that might not be solvable by experimental work. One should perform extensive

analysis of the results in order to reach valuable conclusions about the investigated system. To get better results one should select the most appropriate model for calculations. This depends on the complexity of system and the available computational power [29].



Molecular Modeling scheme

The discovery of new drugs is the main target for pharmacologists and medicinal chemists. Drugs are chemical substances that can be used both for treatment and for diagnosis of a disease. In addition; it can prevent the development of disease in humans, animals and plants. The function of drugs is to inhibit or to enhance certain physiological functions. The biological and pharmacological effects of drugs are either helpful or harmful for living organisms. Drugs interact with specific targets in living organisms such as enzymes, receptors, nucleic acids, channels or other biological macromolecules [30]. The discovery of new pharmacological compounds requires the design and synthesis of drug, studying its physicochemical and biophysical properties in addition to its pharmaceutical functions. These studies improve drug safety and biological activity while reducing adverse side effects. The development of drugs has several strategies. These strategies involve either a change in the shape of the drug in order to fit into its active site receptor or a change in its pharmaceutical properties, which include Absorption, Distribution, Metabolism and Excretion **(ADME)**. These strategies require the synthesis of large number of compounds and substitutions that consume time and money.

Quantitative Structure-Activity Relationship **(QSAR)** is a technique that quantifies the relationship between a physicochemical property of a drug and its biological activity. QSAR is useful

for optimizing the groups that modulate the potency of a drug. It is based on the determination of mathematical equations that express the biological activities in terms of molecular descriptors such as the logarithm of partition coefficient (log P), steric constituent constant (Es) and molar refractivity. QSAR also may make use of structural indexes obtained by quantum mechanics such as Highest Occupied Molecular Orbital **(HOMO)** energies, Lowest Unoccupied Molecular Orbital **(LUMO)** energies, total dipole moments, charge, molecular polarizability, electronegativity and frontier orbital energies [25,27,30-33].

QSAR descriptors are not universal and depend on the nature of chemical structures or process involved. Once a correlation between structure and activity is found for a compound or group of compounds, the computer can be used to make screening in order to select structures with the desired properties. It is possible to select the most promising compounds to synthesize and test in the laboratory. A combination of QSAR and molecular modeling approach is the key for success in Computer Aided Drug Design **(CADD)** and to understand drug-receptor interactions [33].

CASE STUDY DESCRIPTION

In the following section, the use of molecular modeling combined with QSAR to study the binding of different drugs (Nucleotide Inhibitors) to NS5b RdRp of HCV from different genotypes will extensively illustrated. Moreover, a comparison between the binding energies of these drugs and native nucleotides using the same technique is also presented.

Steps of Computer Aided Drug Design (CADD)

Protein sequence analysis

Sequence comparison and analysis is important in rational drug design. As shown in previous work [25,33] the amino acid sequence around the active site moiety GDD and the surrounding environment is conserved (5 Å region around the GDD motif). Since the structure of the active site of polymerase is conserved among different HCV genotypes, it would be possible to target the active site in different genotypes with the same inhibitor. However, some studies on NS5B RdRp with NIs show different results for different genotypes. This may be due to the effect of mutations on the cavity at the active site. These mutations don't occur in the active site environment but aside from it and probably lead to drug disability to inhibit the protein or at least lead to decreased inhibition [4,10,34].

There are different methods that can be used for sequence comparison. For example, using Visualizing Molecular Dynamics **(VMD)** software, sequence alignments may be carried out for the sequences with the help of Clustal W program in multiseq extension or using the web based service of CLUSTALW 2 [25, 33].

Homology modeling

Homology modeling, also called comparative protein modeling or knowledge-based modeling, is the process by which a 3-dimentional model of a target sequence being built based

on a homologue of experimentally solved structure (experimental processes include X-ray crystallography, solution Nuclear Magnetic Resonance **[NMR]** and Electron Microscopy **[EM]**). Homology modeling relies on the identification of one or more known protein structures likely to resemble the structure of the query sequence and on the production of an alignment that maps residues in the query sequence to residues in the template sequence [35]. The sequence alignment and template structure are then used to produce a structural model of the target [36].

A target (or query) sequence is the primary sequence of a protein whose structure has to be modeled. When first loaded in the workspace, it is provisionally drawn as a long helix. A template structure, or simply a template, is an experimentally solved structure used as a scaffold to model the structure of the target sequence. Template sequence is the primary sequence of a template. The quality of the homology model is dependent on the quality of the sequence alignment and template structure. The approach can be complicated by the presence of alignment gaps that indicate a structural region present in the target but not in the template, and by structure gaps in the template that arise from poor resolution in the experimental procedure (usually X-ray crystallography) used to solve the structure.

After modeling, one should check the models for errors using different mechanisms including 3D structure related properties (such as bond angles, length, Ramachandran plots). This process is called "Protein Model Validation". On the web, there are several servers built for helping researchers to check their structures for errors. Structural Analysis and Verification Server **(SAVES)** is one of them, in which the protein three-dimensional structure **(PDB file)** is uploaded and checked by built-in programs which check the PDB file for errors where each program produces its own result. Based on the given results one can judge the validity of his own protein model. For example, SAVES server contains the program PROCHECK which check the stereo-chemical parameters such as Ramachandran plots, main and side chain parameters, residue properties, G-factor dihedrals, main chain bond angle, and bond length[37]. The programVERIFY-3D checks the residual environment [38]. ERRAT program generates overall quality factor of the protein model [39]. PROVE program checks atomic volumes and calculates the atomic Z-score[40]. The program WHATCHECK gives a report for almost all parameters of the uploaded protein structure PDB file [41].

Drug activation

Some drugs are converted to its active form inside target cells. In this study, phosphorylation of Sofosbuvir, IDX-184, R7128 and Ribavirin was performed in silico to become in its active tri-phosphate form [25, 33]. After in silico phosphorylation, the activated Nucleotide Inhibitors **(NIs)** are energy minimized using mechanical chemistry calculation method **(MM3)** followed by the semi-empirical quantum mechanics calculation method **(PM3)**. The use of a low level method (classical mechanical method **(MM3)**) for an initial energy minimization reduces the time of calculation needed by the higher level method **(PM3)** for energy minimization. Infrared

vibrational spectrum is then calculated at PM3 level in order to ensure that an active form of the drug is real (no negative vibrations). After optimization, infrared vibrational spectrum calculation is performed in order to ensure the structures being real.

QSAR descriptors calculation

In some previous studies, Quantitative Structure Activity Relationship **(QSAR)** descriptors are calculated for selected DAA drugs (Sofosbuvir, IDX-184 and R7128) and Ribavirin in addition to their parent nucleotide tri-phosphates (Guanine, Cytosine and Uracil) for comparison [25, 33]. QSAR calculations are carried out at PM3 level using computational chemistry integrated platform software SCIGRESS [42]. The calculated descriptors are: dipole moment, the logarithm of partition coefficient (Log P), electron affinity, molar refractivity, ionization potential, solvent accessible surface area, volume, total energy, heat of formation, Highest Occupied Molecular Orbitals **(HOMO)**, Lowest Unoccupied Molecular Orbitals **(LUMO)** and frontier energy gap (ΔE = LUMO - HOMO).

Drug-protein interaction

The final and very important step in rational drug design is the study of drug-protein interaction. The drug in its active form may form covalent or non-covalent bonds with the active site amino acid or the amino acid around the active site cavity. Some hydrophobic or Van der Walls interactions may stabilize the drug in the protein cavity. The interaction potency between the drug and the protein is the factor that one can depend on when comparing different drugs against specific protein.

One example of drug/protein interaction is the interaction between IDX-184 and NS5B protein model (Figure 4). A hydrogen bond is formed between the drug in its active form (Tri-phosphate) and the amino acid S59 of the polymerase active site environment. In addition, H-bonds are formed between active site environment's amino acids. Another force of interaction arises from weak interactions formed between the two Mg^{+2} ions and both of the triphosphates' oxygen and the two aspartic acids (active site amino acids) oxygen atoms.

Figure 4: The interaction between IDX-184 and the active site of NS5b RdRp D89 and D90 showing the formation of H-bond between the drug and S59 in addition to the coordination bonds between the two Mg+2 and the oxygen atoms of both D89, D90 of the protein and Phosphate group in the drug.

CONCLUSION

Molecular modeling represents a promising technique that had a very high momentum of development in the last decade. This progress was in fact, related to the rapid improvement in the hardware of new computers in the market. In addition, software improvement provided a second source making molecular modeling the best choice in different areas of research.

In presented case study, Computer Aided Drug Design CADD was utilized in order to provide an insight about the binding of Sofosbuvir, IDX-184, R7128 and Ribavirin to HCV NS5b active site. The results showed diversity among different drugs and different genotypes. These findings emerged from both QSAR calculated parameters and interaction energies calculated for the binding of the drugs to the active site of the polymerase. These results were in agreement with the experimental data obtained from patients infected by different genotypes of HCV, where different responses to the same drug were recorded.

CADD results implied that IDX-184 represented a promising drug against all studied genotypes. Also, all of the studied drugs (Sofosbuvir, IDX-184 and R7128) were able to interact more effectively with viral polymerase than ribavirin of the dual therapy. Hence, these drugs were better than ribavirin in competing the nucleotides for binding to HCV polymerase.

GLOSSARY

Active site environment: Describes a 5*Å* region around the active site motif GDD. This active site environment complex includes 12 amino acids (including GDD motif), two Mg^{+2} ions and one of the Nucleotides, NIs or ribavirin.

Frontier energy gap: The difference in energy between HOMO and LUMO. it is high in more stable structures.

Heat of formation: The change in enthalpy accompanying the formation of one mole of a compound from its elements in their natural and stable states, under standard condition of one atmosphere at a given temperature.

Ionization potential: It is the energy required to ionize an atom. High values of the ionization potential means a more stable structure.

Sustained Virologic Response: It is defined as aviremia (Lack of virus in the blood plasma) 24 weeks after completion of antiviral therapy for chronic hepatitis C virus infection.

References

1. Firpi RJ, Nelson DR. Current and future hepatitis C therapies. Arch Med Res. 2007; 38: 678-690.
2. Lemon SM, McKeating JA, Pietschmann T, Frick DN, Glenn JS. Development of novel therapies for hepatitis C. Antiviral Res. 2010; 86: 79-92.
3. Das D, Hong J, Chen SH, Wang G, Beigelman L. Recent advances in drug discovery of benzothiadiazine and related analogs as HCV NS5B polymerase inhibitors. Bioorg Med Chem. 2011; 19: 4690-4703.
4. Yang PL, Gao M, Lin K, Liu Q, Villareal VA. Anti-HCV drugs in the pipeline. Curr Opin Virol. 2011; 1: 607-616.
5. Chamberlain RW, Adams N, Saeed AA, Simmonds P, Elliott RM. Complete nucleotide sequence of a type 4 hepatitis C virus variant, the predominant genotype in the Middle East. J Gen Virol. 1997; 78: 1341-1347.
6. De Francesco R, Tomei L, Altamura S, Summa V, Migliaccio G. Approaching a new era for hepatitis C virus therapy: inhibitors of the NS3-4A serine protease and the NS5B RNA-dependent RNA polymerase. Antiviral Res. 2003; 58: 1-16.
7. Yan S, Appleby T, Larson G, Wu JZ, Hamatake R, et al. Structure-based design of a novel thiazolone scaffold as HCV NS5B polymerase allosteric inhibitors. Bioorg Med Chem Lett. 2006; 16: 5888-5891.
8. Bahgat MM, Ibrahim AA, Abd-Elshafy DN, Mesalam AA, Gewaid HE. Characterization of NS3 protease from an Egyptian HCV genotype 4a isolate. Arch Virol. 2009; 154: 1649-1657.
9. Massariol MJ, Zhao S, Marquis M, Thibeault D, White PW. Protease and helicase activities of hepatitis C virus genotype 4, 5 and 6 NS3-NS4A proteins. Biochem Biophys Res Commun. 2010; 391: 692-697.
10. De Francesco R, Carfí A. Advances in the development of new therapeutic agents targeting the NS3-4A serine protease or the NS5B RNA-dependent RNA polymerase of the hepatitis C virus. Adv Drug Deliv Rev. 2007; 59: 1242-1262.
11. Martell M, Esteban JI, Quer J, Genescà J, Weiner A. Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. J Virol. 1992; 66: 3225-3229.
12. Mayhoub AS. Hepatitis C RNA-dependent RNA polymerase inhibitors: A review of structure-activity and resistance relationships; different scaffolds and mutations. Bioorg Med Chem. 2012; 20: 3150-3161.
13. Suzuki T, Ishii K, Aizaki H, Wakita T. Hepatitis C viral life cycle. Adv Drug Deliv Rev. 2007; 59: 1200-1212.
14. Doublié S, Ellenberger T. The mechanism of action of T7 DNA polymerase. Curr Opin Struct Biol. 1998; 8: 704-712.
15. Chinnaswamy S, Cai H, Kao C. An update on small molecule inhibitors of the HCV NS5B polymerase: effects on RNA synthesis in vitro and in cultured cells, and potential resistance in viral quasispecies. Virus Adapt Treat. 2010; 2: 73-89.

16. O'Farrell D, Trowbridge R, Rowlands D, Jäger J. Substrate complexes of hepatitis C virus RNA polymerase (HC-J4): structural evidence for nucleotide import and de-novo initiation. J Mol Biol. 2003; 326: 1025-1035.

17. Ohno T, Lau JY. The "gold-standard," accuracy, and the current concepts: hepatitis C virus genotype and viremia. Hepatology. 1996; 24: 1312-1315.

18. Sarrazin C, Hézode C, Zeuzem S, Pawlotsky JM. Antiviral strategies in hepatitis C virus infection. J Hepatol. 2012; 56: S88-100.

19. Beaulieu PL, Gillard J, Jolicoeur E, Duan J, Garneau M, Kukolj G, et al. From benzimidazole to indole-5-carboxamide Thumb Pocket I inhibitors of HCV NS5B polymerase. Part 1: indole C-2 SAR and discovery of diamide derivatives with nanomolar potency in cell-based subgenomic replicons. Bioorg Med Chem Lett. 2011; 21: 3658-3663.

20. Thompson AJ, Locarnini SA, Beard MR. Resistance to anti-HCV protease inhibitors. Curr Opin Virol. 2011; 1: 599-606.

21. Gelman MA, Glenn JS. Mixing the right hepatitis C inhibitor cocktail. Trends Mol Med. 2011; 17: 34-46.

22. Perrone P, Daverio F, Valente R, Rajyaguru S, Martin JA, et al. First example of phosphoramidate approach applied to a 4'-substituted purine nucleoside (4'-azidoadenosine): conversion of an inactive nucleoside to a submicromolar compound versus hepatitis C virus. J Med Chem. 2007; 50: 5463-5470.

23. Murakami E, Tolstykh T, Bao H, Niu C, Steuer HM. Mechanism of activation of PSI-7851 and its diastereoisomer PSI-7977. J Biol Chem. 2010; 285: 34337-34347.

24. Chen YL, Tang J, Kesler MJ, Sham YY, Vince R. The design, synthesis and biological evaluations of C-6 or C-7 substituted 2-hydroxyisoquinoline-1,3-diones as inhibitors of hepatitis C virus. Bioorg Med Chem. 2012; 20: 467-479.

25. Elfiky AA, Elshemey WM, Gawad WA, Desoky OS. Molecular modeling comparison of the performance of NS5b polymerase inhibitor (PSI-7977) on prevalent HCV genotypes. Protein J. 2013; 32: 75-80.

26. Chevaliez S, Pawlotsky JM. Interferon-based therapy of hepatitis C. Adv Drug Deliv Rev. 2007; 59: 1222-1241.

27. Cohen NC. Guidebook on Molecular Modeling in Drug Design, Academic press, Inc. 1996.

28. Foresman JB, Frisch A. Exploring Chemistry with Electronic Structure Methods, Gaussian Inc., 2nd ed. 1996.

29. Leach AR. Molecular Modelling Principle and Applications, Addison Wesley Longman Limited, Edinburgh Gate, Harlow, Essex CM20 2JE, England. 2001.

30. Saleh NA, Ezat AA, Elfiky AA, Elshemey WM, Ibraheim M. Theoretical Study on Modified Boceprevir Compounds as NS3 protease inhibitors. J Comput Theor Nanos. 2015; 12: 371-375.

31. Ibrahim M, Saleh NA, Hameed AJ, Elshemey WM, Elsayed AA. Structural and electronic properties of new fullerene derivatives and their possible application as HIV-1 protease inhibitors. Spectrochim Acta A Mol Biomol Spectrosc. 2010; 75: 702-709.

32. Ibrahim M, Saleh NA, Elshemey WM, Elsayed AA. Hexapeptide functionality of cellulose as NS3 protease inhibitors. Med Chem. 2012; 8: 826-830.

33. Elfiky AA, Elshemey WM, Gawad WA. 2'-Methylguanosine prodrug (IDX-184), Phosphoramidate prodrug (Sofosbuvir), Diisobutyryl prodrug (R7128) are better than their parent nucleotides and Ribavirin in Hepatitis C Virus inhibition: A Molecular Modeling study. J Comput Theor Nanos. 2015; 12: 376-386.

34. Chevaliez S, Asselah T. Mechanisms of non-response to antiviral treatment in chronic hepatitis C. Clin Res Hepatol Gastroenterol. 2011; 35: S31-41.

35. Elshemey WM, Elfiky AA, Gawad WA. Correlation to protein conformation of Wide-angle X-ray Scatter parameters. Protein J. 2010; 29: 545-550.

36. Kelley LA, Sternberg MJ. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc. 2009; 4: 363-371.

37. Laskowski RA, Moss DS, Thornton JM. Main-chain bond lengths and bond angles in protein structures. J Mol Biol. 1993; 231: 1049-1067.

38. Lüthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature. 1992; 356: 83-85.

39. Colovos C, Yeates TO. Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci. 1993; 2: 1511-1519.

40. Pontius J, Richelle J, Wodak SJ. Deviations from standard atomic volumes as a quality measure for protein crystal structures. J Mol Biol. 1996; 264: 121-136.

41. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. Nature. 1996; 381: 272.

42. Stewart JJP. SCIGRESS, Version 2.9.0, Fujitsu Limited, United States. 2009.

Austin Publishing Group

Application of Structure and Ligand-Based Drug Design for Finding Lead Compounds from Natural Product Source: Case of Influenza Targeted

Muchtaridi*

Department of Pharmaceutical Analysis and Medicinal Chemistry, Faculty of Pharmacy, Universitas Padjadjaran, Sumedang

***Corresponding author:** Muchtaridi, Department of Pharmaceutical Analysis and Medicinal Chemistry, Faculty of Pharmacy, Universitas Padjadjaran; Jl. Bandung-Sumedang KM 21, Jatinangor, 45363, Sumedang, Email: muchtaridi@unpad.ac.id

Published Date: December 01, 2016

INTRODUCTION

Over the past 25 years, the discovery and development of novel lead drugs is conducted by rational drug design [1]. Rational drug design process is time consuming, expensive, and requires consideration of many aspects [2]. Computational techniques are applied in the rational drug design for the purpose of discovering molecules that can be very rapidly developed into an effective treatment [3].

The use of computational techniques has been shown to increase the efficiency of drug discovery and development [4, 5]. Computer-aided molecular design (**CAMD**, also called as in silico or Computer-Aided Drug Design **CADD**) is being applied to expedite and assist hit-to-lead selection, hit identification, optimize the Absorption, Distribution, Metabolism, Excretion **(ADME)** and profile toxicity [6].

As shown Figure 1. CADD can be divided into; (1) ligand based design, (2) structure based design, and (3) de novo design. Various methods of ligand-based drug design (LBDD) can be applied, if protein structures are unknown, such as the methods of Quantitative Structure Activity Relationship (QSAR) and pharmacophore modeling [4, 6]. The knowledge of ligand properties, such as pharmacological effect and bioactivity, is important in LBDD.

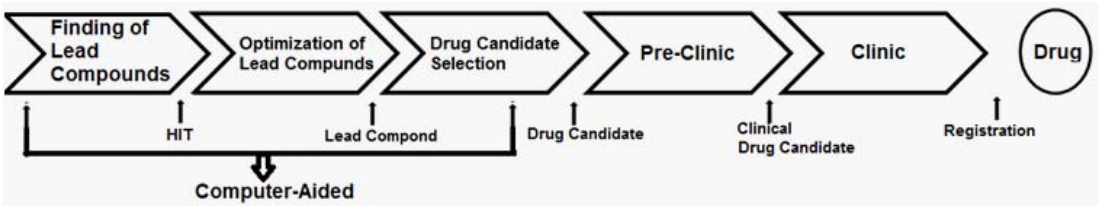


Figure 1: Stages of Drug Discovery and Development.

Computational approaches, including QSAR, pharmacophore modelling, and database mining, can be applied to a ligand set with known activity [4]. Retrieval of 3D structures from database is similar to a 2D similarity searching. However, 3D similarity searching reduce the problem of conformational flexibility [7]. The success stories of LBDD approaches in facilitating drug discovery have been reported by Kubinyi [8, 9].

Table 1: CADD Methods Resume.

	Known Ligands	Unknown Ligand
Known Target	**Structure-based drug design (SBDD)** Protein-ligand docking, molecular dynamics, homology modeling	***De novo* design**
Unknown Target	**Ligand-based drug design (LBDD)** *One or more ligands* 1. Similarity searching 2. Pharmacophore searching *Many ligands (more than 20)* 1. Quantitative Structure-Activity Relationships	**CADD of no use** Need experimental data of some sort Can apply ADMET filters

In SBDD, structural knowledge obtained from ligand–protein complexes (X-ray crystallography or NMR data) can primarily facilitate the design of focused structure-based libraries by optimizing ligand–receptor complementary interactions, in an effort to increase potency and specificity [10,11]. The applications of SBDD include the discovery of potent and selective HIV protease inhibitors [12,13], thrombin inhibitors [14,15], breast cancer [16] and neuraminidase inhibitors [17]. Recent example of this method, discovery of peramivir (BCX-1812), which was based on the structure based method utilizing the crystallography structure of a highly conserved NA active site and its substrate interactions[18,19].

On the other hand, De novo method is practically used when the ligand is unknown, with a known target. De novo ligand design will be able to test many structures in a short period of time and arrange them into a ranked list based on an accurate prediction of binding free energies since the latter reflects actual binding propensities [20-22].

One of CADD tools, which are the most popular in the last 10 years, is virtual screening [23]. Both LBDD and SBDD approaches are powerful technologies which can be supplemented by virtual screening (VS) for lead identification and optimization [4]. However, the structure-based and ligand-based techniques can be performed in the early stages of drug discovery process and help in discovery of lead compounds (as shown in Figure 2) making an initial basis for further modifications to improve pharmacokinetics, solubility, selectivity, potency or stability. Structure and Ligand-based techniques can be applied to explore the mechnaisms of ligands selectivity against their targets.

In this study, we demonstrate application of the structure-based drug design methods for investigation of neuraminidase as drug target.

NEURAMINIDASE AS A TARGET FOR DRUG DISCOVERY USING SBDD TECHNIQUE

Neuraminidase

Neuraminidase (NA) is responsible for cleaving sialic acid in terminal receptors, releasing new viruses from infected cells. NAs are found particularly in diverse virus families and bacteria, as well as in protozoa, some invertebrates and mammalian [24, 25]. They have differences in binding affinity and substrate preference; however they have conserved domains and structural similarities [24]. NA plays a vital role in influenza virus

replication, and has a conserved active site residues, thus inhibition of NA can delay the release of virus progeny from infected cells. This will reduce the virus population and will give time for the immunity of the host cell in the body to eliminate the virus [26]. NA hydrolyzes α-2,3-sialic acid from sugar (galactose), and it is also involved in the hydrolysis at α-2,6-sialic acid-galactocyl,. There are nine subtypes of neuraminidase from influenza A viruses (N1-N9) [27]. Type A influenza neuraminidases form two genetically distinct groups: group 1 consists of subtypes N1, N4, N5 and N8, while group-2 consists of N2, N3, N6, N7, and N9. Group-1 has a 150-loop cavity adjacent to the active site that serves as a gateway for the ligand to interact with NA [28]. The cavity is suitable for the active site in the development of new anti-influenza drugs [29].

The active site of NA has highly conserved active residues which are very specific to the sialic acid as the natural ligand. NA active site contains 18 residues (6 basic, 7 acidic, 3 polar, and 2 hydrophobic) [30]. Based on the chemical bonding and interaction, NA active site can be divided into sub-pockets (Figure 1.2). Subsite 1 (S1) consists of triarginyl cluster (Arg118, Arg292, and Arg371), which has a pocket of positive charge; thus it will interact with the carboxylic groups of the ligand [31]. S1 is called basic pocket which is important for designing lead compound for NA inhibitors [32]. S2 subsite is negatively charged and is composed of Glu 119 and Glu 227, and it interacts with the amine group on the acetamido of sialic acid. S3 consists of Trp178 and Ileu222 and has hydrophobic properties. The two residues are adjacent to Arg152 that binds to the water

molecules. S4 consists of Ala246 and Arg224, which are adjacent to the Ileu222 pocket and it is unoccupied by the functional groups of sialic acid [33]. The pocket accommodates a methyl group from SA and Neu5Ac2en (DANA) [32]. S4 is a new target for the development of new NA inhibitors. S5 has a unique pocket with mixed polarity environment depending on the incoming ligand. This site consists of carboxylate of Glu276 (trans-conformation) and methyl of Ala 246. During enzymatic reaction, Glu276 and Glu277 form hydrogen bonds with Tyr406 to stabilize the oxocarbonium ion with sialic acid. Glu276 interacts with O8-O9 in glycerol (SA) [34]. In addition to these amino acid residues, Asp151 has also an important role but not defined on the S1-S5 sites. This carboxylic residue does not make direct contact with DANA, but is believed to play an important role in catalysis by polarizing the bond α-2,3-sialic acid-glycosidic. Asp115 with Glu119 and Glu227 are also involved in sialic acid hydrolysis through the involvement of water molecules [34].

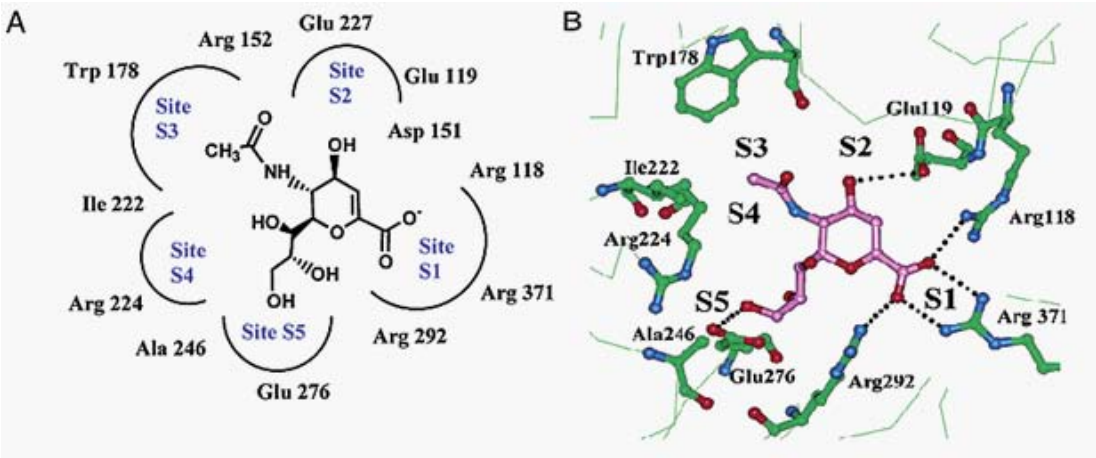


Figure 2: The interaction of DANA inside a neuraminidase active site in (a) 2D and (b) 3D representative (taken from PDB ID: 1NNB) (from Stoll et al., 2003).

Structure-Based Drug Design (SBDD)

As mentioned above, SBDD is a powerful technique in the process of discovery and development of drug. SBDD approach is useful for evaluating the complementarities and predicting the possibility of binding modes and affinities between ligands and their macromolecular receptors [4]. The availability of X-ray crystal structures of the influenza virus NA with and without a ligand such as α-Neu5Ac and Neu5Ac2en [35-37] provides the key in designing NA inhibitors. Edmond and co-workers [38] together with Meindle's group [39] started a random screening method in drug discovery. This method is based on guided activity that focused on trial and error, but the method doesn't work as the compounds easily produce drug resistance. Goodford [40] calculated the interaction energy between ligand and the target using computational methods and found that target – ligand interaction can be predicted by software programs such as GRID [40]. The

GRID program has been used by on Itztein et al. [31] to design NA inhibitors in SBDD approach. Potent interaction between sialic acid **(SA)** and NA in the complex crystal structure is the basis for design of NA inhibitors [41].

It is possible to design highly potent NA inhibitors with SBDD, as has been shown by the discovery of zanamivir (ZANA) [34]. Using GRID [40,42] the active site of NA is explored for the ability to accommodate a variety of groups such as carboxylates, amine, methyl and phosphate functional groups to get a potent and effective inhibiting NA [31]. Several compounds have been successfully modified and optimized based on charge and shape of the character of active sites through SBDD methods, such as ZANA [31] and OTV [43]. Based on the results of computational chemistry, von Itztein et al. [31] replaced the hydroxyl group at C-4 from the Neu5Ac2en with amine base groups into 4-amino-4-deoxy-Neu5Ac2en (Figure 3(a)) and further replaced with a guanidino group (ZANA) (Figure 3(b)). Based on these data, C-4 group on the guanidino of ZANA successfully interact with carboxylic groups on the site active residues (Glu119 and Glu227) which leads to better inhibition of NA of Neu5Ac2en. The importance of NA in the history of the pathogenesis of influenza virus infection and the properties of the active side residue which is highly conserved lead to a concrete reason to design of small molecule, which is selective and effective towards NA.

The glycerol moieties of ZANA interact with the active site of NA in the way similar to DANA. The in silico results show that the replacement of glycerol with a more hydrophobic group makes the ligand more stable in solid form (oral administration), whereas ZANA is stable only in the form of solution (intravenous). In addition, QSAR studies have shown that the replacement of glycerol with considerations chain length, branches, and stereochemistry of alkyl groups also improve the inhibition of NA. This is the basis in designing of OTV (GS4071) (Figure 4(a)). Kim et al. performed optimization by the replacement of glycerol with 3-pentyl ether but maintaining acetamido and amino groups in the GS 4071[43]. In this discovery, GS4104 (Figure 4(b)) has been developed for the purpose of drug formulations, which is the ester derivative of OTV [44-46].

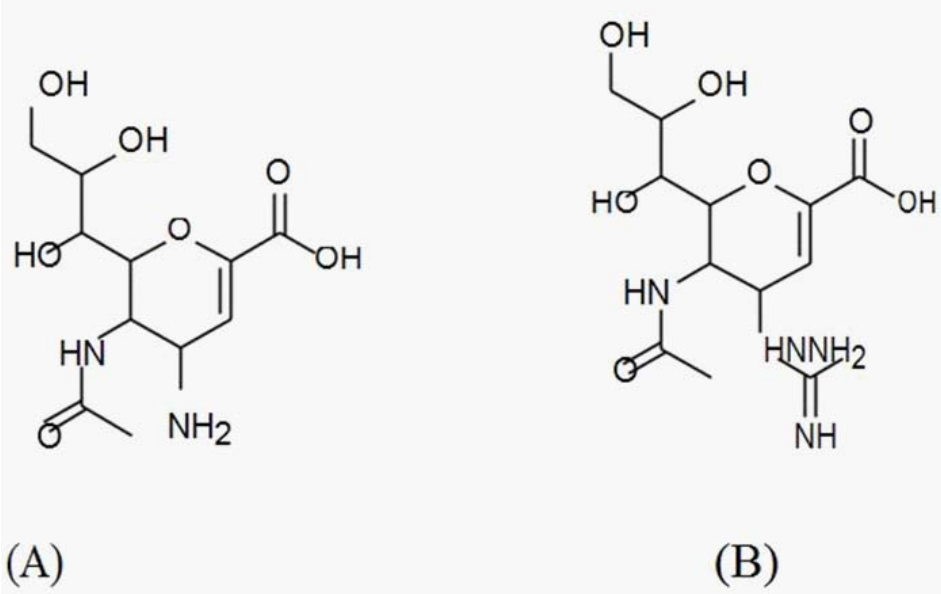


Figure 3: The structures of (a) 4-amino-4-deoxy-Neu5Ac2en and (b) zanamivir.

Based on the success in discovery of ZANA [31] and OTV [44], SBDD has played an important role in the discovery of other NA inhibitors. However, the constant threat of pandemic avian influenza [47] and the emergence of strains resistant to OTV (Tamiflu) make the development of new effective NA inhibitors important.

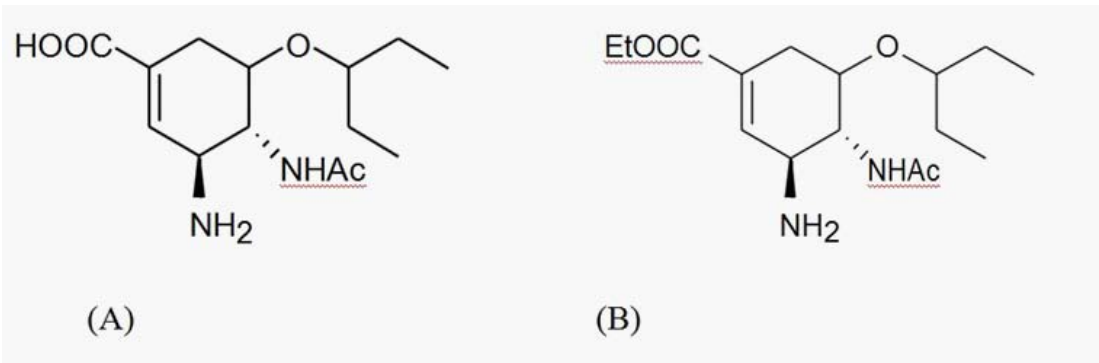


Figure 4: (A) oseltamivir carboxylate (GS 4071) (B) oseltamivir (GS 4104).

Molecular docking is one of the useful SBDD methods. The availability of three-dimensional structure of NA has played an important role in SBDD methods to discover new inhibitors of neuraminidase. Varghese et al. (1983) determined the first three-dimensional structure of NA (N2 subtype structure) using X-ray crystallography with 2.9 Å resolution [36].They also solved the first complex structure of NA-sialic acid in 1992 (PDB id; 2bat) [37]. 2HU4 (neuraminidase from H5N1) from Protein Data Bank (PDB) was used in molecular docking simulation with Autodock 3.5. In this study were successfully screened 3000 natural product compounds. From

five plants, 12 compounds were isolated which show neuraminidase inhibiting activity, including two compounds with IC 50 values less than 92 μM [48]. Whereas, 3B7E (neuraminidase from H1N1) and 3NSS (neuraminidase from H5N1 mutant) were employed in molecular docking simulation (Autodock 4.2) to screen out 113 natural product compounds and the some natural compounds screened were assayed by MUNANA assay to prove the in silico concept as mention above. Catechin, epicatechin, galocatechin and gallic acid were tested against N1 of neuraminidase (C. perpringens) as shown in Figure 6; the IC50 value of catechin was 93.92 μM. Epicatechin (18), galocatechin (19), and gallic acid (20) had 137.1 μM, 165.1 μM, and 205.7 μM, respectively [49].

LIGAND-BASED DRUG DESIGN: DISCOVERY OF NEURAMINIDASE INHIBITORS

Recently, sialic acid and Neu5Ac2en derivatives have been synthesized and evaluated for their influenza virus sialidase inhibitory activity. At least 268 derivatives of Neu5Ac2en have been synthesized up to date (www.bindingDB.org) [50, 51]. Molecular alignments of Neu5Ac2en derivatives with known activity (IC50 or Ki) can be employed as basic data in Ligand-Based Drug Design **(LBDD)**. The crucial component in the ligand-based approaches is the need to superimpose, or align, a series of active ligands, ideally to mimic the way in which they would be overlaid in the binding site. Such superimpositions form the basis for techniques such as 3D database searching, 3D quantitative structure–activity relationship (QSAR), and receptor modelling [52].

Every active ligand will have the key pharmacophores that influence the biological activity. For example, Neu5Ac2en ligand has carboxylic groups acting as the negative ionizable feature and this feature contributesto charge-charge binding interaction of the ligand and the Arginine triad (371, 292, and 118) [34].

Pharmacophore features are generated from molecular alignments of ligand analogues based on steric, electronic, function-determining points for an optimal interaction with the relevant pharmacological target. There are many ways to generate pharmacophores. In one approach, 3D and predictive pharmacophores can be created automatically from the most active ligand set as the basic information [53]. For example, (refer to chemical structure in Figure 1.7), sialic acid (Figure 1.7(b)) is the natural ligand that is the least active towards neuraminidase (NA) (IC50 1000 mM). The compound has the least capability to inhibit NA, but it has good selectivity, whereas DANA which is modified from sialic acid by removing hydroxyl group at C2 (Figure 5b), was more active to inhibit NA (IC50 1 mM). Hydroxyl group at C4 is replaced with guanidine in ZANA (Figure 1.7(b)) making the compound more active than DANA against NA. DANA and ZANA, produced three feature; negative ionisable (in carboxylic acid), hydrogen bond donor (in OH or NH2 at C4 and C6), and hydrogen bond acceptor (in acetonimide at C5). Hiphop in CATALYST generates common features and produce three features as discussed above and determine inter features distance. In recent time, hydrophobic groups are attached to Neu5Ac2en at C6 to replace glycerol in DANA or ZANA that is exemplified by OTV in Figure 5 (c).

Several successful studies in discovery of new inhibitors of neuraminidase using LBDD have been reported where medicinal chemists utilised the pharmacophore models and QSAR [54-57]. Zhang et al.[57] generated the best hypogen model of pharmacophore from 22 NA inhibitors structures, consisting of 2 Hydrogen Bond Donors (HBD), hydrophobic (HY), negative ionizable, and positive ionizable features. In addition, five best pharmacophore models that emerged in the optimal QSAR equations show the existence of several different ligand-NA binding modes in the NA binding pocket [56].

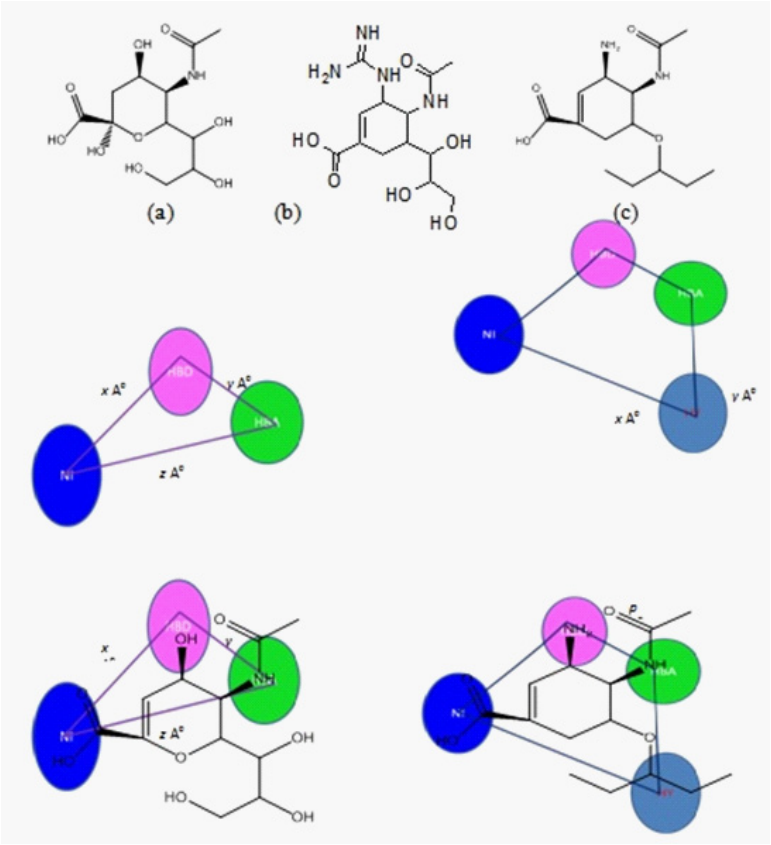
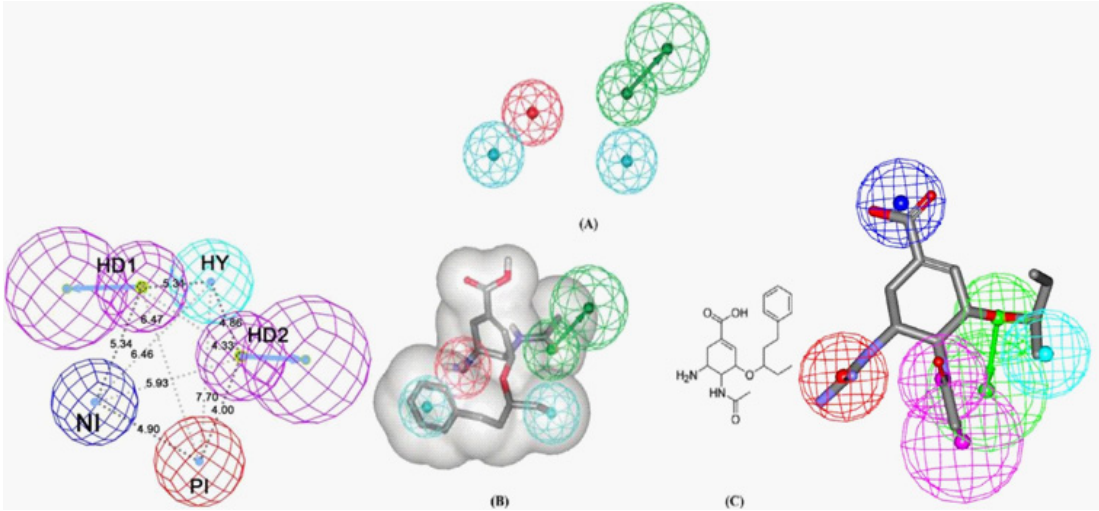


Figure 5: Common pharmacophore features of sialic acid derivatives (a) sialic acid structure. (b) zanamivir structure (c) oseltamivir.

Nevertheless, the resistance of influenza virus to existing NA inhibitors [58] suggested medicinal chemists an idea to attach some functional groups in sialic acid derivatives, that required development of the models with the last pharmacophore features. For example, Zhang et al.[57] added one hydrophobic feature and one hydrogen bond acceptor in their models, while Hammad et al. [56] attached two hydrophobic feature and one hydrogen bond donor in one of their five models. Hammad et al. (2009) used 181 NA inhibitors in their study.



(a) (b) (c)

Figure 6: (a) Hypogen pharmacophore model consisting of 2 HBD, 1 HY, PI, and NI, that was generated from 22 NA inhibitor structures [57]. (b) One of five hypogen pharmacophore models was generated from 181 selected NA inhibitor structures. The model included 2 HY, one HBA, and PI. (c) One of four hypogen best pharmacophore models was generated from 232 selected NA inhibitor structures. Pharmacophore features are color coded: magenta - hydrogen bond donor (HBD), green - hydrogen bond acceptor (HBA), blue – hydrophobic feature (Hy), blue – Negative Ionizable (NI), red - Positive ionizable (PI).

Muchtaridi et al. (2014) have created 3D-pharmacophore models with features similar to those described above. The best models were applied to successfully screen natural product compounds from the NADI database [59]. Hammad et al. [60] resumed that all the pharmacophore models discovered have HBD, HBA, NI, and PI. Pharmacophore model A-5-5 includes only 3 HBA and NI while model A-8-1 consists of more complete features (HBD, HBA, HY, NI, and PI). Model B-3-2 contains 2HBD, 2HY, and NI while model C-1-2 consist 2HBD and 2 HY.

3D-pharmacophore model prepared by Zhang et al. [57] includes 2 HBD, PI, NI, and HY. In this model, OTV was used as the most active compound which is similar to T2S202 model. However, the distance of inter-features in Zhang's model is around 4.0-6.47 Å A, while

T2S202 model has distance about 3.487 - 7.826 Å, thus it might indicate that there are conformational differences of OTV between both models. For example, NI and PI of Zhang's and T2S202 models mapped into same chemical groups of OTV (10) while the distance of PI and NI in the both models were different (Zhang's:4.90 Å; T2S202:6.54 Å).

VIRTUAL SCREENING

Virtual Screening **(VS)** is a method which attempts to rank candidate molecules in descending order by likelihood of biological activity, hence reducing the number of compounds for experimental evaluations [61]. Suggested, that this method can improve the reliability of the measurements. VS are based on computer filtering tools to effectively eliminate inactive small molecules and find lead compounds. The VS is used for browsing databases and molecules fitting either an established pharmacophore model or a three dimensional (3D) structure of a macromolecular target [62]. Both LBDD (pharmacophore) and SBDD (docking) approaches can be utilized in virtual screening for lead identification and optimization [23, 63, 64]. Success stories of these methods have been reviewed by Villoutreix et al. (2000), Kubinyi et al (2006) and Gosh et al. (2006). In neuraminidase targeted case, Ikram et al., (2015) employed docking for virtual screening to find neuraminidase inhibitors from natural products, while Muchtaridi et al. (2014) used combination of pharmacophore-docking approaches to find the lead compounds. Hammad et al. used their pharmacophore models (Figure 3b) to find compounds from NCI potentially active against NA. They discovered the top hits based on pharmacophore ranking that showed an in vitro IC50 value of 1.8 μM [49].

CONCLUSION AND FUTURE PERSPECTIVES

Combination of the in silico methods of and bioassay-guided isolation were applied to find potent inhibitors of the neuraminidase. Among in silico methods, pharmacophore modeling and molecular docking were found particularly useful.

AknowledgmentI gratefully acknowledge the Rector of Universitas Padjadjaran and Minister of Research-Tchnology and Higher Education, Indonesia, for funding this project through PUPT 2015.

I thank to Prof. Habibah A. Wahab, University Sains Malaysia, for the part funding and the computing facilities used in this project.

References

1. Mavromoustakos T, Durdagi S, Koukoulitsa C, Simcic M, Papadopoulos MG. Strategies in the rational drug design. Curr Med Chem. 2011; 18: 2517-2530.
2. Mandal S, Moudgil M, Mandal SK. Rational drug design. Eur J Pharmacol. 2009; 625: 90-100.
3. Srinivasa Rao V, Srinivas K. "Modern drug discovery process: An in silico approach". J Bioinfo Seq Anal. 2011; 3: 89-94.
4. Zhang S1. Computer-aided drug discovery and development. Methods Mol Biol. 2011; 716: 23-38.
5. Marshall GR. Computer-aided drug design. Annu Rev Pharmacol Toxicol. 1987; 27: 193-213.
6. Kapetanovic IM1. Computer-aided Drug Discovery and Development **(CADDD)**: in silico-chemico-biological approach. Chem Biol Interact. 2008; 171: 165-176.
7. Terfloth L, Gasteiger J. Electronic Screening: Lead Finding From Database Mining. In: The Practice of Medicinal Chemistry. Edited by Wermuth CG. London: Elsevier; 2003: 131-415.
8. Kubinyi H. Success Stories of Computer-Aided Design. In: Computer applications in pharmaceutical research and development. Edited by Ekins S. Hoboken NJ. Wiley-Interscience; 2006: 377-424.

9. Kubinyi H. QSAR: Hansch analysis and related approaches. Weinheim; New York: VCH. 1993.

10. Orry AJ, Abagyan RA, Cavasotto CN. Structure-based development of target-specific compound libraries. Drug Discov Today. 2006; 11: 261-266.

11. Hubbard RE. Structure-based drug discovery and protein targets in the CNS. Neuropharmacology. 2011; 60: 7-23.

12. Hubbard RE. Structure-based drug discovery and protein targets in the CNS. Neuropharmacology. 2011; 60: 7-23.

13. Rubin B, Laffan RJ, Kotler DG, O'Keefe EH, Demaio DA. SQ 14,225 (D-3-mercapto-2-methylpropanoyl-l-proline), a novel orally active inhibitor of angiotensin I-converting enzyme. J Pharmacol Exp Ther. 1978; 204: 271-280.

14. Wlodawer A, Vondrasek J. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. Annu Rev Biophys Biomol Struct. 1998; 27: 249-284.

15. Wagner J, Kallen J, Ehrhardt C, Evenou JP, Wagner D. Rational design, synthesis, and X-ray structure of selective noncovalent thrombin inhibitors. J Med Chem. 1998; 41: 3664-3674.

16. Wagner J, Kallen J, Ehrhardt C, Evenou JP, Wagner D. Rational design, synthesis, and X-ray structure of selective noncovalent thrombin inhibitors. J Med Chem. 1998; 41: 3664-3674.

17. Böhm HJ, Banner DW, Weber L. Combinatorial docking and combinatorial chemistry: design of potent non-peptide thrombin inhibitors. J Comput Aided Mol Des. 1999; 13: 51-56.

18. Muchtaridi M, Yusuf M, Diantini A, Choi SB, Al-Najjar BO. Potential activity of fevicordin-A from Phaleria macrocarpa (Scheff) Boerl. seeds as estrogen receptor antagonist based on cytotoxicity and molecular modelling studies. Int J Mol Sci. 2014; 15: 7225-7249.

19. Itzstein vM, Wu WY, Kok GB, Pegg MS, Dyason JC, et al: Rational design of potent sialidase-based inhibitors of influenza virus replication. Nature. 1993; 363: 418-423.

20. Young D, Fowler C, Bush K. RWJ-270201 (BCX-1812): a novel neuraminidase inhibitor for influenza. Philos Trans R Soc Lond B Biol Sci. 2001; 356: 1905-1913.

21. Babu YS, Chand P, Bantia S, Kotian P, Dehghani A, El-Kattan Y, Lin TH, Hutchison

22. TL, Elliott AJ, Parker CD, et al: BCX-1812 (RWJ-270201): discovery of a novel, highly potent, orally active, and selective influenza neuraminidase inhibitor through structure-based drug design. J Med Chem. 2000, 43: 3482-3486.

23. DeWitte RS, Shakhnovich EI: SMoG. de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. J Am Chem Soc. 1996, 118:11733-11744.

24. Böhm HJ. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. J Comput Aided Mol Des. 1998; 12: 309-323.

25. Hartenfeller M, Schneider G. De novo drug design. Methods Mol Biol. 2011; 672: 299-323.

26. Lengauer T, Lemmen C, Rarey M, Zimmermann M. Novel technologies for virtual screening. Drug Discov Today. 2004; 9: 27-34.

27. Schwerdtfeger SM, Melzig MF. Sialidases in biological systems. Pharmazie. 2010; 65: 551-561.

28. Sander-Wewer M, Schauer R, Corfield AP. Substrate specificity of viral, bacterial and mammalian sialidases with regard to different N,O-acetylated sialic acids and GM1. Adv Exp Med Biol. 1982; 152: 215-222.

29. Sander-Wewer M, Schauer R, Corfield AP. Substrate specificity of viral, bacterial and mammalian sialidases with regard to different N,O-acetylated sialic acids and GM1. Adv Exp Med Biol. 1982; 152: 215-222.

30. Garman E, Laver G. Controlling influenza by inhibiting the virus's neuraminidase. Curr Drug Targets. 2004; 5: 119-136.

31. Liu C, Eichelberger MC, Compans RW, Air GM. Influenza type A virus neuraminidase does not play a role in viral entry, replication, assembly, or budding. J Virol. 1995; 69: 1099-1106.

32. Russell RJ, Haire LF, Stevens DJ, Collins PJ, Lin YP. The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. Nature. 2006; 443: 45-49.

33. Rudrawar S, Dyason JC, Rameix-Welti MA, Rose FJ, Kerry PS. Novel sialic acid derivatives lock open the 150-loop of an influenza A virus group-1 sialidase. Nat Commun. 2010; 1: 113.

34. Colman PM. Influenza virus neuraminidase: structure, antibodies, and inhibitors. Protein Sci. 1994; 3: 1687-1696.

35. von Itzstein M, Dyason JC, Oliver SW, White HF, Wu WY. A study of the active site of influenza virus sialidase: an approach to the rational design of novel anti-influenza drugs. J Med Chem. 1996; 39: 388-391.

36. Taylor GL. Influenza Virus Neuraminidase Inhibitors. In: Handbook of Cell Singaling,. Edited by R.A. B, Dennis EA: Academic Press Inc. 2009: 103-110.

37. Stoll V, Stewart KD, Maring CJ, Muchmore S, Giranda V. Influenza neuraminidase inhibitors: structure-based design of a novel inhibitor series. Biochemistry. 2003; 42: 718-727.

38. Taylor NR, von Itzstein M. Molecular modeling studies on ligand binding to sialidase from influenza virus and the mechanism of catalysis. J Med Chem. 1994; 37: 616-624.

39. Colman PM, Varghese JN, Laver WG. Structure of the catalytic and antigenic sites in influenza virus neuraminidase. Nature. 1983; 303: 41-44.

40. Varghese JN, Laver WG, Colman PM. Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 A resolution. Nature. 1983; 303: 35-40.

41. Varghese JN, McKimm-Breschkin JL, Caldwell JB, Kortt AA, Colman PM. The structure of the complex between influenza virus neuraminidase and sialic acid, the viral receptor. Proteins. 1992; 14: 327-332.

42. Edmond JD, Johnston RG, Kidd D, Rylance HJ, Sommerville RG. The inhibition of neuraminidase and antiviral action. Br J Pharmacol Chemother. 1966; 27: 415-426.

43. Meindl P, Tuppy H. [2-Deoxy-2,3-dehydrosialic acids. II. Competitive inhibition of Vibrio cholerae neuraminidase by 2-deoxy-2,3-dehydro-N-acylneuraminic acids]. Hoppe Seylers Z Physiol Chem. 1969; 350: 1088-1092.

44. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem. 1985; 28: 849-857.

45. Wade RC. 'Flu' and structure-based drug design. Structure. 1997; 5: 1139-1145.

46. Goodford PJ. Drug design by the method of receptor fit. J Med Chem. 1984; 27: 558-564.

47. Kim CU, Lew W, Williams MA, Liu H, Zhang L. Influenza neuraminidase inhibitors possessing a novel hydrophobic interaction in the enzyme active site: design, synthesis, and structural analysis of carbocyclic sialic acid analogues with potent anti-influenza activity. J Am Chem Soc. 1997; 119: 681-690.

48. Lew W, Chen X, Kim CU. Discovery and development of GS 4104 (oseltamivir): an orally active influenza neuraminidase inhibitor. Curr Med Chem. 2000; 7: 663-672.

49. Lew W, Escarpe PA, Mendel DB, Sweeny DJ, Kim CU. Stereospecific synthesis of a GS 4104 metabolite: determination of absolute stereochemistry and influenza neuraminidase inhibitory activity. Bioorg Med Chem Lett. 1999; 9: 2811-2814.

50. Li W, Escarpe PA, Eisenberg EJ, Cundy KC, Sweet C. Identification of GS 4104 as an orally bioavailable prodrug of the influenza virus neuraminidase inhibitor GS 4071. Antimicrob Agents Chemother. 1998; 42: 647-653.

51. W, Williams M, Zhang L, et al: Identification of GS 4104 as an orally bioavailable prodrug of the influenza virus neuraminidase inhibitor GS 4071. Antimicrob Agents Chemother. 1998; 42: 647-653.

52. Monto AS. The threat of an avian influenza pandemic. N Engl J Med. 2005; 352: 323-325.

53. Ikram NK, Durrant JD, Muchtaridi M, Zalaludin AS, Purwitasari N. A virtual screening approach for identifying plants with anti H5N1 neuraminidase activity. J Chem Inf Model. 2015; 55: 308-316.

54. Muchtaridi M, Aliyudin A, Holik HA. Potential Activity of Some Natural products Compounds as Neuraminidase Inhibitors Based on Molecular Docking Simulation and *In Vitro* Test. J App Pharm Sci. 2015; 5: 065-073.

55. Chen X, Lin Y, Liu M, Gilson MK. The Binding Database: data management and interface design. Bioinformatics. 2002; 18: 130-139.

56. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res. 2007; 35: D198-201.

57. Quintus F, Sperandio O, Grynberg J, Petitjean M, Tuffery P. Ligand scaffold hopping combining 3D maximal substructure search and molecular similarity. BMC Bioinformatics. 2009; 10: 245.

58. Kurogi Y, Güner OF. Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. Curr Med Chem. 2001; 8: 1035-1055.

59. Gong JZ, Liu Y, Xu WF. Pharmacophore model of influenza neuraminidase inhibitors--a systematic review. Pharmazie. 2009; 64: 627-632.

60. Chen CY, Chang YH, Bau DT, Huang HJ, Tsai FJ. Ligand-based dual target drug design for H1N1: swine flu-a preliminary first study. J Biomol Struct Dyn. 2009; 27: 171-178.

61. Abu Hammad AM, Taha MO. Pharmacophore modeling, quantitative structure-activity relationship analysis, and shape-complemented in silico screening allow access to novel influenza neuraminidase inhibitors. J Chem Inf Model. 2009; 49: 978-996.

62. Zhang J, Yu K, Zhu W, Jiang H. Neuraminidase pharmacophore model derived from diverse classes of inhibitors. Bioorg Med Chem Lett. 2006; 16: 3009-3014.

63. Puzelli S, Facchini M, Di Martino A, Fabiani C, Lackenby A. Evaluation of the antiviral drug susceptibility of influenza viruses in Italy from 2004/05 to 2009/10 epidemics and from the recent 2009 pandemic. Antiviral Res. 2011; 90: 205-212.

64. Muchtaridi M, Sy Bing C, Abdurrahim AS, Wahab HA. Evidence of Combining Pharmacophore Modeling-Docking Simulation for Screening on Neuraminidase Inhibitors Activity of Natural Product Compounds. Asian J Chem. 2014; 26: S59-S63.

65. Abu Hammad AM, Taha MO. Pharmacophore modeling, quantitative structure-activity relationship analysis, and shape-complemented in silico screening allow access to novel influenza neuraminidase inhibitors. J Chem Inf Model. 2009; 49: 978-996.

66. Shoichet BK. Virtual screening of chemical libraries. Nature. 2004; 432: 862-865.

67. Rollinger JM, Stuppner H, Langer T. Virtual screening for the discovery of bioactive natural products. Prog Drug Res. 2008; 65: 211, 213-249.

68. Xu H, Agrafiotis DK. Retrospect and Prospect of Virtual Screening in Drug Discovery. Curr Top in Med Chem. 2002; 2: 13005-11320.

69. Vyas V, Jain A, Jain A, Gupta A. Virtual Screening: A Fast Tool for Drug Design. Sci Pharm. 2008; 76: 333-360.

Austin Publishing Group

Molecular Dynamics of E. coli Undecaprenyl Diphosphate Synthase: Asymmetry in a Homodimer

Newhouse E I[1]*, Alam M[2] and Mukhametov A[3]

[1]Advanced Studies in Genomics, Proteomics and Bioinformatics, University of Hawaii at Manoa, Honolulu, USA

[2]Department of Microbiology, University of Hawaii at Manoa, Honolulu, USA

[3]Department of Computer-Aided Molecular Design, Institute of Physiologically Active Compounds of the Russian Academy of Sciences, Russia

***Corresponding author:** Newhouse EI, Advanced Studies in Genomics, Proteomics and Bioinformatics, University of Hawaii at Manoa, 2565 McCarthy Mall, Keller 319, Honolulu, HI 96822, USA, Email: einew@hotmail.com

Published Date: December 01, 2016

ABSTRACT

Undecaprenyl Diphosphate Synthase **(UPPS)** is of interest as a target for antibiotic development. It catalyses the synthesis of undecaprenyl diphosphate, a C_{55} isoprene lipid carrier required during synthesis of peptidoglycan, a cell wall component, in many bacteria. Its substrates are isoprenyl diphosphate and farnesyl diphosphate. In this work, we performed molecular dynamics calculations on the apo-protein, substrate-bound protein, and product-bound protein from *E. coli* to understand better the mechanism of substrate entry and product release from the reaction cavity, which has been found, from crystal structures, to open via separation between two of the α-helices, α2 and α3. Our work confirms significant volume fluctuations in the reaction cavity over simulation time, accompanied by significant volume differences between the dimer subunits at a given time step. These appear not to be from overall protein backbone conformational changes,

but primarily from reorientation of the α3 helix and from sidechain reorientations in residues lining the cavity surface. We have also observed that there are salt bridges between the three α-helices (α2, α3 and α4) comprising the outer wall of the reaction cavity; they limit the ability of these helices to separate as required for substrate entry and product release. These salt bridges are found in all the UPPS crystal structures obtained to date, though not at conserved positions, and may control protein activity.

Keywords: *Ditrans*; *Polycis*-undecaprenyl-diphosphate synthase [(2*E*,6*E*)-farnesyl-diphosphate specific]; Undecaprenyl synthase; Molecular dynamics; Homodimer; Asymmetry

Abbreviations: FPP: Farnesyl Pyrophosphate; **FPS:** Farnesyl Thiopyrophosphate; **IPP:** Isopentylpyrophosphate; **NPT** Thermodynamical Condition of Constant Number, Pressure and Temperature; **NVT:** Thermodynamical Condition of Constant Number, Volume, and Temperature; **PBC:** Periodic Boundary Conditions; **PCA** Principal Component Analysis; **PDB:** Protein Data Bank; **UPPS:** *Ditrans*, *Polycis*-Undecaprenyl-Diphosphate Synthase

INTRODUCTION

Undecaprenyl diphosphate synthase superfamily members catalyze the synthesis of straight-chain isoprenes. We consider here *ditrans, polycis*-undecaprenyl-diphosphate synthase [(2*E*,6*E*)-farnesyl-diphosphate specific] (EC 2.5.1.31) (UPPS) from *E. coli* which catalyses the synthesis of Undecaprenyl Pyrophosphate **(UPP)**, a C_{55} compound used to synthesize a lipid carrier by a wide range of bacteria during peptidoglycan biosynthesis[1-3]. As the protein is an integral part of cell wall synthesis, it is thus of interest as a target for antibiotic development [4]. UPPS adds eight Isoprenyl Units **(IPP)** to Farnesyl Pyrophosphate **(FPP)**. The substrate pyrophosphate head groups are bound by a "P-loop" consisting of residues G-N-G/R-R [5], which are numbered 27 to 30 in *Micrococcus luteus* UPPS (The third residue in the P-loop is not strictly conserved and is either G or R, hence 'G/R'). Several crystal structures for UPPS with or without substrate analogs, as well as with bound inhibitors, are in the Protein Data Bank [4-13]. Isoprenyl addition in UPPS has been experimentally demonstrated to occur via an associative SN_2 mechanism [14]. Experiments have also determined that in *E. coli* UPPS, FPP binds first [15], followed by IPP. UPPS is a 'metal activated' enzyme; Mg ion is required for activity [16], though not incorporated into the crystal structure. UPPS is also activated by lipids or anionic detergents [17]. In each reported structure with bound molecules, the compounds are so deep within the protein that it is obvious significant conformational rearrangement must occur in order for binding or release to be possible.

METHODS

Substrate-bound model: The series PDB 1X06, 1X07, 1X08, and 1X09 was used to prepare a substrate-bound model for molecular dynamics. As coordinates for only one protein monomer unit were deposited, the second unit was generated from the crystal symmetry data with Chimera [18]. 1X08 and 1X09 were D26A mutants, which were back-mutated using Maestro (Schrodinger Inc.,

Portland OR USA) software. The fragment of FPS (farnesyl thiopyrophosphate, an analog of FPP with sulfur instead of oxygen linking the hydrocarbon to phosphorus) far from the pyrophosphate binding site was deleted. 1X08 had a complete FPS chain, but the IPP was not resolved. Thus the IPP coordinates from 1X09 were used (1X09 lacked the FPS hydrocarbon chain). The partial substrate coordinates in 1X08 and 1X09 were mutually aligned using the Multiseq plugin in VMD [19] to align the surrounding protein, and combined into a unified set for future use. Protonation states were determined using PDB2PQR [20] in the presence of substrates. Partial charges on the substrate molecules for use with PDB2PQR were calculated using the Swissparam server [21]. Hydrogens were added, and the model was solvated and ionized (neutralized and NaCl added to 0.15M) using the Maestro GUI. The model system was equilibrated locally using the default settings in Desmond [22]: minimization with restrained solute, full minimization, Berendsen NVT molecular dynamics at 10K, Berendsen NPT dynamics at 10K, Berendsen NPT simulation at 300K with restrained solute, Berendsen NPT simulation at 300K with no restraints, then finally, 100psec of molecular dynamics. Simulations were carried out to ~80nsec using Desmond 2.0.4 using the final equilibrated structure. The use of periodic boundary conditions to compensate for the small size (relative to real life) of simulation systems resulted in some frames having one of the monomer units in a unit cell adjacent to its partner instead of in the same unit cell. There are software tools to correct the situation; the process is called "unwrapping". The trajectory was unwrapped using PBC Tools [23]. The PBC Tools center selection was varied until a wrapping was achieved in which only ~6% of the frames contained wrapping errors. These errors occurred between 48.4 and 60.92 n sec, with four additional bad frames from 63.59 to 63.63 nsec inclusive. These frames were excluded from the analyses of cavity volumes.

Apo-protein model: PDB 3QAS, apo-UPPS, had the 72-83 loop fully resolved in only one of the protein dimer units. There were two choices for constructing a model system containing both loops: perform a symmetry operation on a copy of the fully-resolved unit to superimpose it on the partially-resolved unit, or to perform the same symmetry operation only on the missing residues of the flexible loop and edit them into the gap. The first choice resulted in a system with a large positive free energy of dimerization as computed from an 8-nsec NAMD [24] simulation using the MM/PBSA method implemented in AMBER11 [25]. (See next paragraph). Consequently, the second method was adopted, and the missing residues edited into the chain with the gap. This was done by selecting residues 71-90 (residues 72-89 were not resolved) and using Maestro's superimpose atoms feature to superimpose the Cα of residues 71 and 90 of the loop to be inserted on their counterparts at the limits of the protein gap, and deleting one copy of the duplicated residues 71 and 90. The energies of the 'splices' were minimized using the loop minimization feature of Maestro. This second model was processed as described above, and both Desmond and NAMD simulations rerun. MM/GBSA calculated a favorable dimerization free energy for the NAMD trajectory. Desmond 3.0.1 was used for 80nsec of simulation. This version has improved routines to prevent protein wrapping to other periodic cells, and PBC Tools was not required.

Checking the protein-protein dimerization free energy: In order to check the order of magnitude of ΔG_{dimer} expected for UPPS, we computed an 8-nsec test trajectory from PDB 2VG4, a system for which coordinates for both dimer units with all residues resolved are available. It gave -127 ± 15 Kcal/mol for ΔG_{dimer}. (250 points equally spaced between 4 and 8nsec were used for the enthalpy term, and 20 points equally spaced along the same interval were used for the entropy term). The same calculation was performed on a 2VG4-based dimer obtained by rotating one of the monomer units, resulting in -104 ± 12 Kcal/mol. This difference indicates that local symmetry-breaking in dimer interfaces may stabilize dimer interactions. ΔG_{dimer} for 3QAS dimer with the 72-89 gap filled in as described in the previous paragraph gave ΔG_{dimer}= -99 ± 9 Kcal/mol.

Product-bound UPPS model: Product-bound UPPS was also simulated using the coordinates of PDB 1X06. The build function of the Schrodinger Maestro GUI was used to build the C_{55} product. The molecular mechanics minimization routine in Maestro was used to minimize the resulting structure. As reported by Chen et al. [26], the minimized structure is very compact compared to the extended form in which one would draw it. The superimpose atoms function of Maestro was used to superimpose the pyrophosphate group of the product structure over that in FPS. The resulting steric clashes were displayed ('ugly' contacts in Maestro terminology) using the measure contacts feature of Maestro. Dihedral angles were manually adjusted to reduce the number of these 'ugly' contacts. As dihedrals were adjusted, an effort was made to direct the long axis of the three trans-double bonds toward L137, as L137 has been determined from mutation studies to control product chain length [7]. Once the 'ugly' contacts had been reduced to a small number, the resulting structure was minimized with protein backbone constraints within Maestro. Only one 'ugly' contact of the original 80 remained. This structure was used for molecular dynamics simulation, and equilibrated as previously. Production MD was done with Desmond 3.0.1.

Domain analysis was performed with the DynDom server [27]. Cavity volumes and substrate volumes were computed with 3V [28]. 3V computations were performed with the surface-limiting probe set to 6 Å and the concave cavity surface probe set to 2 Å. The grid spacing was 0.5 Å, and the minimum volume to be selected as a candidate cavity was set to 500Å^3, based on the substrate volume of ~750Å^3. These settings calculated two cavities in 98% of the trajectory frames of the substrate-bound protein simulation. 3V sometimes identified the cavity on chain A first, sometimes the one on chain B was identified first. It was not related to the relative cavity sizes. Consequently, it was not possible to plot the time course of cavity sizes for a given chain. Instead, we plotted (data not included) the distribution of cavity sizes on both chains, and the differences in volume between the two cavities in a given trajectory frame. Cavity volumes were visualized using UCSF Chimera [18]. When a rendering of the opening of a cavity was desirable, the Castp server [29] was used with the default probe radius of 1.4 Å^3. VMD [19] was used for other analyses.

Principal component analysis **(PCA)** was done using the AMBER ptraj utility, and plotted (data not included) with gnuplot [30]. Asymmetry scores were calculated as in [31]. Sequence alignments were done with the ClustalW server [32].

RESULTS AND DISCUSSION

UPPS Mechanism from Crystal Structures

UPPS is a homodimer with seven α helices and six β strands in each unit (Figure.1). Domain analysis [27] identifies regions of the protein that move as a unit, as well as the sections that must move the most in order to permit these motions to occur. It requires input of two structures. Since crystal structures of substrate-bound (FPS/IPP) protein (1X08) and apo-protein (3QAS) were available, domain analysis was possible. The result indicated that residues 17-68 and 97-237 superimpose (fixed domain). Residues 72- 92 were most different in the two structures (mobile domain). These residues form the loop connecting helices α2 and α3 and part of the α3 helix. The bending regions are residues 68-71 and 93-96. The difference in position of residues 72-92 between apo-protein and substrate-bound protein were determined to consist of an 84° rotation and a translation of 2.9 Å. These relationships have been previously noted and reported qualitatively, without the domain analysis which clarifies the relationships between fixed and moving segments of the chain. The comparatively long FPP molecule is thought to enter the reaction cavity via space between the α2 and α3 helices [6].

The domain analysis is thus in agreement with the observation that residues 72-83 are often not resolved in apo-protein crystal structures [15], and change orientation on substrate binding or inhibition. Residues 74, 75, 77, and 81 from the flexible loop are involved in catalytic activity, as demonstrated by mutations [15]. The substrate-binding pocket volume is ~15% larger in apo-UPPS (980Å^3) than in substrate-bound UPPS (850 Å^3). In Figure.1, the substrates have been positioned into apo-UPPS by superimposing the combined substrate coordinates from PDB 1X08 and 1X09 over 3QAS. Although the pockets in both monomers have openings between the α2 and α3 helices, (the yellow surface at the right of Figure. 1, oriented roughly perpendicular to the page) these openings are shorter than FPP. FPP binding may thus be dependent on transient, even more open, structures. These openings are even less a fit for the C_{55} product.

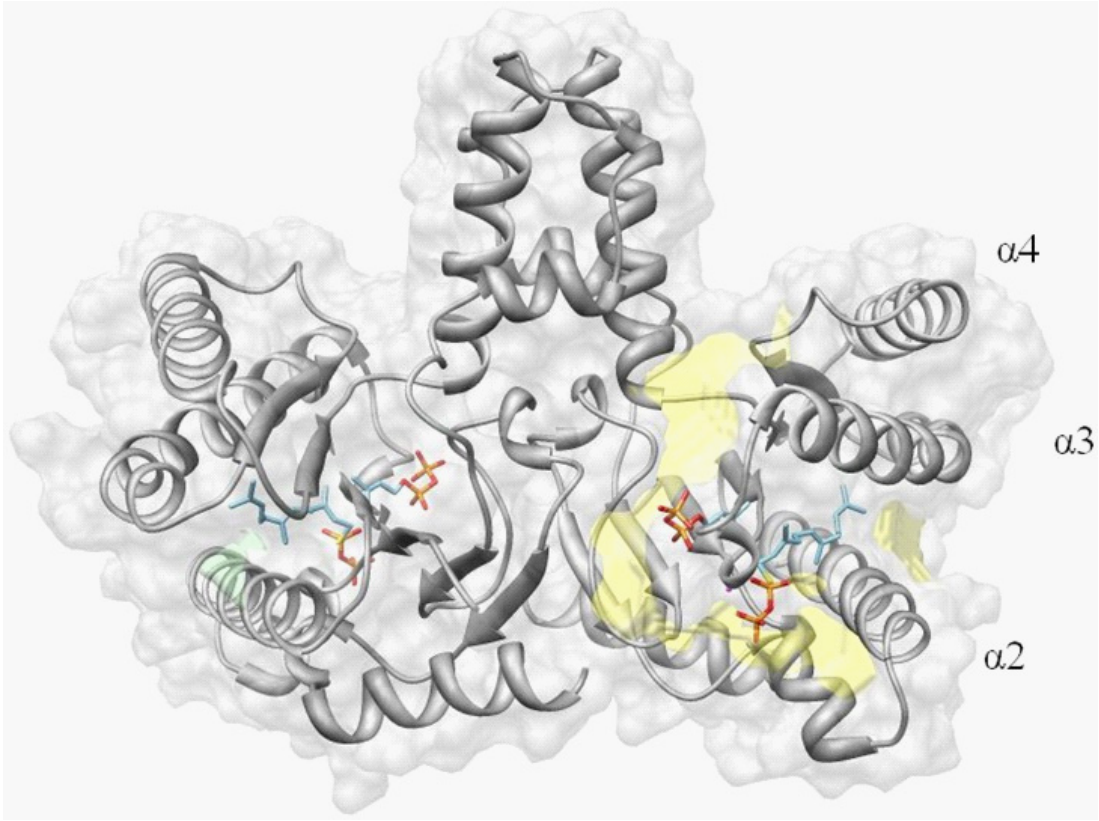


Figure 1: Apo-UPPS (PDB 3QAS) with superimposed substrates FPP and IPP from PDB 1X08. The surface/pocket intersections are shown. The protein surface is transparent light gray, surface/pocket intersections are colored yellow and green.

The Protein Data Bank contains several *E. coli* UPPS structures bound with inhibitors for which coordinates of both protein molecules in the dimer have been deposited (some are PDB 2E98, 2E99, 2E9A, 2E9C, 2E9D) [10]. In these structures, more than one molecule of inhibitor binds, with unequal numbers of inhibitors in the two dimer units. PDB 1UEH has product-analog detergent bound to only one of the dimer units [33]. This suggested asymmetry with functional implications, so the global asymmetry score, 0.7, was computed using those residues from 20 to 235 which were resolved in both dimer units of PDB 3QAS. A recent study of asymmetry in homodimers indicated that 76% of the homodimers studied had global asymmetry scores ≤ 0.4 ('highly symmetrical'), while ~90% had global asymmetry scores ≤ 1. Proteins with scores from 1-3 were classified as having 'limited symmetrical organization' and those with scores > 5 as having 'gross profound asymmetry' [31]. UPPS thus appears to be on the border between symmetrical and modestly asymmetrical homodimers.

There are also openings of significant surface area between helices 2, 3, and 4 and the boundary between the two monomer units in 3QAS. These last may provide an entry point for IPP,

which binds close enough to the surface for the pyrophosphate head group to protrude slightly (blue patch in Figure.4a). It is difficult to picture MgIPP entering via the gap between α2 and α3 in the presence of FPP, and working its way around the FPP to the IPP binding site. There are surface patches of both negative charge (attractive to Mg) and positive charge (attractive to the pyrophosphate) on the surface near the MgIPP and FPP binding sites.

Molecular Dynamics Simulations – apo-UPPS

The protein was stable over the course of the simulation time as apparent from the backbone RMSD (1.9 ± 0.3Å). Principal component analysis indicated that, although significant regions of component space were visited, each monomer sampled a slightly different region in a non-uniform manner; these simulations have not converged in 80nsec, which was the simulation time feasible for us.

In the apo-protein constructed for molecular dynamics simulation, the two cavities had volumes of 989 and 1003 $Å^3$. This slight asymmetry is possibly due to reconstructing the 72-83 loop in chain B, and increased markedly during the simulation. Also, only 75% of 2070 trajectory frames analyzed retained two detectable cavities. In frames with one cavity which were examined visually, the cavities have coalesced. The peak in the cavity volume distributions is between 501 and 750$Å^3$, while the most frequently observed difference between the volumes of the two cavities in a given frame is between 501 and 1000 $Å^3$ (40% of the frames). Thus it was most common for one of the cavities to be rather compact, with the other larger by 50% or more.

Figure.2 is a view of the two cavities in frame 3184 (retrieved after 31.84 ns of production MD), the frame with the maximum calculated cavity volume difference. The large cavity on chain A also has a large area in common with the protein surface (Figure.2a), which continues between helices α2 and α3. In Figure.2b a roughly oval opening between the α3 and α4 helices is also apparent. The overall backbone RMSD for residues 20 to 235 (chain B compared to chain A) has increased from 0.66A in the crystal structure to 1.36 in this frame. The RMSD difference between chains A and B for helices α2, α3, and α4 is, respectively, 1.34, 1.25 and 1.42. The RMSD difference for the loop 72-83 is 4.24. The global asymmetry score for this frame is 1.5. The variability in reaction cavity volume may be related to cavity size requirements for accommodating growing product, and seems to depend more on side chain reorientation than on backbone flexibility, given the modest RMSD between helices α2, α3, and α4 in the two chains in this frame, and the modest asymmetry score. Inspection of this frame revealed several residues at the cavity walls with significantly different orientations in the two chains.

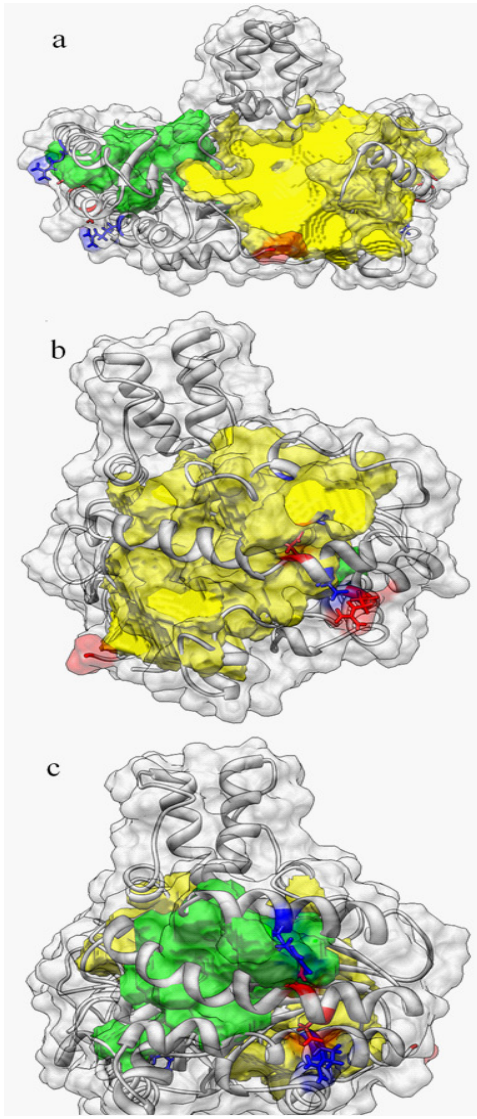


Figure 2: Apo-UPPS trajectory frame 3184 (retrieved after 31.84 ns of production MD), the frame with the greatest volume difference between cavities. The protein surface is rendered in transparent gray. The N-terminal of chain B is highlighted in red. The chain A cavity (volume is 2522 Å^3) is outlined in yellow (right side in panel a, forward in panel b), that in chain B green (volume is 568 Å^3) (left side in panel a, forward in panel c). The brighter the color of the cavity, the closer to the protein surface it is. The electrostatically-interacting residues R51/E96 and D94/R123 are highlighted in blue (R) or red (D,E).

It was apparent from examination of a random selection of trajectory frames that cavity volumes fluctuate in directions that do not contribute to the ability of substrates/products to enter/exit the protein. Spacing between the α2 and α3 helices has been considered the key

difference between an open (able to accept substrate or release product) and closed (reacting) structure [6,7]. The occurrence of cavity intersections with the protein surface between helices α2 and α3 or α3 and α4 did not seem to correlate with maximal overall RMSD from the initial frame. However, when PDB 1X06 was used as reference, and backbone RMSD calculated for residues 84-91 (helix α3), we identified 84 frames with RMSD > 5.0 Å. Spot-check of random frames in this group indicated that such frames do have openings between helices α2 and α3. The reaction cavity volume fluctuations thus appear to result from both motion in the α3-helix and sidechain reorientations along the entire reaction cavity boundary.

It became evident from examining the trajectory that residues R51 and E96, as well as D94 and R123 form salt bridges linking α3 helix with α2 and α4. These pairs also have their side chains oriented toward each other in the crystal structure. Salt bridge analysis of the trajectory indicated that the distance between an N on R51 and O on E96 was 3.3 ± 0.2Å on both chains. These residues are located near the center of the two helices. The α3 and α4 helices were similarly tied together by D94 and R123 for 80% of the trajectory on chain B, (average N-O distance 3.4 ± 0.2Å) but only for 30% of the trajectory on chain A (average N-O distance 5.4 ± 1.3 Å). D94 and E96 are also two of the residues in one of the bending regions identified in the DynDom analysis. These interactions, particularly R51/E96, may, by their presence or absence, serve to regulate enzyme activity by controlling substrate entry and/or product exit.

The positions 51, 94, 96, and 123 are not conserved across UPPS. However, the available crystal structures from five different species all had salt bridges across the α2 and α3 helices, and only one (*Micrococcus luteus*) lacked a bridge between the α3 and α4 helices. In *Helicobacter pylori* there are two basic residues in the α2 helix oriented properly for bridging α2 to α3. In *Mycobacterium tuberculosis* there are two basic residues oriented properly for bridging α3 to α4, while in *Campylobacter jejuni* there is one basic residue oriented so it could bridge to two acidic residues. The bridges occur across slightly different portions of the helices, as is evident from the aligned sequence segments shown in Table 1.

Table 1: Partial alignments of helices α2-α4 in UPPS sequences for which there are PDB entries, showing the positions of salt-bridging residues. Residues numbered as in the PDB entries. Those which are bold and underlined bridge the α2 and α3 helices, those which are italicized and highlighted bridge the α3 and α4 helices. Numbers above the alignments indicate the α2, α3, and α4 helices in *E. coli* UPPS. 1X06 is from E. coli, 1F75 from *Micrococcus luteus*, 2DTN from *Helicobacter pylori*, 2VG4 from *Mycobacterium tuberculosis*, and 3UGS from *Campylobacter jejuni*.

```
                              22222222222222222222
1X06   23 IIMDGNGRWAKKQGKIRAFGHKAGAKSVRRAVSFAANNGIEALTLYAFSS  72
1F75   26 IIMDGNGRWAKQKKMPRIKGHYEGMQTVRKITRYASDLGVKYLTLYAFST  75
2DTN   10 IIMDGNGRWAKLKNKARAYGHKKGVKTLKDITIWCANHKLECLTLYAFST  59
2VG4   73 IVMDGNGRWATQRGLARTEGHKMGEAVVIDIACGAIELGIKWLSLYAFST 122
3UGS   12 VVMDGNRRWARAKGFLAKLGYSQGVKTMQKLMEVCMEENISNLSLFAFST  61

             333333333333333333333             444444444444444
1X06   83 SALMELFVWALDSEVKSLHRHNVRLRIIGDTSRFNSRLQERIRKSEALTA 132
1F75   86 NYLMKLPGDFLNTFLPELIEKNVKVETIGFIDDLPDHTKKAVLEAKEKTK 135
2DTN   70 DFLMKMLKKYLKDERSTYLDNNIRFRAIGDLEGFSKELRDTILQLENDTR 119
2VG4  136 RFLMGFNRDVVRRRRDTLKKLGVRIRWVGSRPRLWRSVINELAVAEEMTK 185
3UGS   70 DFIFELLDRCLDEALEKFEKNNVRLRAIGDLSRLEDKVREKITLVEEKTK 119
```

With FPP binding to UPPS first, it is clear that IPP must enter via an opening between the FPP binding site and the outer part of the reaction cavity formed by the helices α2, α3, and α4. Figure.3 shows a frame selected from the apo-protein trajectory for its large opening to the surface in the vicinity of residues within 4Å of the IPP pyrophosphate in PDB 1X09. Conformations similar to this one probably allow the comparatively small IPP molecule to enter after the much larger FPP has bound.

MD simulations of substrate-bound UPPS

The substrate-bound protein was also stable over the course of the simulation (backbone RMSD 1.8 ± 0.3Å). The volume distribution was much narrower than for the apo-protein. Less than 1% of frames had one cavity (vs. 25% in apo-protein). Less than 1% of volumes were greater than 1500 $Å^3$ (vs. 26%). For reference, the substrates – MgIPP and FPS – had an average combined volume of 780 ± 10 $Å^3$, with a maximum of 820 and a minimum of 745 $Å^3$. The minimum cavity volume measured was 555 $Å^3$. 22% of the cavities were less than 750 $Å^3$ (vs. 36%). The substrate protrudes slightly from the protein surface (Figure.3). The most frequent volume range was 751-1000 $Å^3$, higher than in the apo-protein, and the most frequent volume difference was less than 250 $Å^3$, with a maximum volume difference of 740 $Å^3$, also less than in apo-UPPS. Substrates decrease protein flexibility in MD, consistent with the observation that the flexible 72-83 loop is better-resolved in crystals of protein bound with small molecules, whether substrate analogs,

inhibitors, or detergent, than in apo-protein. The protein structure has tightened to such a degree on binding substrate that there are few connections between the reaction cavity and protein surface, as apparent in Figure.4.

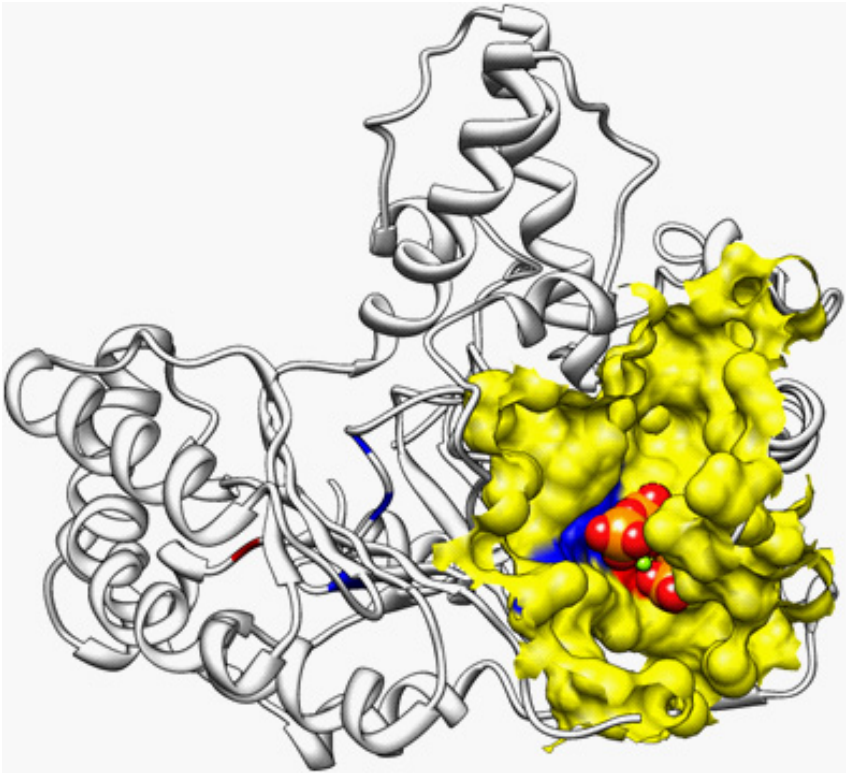


Figure 3: apo-UPPS with substrate binding cavity. Substrate binding cavity is shown in yellow surface representation. Of the charged residues within 4Å of IPP in PDB 1X09, D26, R194, R200, and R202, only parts of some of the positively charged residues are visible in this view (blue surface to the left of the substrate molecules). The substrates are represented by spheres, with Mg the small green sphere between IPP (upper red spheres – oxygen and orange spheres - phosphorus) and FPS (lower). Source of image: trajectory frame 3184 from MD simulation (retrieved after 31.84 ns of production MD).

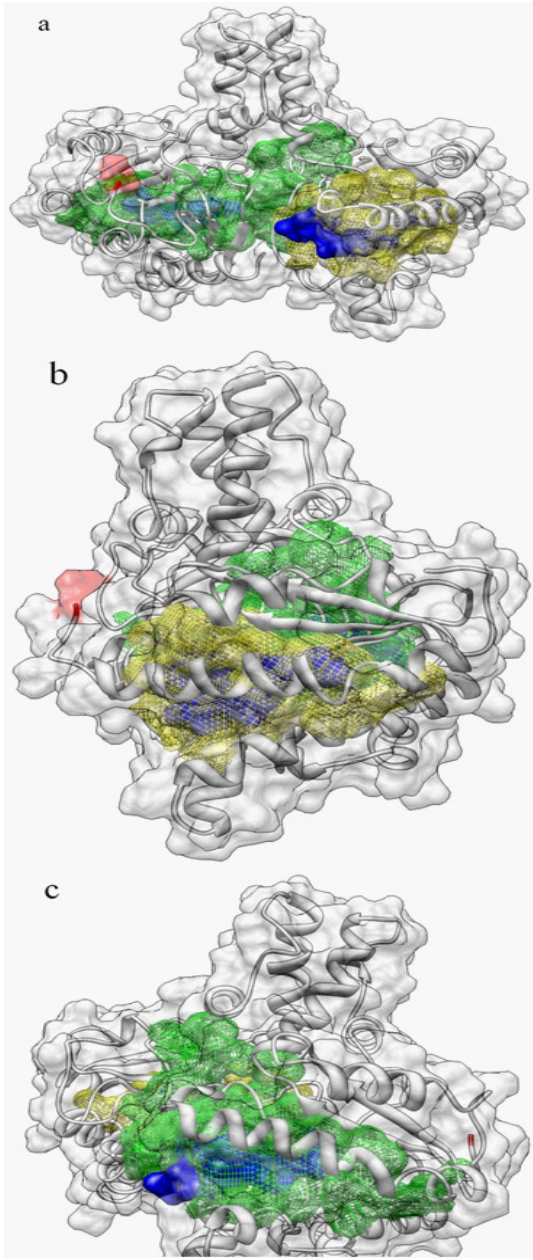


Figure 4: UPPS with bound substrate analogs, with maximum difference in cavity volumes (740 Å3). The N-terminal of chain B is highlighted red. The surface of the chain A cavity is yellow (volume 770 Å3), that of the chain B cavity green (volume 1510 Å3). The substrate analog surfaces are rendered blue. The protein surface has been added in transparent gray. The grayish cast in the cavity surface colors indicates they are all below the surface. Only the bright blue 4a and 4c signifies that some of the substrates protrude from the surface. Source of image: trajectory frame 8016 from MD simulation (retrieved after 80.16 ns of production MD).

MD simulations of product-bound UPPS

In view of the reduction in reaction cavity size induced by substrates, we performed a short simulation (20nsec) with two bound product analogs. Backbone RMSD (1.9 ± 0.1 Å) indicated the model was stable over the trajectory. As this trajectory was so short, we did not plot reaction cavity volume distributions. The average volume of product was 1480 ± 20Å^3; well within the range observed in the apo-protein trajectory. The product volume is roughly double that of the initial substrates, which have about half the number of atoms as the product. Figure.5 shows a frame with a greater exposure of product than was generally observed across the trajectory.

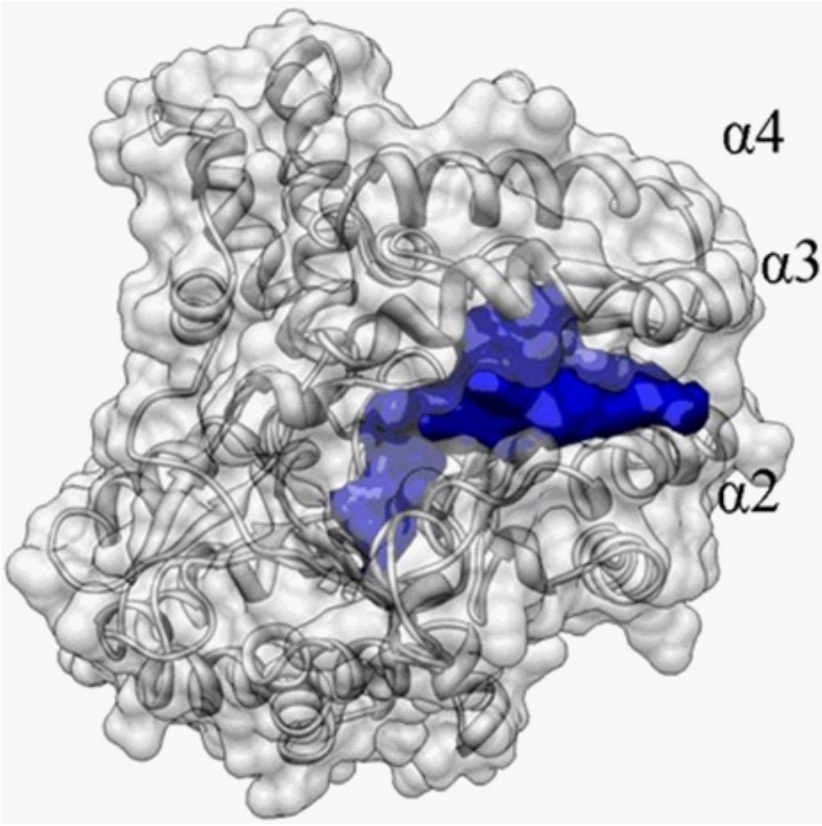


Figure 5: UPPS with modeled product. The view is rotated to show best the surface exposure of product between helices α2 and α3. Bright blue indicates exposed product surface; hazy blue indicates that the product is inside the protein surface, rendered in transparent gray. Source of image: frame 702 from trajectory of MD simulation (retrieved after 7.02 ns of production MD).

SUMMARY AND CONCLUSIONS

Previous MD simulations [4] showed that the reaction cavity in apo-UPPS varied significantly in size, as did ours. Comparison of the volumes with protein surface indicated that the protein may only transiently be able to accept substrates; this may be a mechanism for control of activity.

Salt bridges between R51/E96 and R94/D123 may limit the spacing between the three α-helices (α2, α3 and α4) forming the outer part of the reaction cavity; similar bridges occur in other UPPS, but not at the homologous positions. Transient absence of these salt bridges may thus contribute significantly to conformations able to accept substrate or release product, as motion of the α3 helix of this set is crucial for substrate entry and product release. The most frequently observed conformation of apo-enzyme had one cavity significantly larger in volume (1251-1500 $Å^3$) than the average substrates volume (780 $Å^3$), and one cavity (501-750 $Å^3$) too small or only marginally large enough to accept substrates, consistent with crystal structures having different numbers of inhibitors, substrates, or product analogs in each dimer unit. Backbone RMSD computed across the apo-UPPS trajectory with the substrate-bound crystal structure as reference showed only modest differences, while frames with RMSD > 5Å in the α3-helix were identified. Thus, the computed reaction cavity volume differences seem to originate in a combination of significant differences in sidechain orientation between the dimer units, and α3-helix motion. A model system with substrates in both dimer units was stable under our simulation conditions, as was a model with two product molecules, consistent with reaction cavity volume differences depending more on local orientation than overall protein reorientation. It is unclear what the physiological implications of the structural heterogeneity and asymmetry in UPPS are. We hope this work stimulates further studies.

ACKNOWLEDGEMENTS

We are grateful for a generous grant (TG-MCB100110) of supercomputer time from XSEDE. We are also grateful to Dr. Olaf Lenz for advice regarding use of PBC Tools, and to Dr. Neil Voss for assistance with 3V.

References

1. Teng KH, Liang PH. Undecaprenyl diphosphate synthase, a cis-prenyltransferase synthesizing lipid carrier for bacterial cell wall biosynthesis. Mol Membr Biol. 2012; 29: 267-273.
2. Teng KH, Liang PH. Structures, mechanisms and inhibitors of undecaprenyl diphosphate synthase: a cis-prenyltransferase for bacterial peptidoglycan biosynthesis. Bioorg Chem. 2012; 43: 51-57.
3. Manat G, Roure S, Auger R, Bouhss A, Barreteau H. Deciphering the metabolism of undecaprenyl-phosphate: the bacterial cell-wall unit carrier at the membrane frontier. Microb Drug Resist. 2014; 20: 199-214.
4. Sinko W, de Oliveira C, Williams S, Van Wynsberghe A, Durrant JD. Applying molecular dynamics simulations to identify rarely sampled ligand-bound conformational states of undecaprenyl pyrophosphate synthase, an antibacterial target. Chem Biol Drug Des. 2011; 77: 412-420.
5. Fujihashi M, Zhang Y-W, Higuchi Y, Li X-Y, Koyama T, et al. Crystal structure of cis-prenyl chain elongating enzyme, undecaprenyl diphosphate synthase. Proceedings of the National Academy of Sciences USA. 2001; 98: 4337-4342.
6. Guo R-T, Ko T-P, Chen AP-C, Kuo C-J, Wang AH-J, et al. Crystal Structures of Undecaprenyl Pyrophosphate Synthase in Complex with Magnesium, Isopentenyl Pyrophosphate, and Farnesyl Thiopyrophosphate. J Biol Chem. 2005; 277: 20762–20774.
7. Ko T-P, Chen Y-K, Robinson H, Tsai P-C, Gao Y-G, et al. Mechanism of Product Chain Length Determination and the Role of a Flexible Loop in Escherichia coli Undecaprenyl-pyrophosphate Synthase Catalysis. J Biol Chem. 2001; 76: 47474-47482.
8. Kuo C-J, Guo R-T, Lu I-L, Liu H-G, Wu S-Y, et al. Structure-Based Inhibitors Exhibit Differential Activities against Helicobacter pylori and Escherichia coli Undecaprenyl Pyrophosphate Synthases. J Biomed Biotechnol. 2008; 2008: 1-6.
9. Chang S-Y, Ko T-P, Chen AP-C, Wang AH-J, Liang P-H. Substrate binding mode and reaction mechanism of undecaprenyl pyrophosphate synthase deduced from crystallographic studies. Protein Science. 2004; 13: 971-8.

10. Guo RT, Cao R, Liang PH, Ko TP, Chang TH. Bisphosphonates target multiple sites in both cis- and trans-prenyltransferases. Proc Natl Acad Sci U S A. 2007; 104: 10022-10027.

11. Sinko W, Wang Y, Zhu W, Zhang Y, Feixas F. Undecaprenyl diphosphate synthase inhibitors: antibacterial drug leads. J Med Chem. 2014; 57: 5693-5701.

12. Danley DE, Baima ET, Mansour M, Fennell KF, Chrunyk BA. Discovery and structural characterization of an allosteric inhibitor of bacterial cis-prenyltransferase. Protein Sci. 2015; 24: 20-26.

13. Zhu W, Wang Y, Li K, Gao J, Huang CH. Antibacterial drug leads: DNA and enzyme multitargeting. J Med Chem. 2015; 58: 1215-1227.

14. Lu YP, Liu HG, Teng KH, Liang PH. Mechanism of cis-prenyltransferase reaction probed by substrate analogues. Biochem Biophys Res Commun. 2010; 400: 758-762.

15. Chen YH, Chen AP, Chen CT, Wang AH, Liang PH. Probing the conformational change of Escherichia coli undecaprenyl pyrophosphate synthase during catalysis using an inhibitor and tryptophan mutants. J Biol Chem. 2002; 277: 7369-7376.

16. Takahashi I, Ogura K. Prenyltransferases of Bacillus subtilis: Undecaprenyl Pyrophosphate Synthetase and Geranylgeranyl Pyrophosphate Synthetase. The Journal of Biochemistry. 1982; 92: 1527-37.

17. Allen CM Jr, Muth JD. Lipid activation of undecaprenyl pyrophosphate synthetase from Lactobacillus plantarum. Biochemistry. 1977; 16: 2908-2915.

18. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004; 25: 1605-1612.

19. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph. 1996; 14: 33-38, 27-8.

20. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res. 2004; 32: W665-667.

21. Zoete V, Cuendet MA, Grosdidier A, Michielin O. SwissParam: a fast force field generation tool for small organic molecules. J Comput Chem. 2011; 32: 2359-2368.

22. Bowers KJ, Chow E, Xu H, Dror RO, Eastwood MP, Gregersen BA, et al., editors. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. ACM/EE Conference on Supercomputing (SC06); 2006 November 11-17, 2006; Tampa Fl.

23. Henin J, Lenz O, Mura C, Saam J. PBCTools Plugin User's Guide Verion 2.5

24. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E. Scalable molecular dynamics with NAMD. J Comput Chem. 2005; 26: 1781-1802.

25. Case DA, Cheatham TE 3rd, Darden T, Gohlke H, Luo R. The Amber biomolecular simulation programs. J Comput Chem. 2005; 26: 1668-1688.

26. Chen AP-C, Chang S-Y, Lin Y-C, Sun Y-S, Chen C-T, Wang AH-J, et al. Substrate and product specificities of cis-type undecaprenyl pyrophosphate synthase. Biochem J. 2005; 386: 169-76.

27. Hayward S, Berendsen HJ. Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. Proteins. 1998; 30: 144-154.

28. Voss NR, Gerstein M. 3V: cavity, channel and cleft volume calculator and extractor. Nucleic Acids Res. 2010; 38: W555-562.

29. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res. 2006; 34: W116-118.

30. Williams T, Kelley C. Gnuplot 2.0.

31. Swapna LS, Srikeerthana K, Srinivasan N. Extent of structural asymmetry in homodimeric proteins: prevalence and relevance. PLoS One. 2012; 7: e36688.

32. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G. ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res. 2012; 40: W597-603.

33. Chang SY, Ko TP, Liang PH, Wang AH. Catalytic mechanism revealed by the crystal structure of undecaprenyl pyrophosphate synthase in complex with sulfate, magnesium, and triton. J Biol Chem. 2003; 278: 29298-29307.

Austin Publishing Group

Molecular Dynamics Simulations and Molecular Docking Approaches in Endoinulinase Chemical Modification

Torabizadeh H*

[1]Department of Food Science and Technology, Institute of Chemical Technology, Iran

***Corresponding author:** Torabizadeh H, Department of Food Science and Technology, Institute of Chemical Technology, Iranian Research Organization for Science and Technology (IROST), Azadegan Highway, North to South, Ahmadabad Mostofie, Enghelab Street, Shahid Ehsanirad Street, Tehran, Iran. Tel: +98 21 56276026; Fax: +98 21 56276265; Email: htoraby@alumni.ut.ac.ir

Published Date: December 01, 2016

ABSTRACT

Molecular dynamics simulations and molecular docking approaches were employed to explain the observed structural and functional thermostabilization of endoinulinase (EC 3.2.1.7) through semi-rational chemical modification of surface accessible lysine residues by Pyridoxal-5′-Phosphate **(PLP)** and ascorbate reduction. Improved stability was observed on modifications in the absence or presence of inulin, which indicates storage or functional thermostabilization, respectively. Comparisons have been made between non-modified and modified enzyme by the determination of Tm as an indicator of structural stability, temperature-dependent half-lives (t1/2), and energy barrier of the inactivation process, in a storage thermostability approach. These parameters coincided well with the observed stabilization of the engineered enzyme. The molecular dynamics simulations and molecular docking results revealed that, the establishment of intramolecular interactions between the covalently attached PLP–Lys381 and Arg526 and Ser376 residues can be the origin of the intramolecular contacts in the modified enzyme.

Keywords: Endoinulinase; Molecular dynamics simulations; Molecular docking; Chemical modification; Thermostability

Abbreviations: PLP: Pyridoxal-5′-Phosphate; **Tm:** Transition Midpoint Temperature; **t1/2:** Half-life; **Ea(in):** Activation Energy of Denaturation/Inactivation; **HFS:** High-Fructose Syrup; **ASA:** Accessible Surface Area (in angstrom squared); **DNS:** Dinitrosalicylic Acid; **DSC:** Differential Scanning Calorimetry; **MD:** Molecular Dynamics**; SPC:** Simple Point Charges; **RMSD:** Root Mean Square Deviation

INTRODUCTION

Inulinases are comprised of endo-inulinase (EC 3.2.1.7) and exo-inulinase (EC 3.2.1.80). Inulinases belong to the glycoside hydrolase family GH32 [1,2] based on amino acid sequence comparisons and the presence of conserved amino acid domains. They act as β-fructan fructanohydrolases and hydrolyze inulin to produce High-Fructose Syrup (HFS) and fructo-oligosaccharide. These products are important ingredients in the food and pharmaceutical industries. The structure of inulin consists of a linear polyfructan with β2→1 linkages between fructose residues and a terminal sucrose moiety [3]. Inulin is a storage polysaccharide and is accumulated in the underground tubers of many plants, including the Jerusalem artichoke *(Helianthus tuberosus)*, chicory *(Cichorium intybus),* dahlia *(Dahlia pinnata)*, and dandelion (Taraxacum officinale) [4]. Because of solubility limitation and microbial contamination of inulin, the industrial hydrolysis of inulin is carried out at ≥50 ˚C, which is necessary to obtain an appropriate hydrolysis rate. At this temperature, most of the inulinases lose their activity after a few hours. Therefore, there is a growing interest in introducing inulinases with improved thermostability [5]. The thermal stability of enzymes is an important parameter in enzymatic processes as it determines the limits for use and reuse of the enzyme, and therefore affects the cost. In our previous work, we reported the semi-rational modification of endo-inulinase by using Pyridoxal 5′-Phosphate **(PLP)** [6] as a specific modifier of lysine residues [2]. We have reported that lysine residues are more accessible at the surface of the C-terminal domain compared to the N-terminal domain of the inulinase based on the calculated Accessible Surface Areas **(ASA)** of 123.1 and 74.1 Å2, respectively [6]. Even though, the PLP modifications have brought about enzyme inactivation in several studies [7–16], we reported a case in which not only the enzyme activity was retained but also thermostability was improved. In this reaction, a Schiff base is formed between the ε-amino group of lysine as a nucleophile part and the aldehyde group of PLP [17–19]. The selectivity of the reaction, the spectral properties of the modified product, and the reversibility of the reaction are among the advantages of this chemical modification strategy [6,19]. Although, sodium borohydride is routinely used as a Schiff base-reducing agent, a safe and efficient reduction method using ascorbic acid as a novel Schiff base-reducing agent has been used. In this study, the knowledge of modeling, kinetics, and thermodynamics have been brought together to explain the endo-inulinase stabilization process by using PLP and ascorbate as the modification and reduction agents, respectively.

MATERIALS AND METHODS

Materials

Ascorbic acid was purchased from Acros Organics (Morris Plains, NJ, USA). Endoinulinase (27 U/mg), inulin (chicory inulin), and pyridoxal 5′-phosphate were obtained from Fluka (Switzerland) and Dinitrosalicylic Acid **(DNS)** and other chemicals were purchased from Sigma Chemical Company (St. Louis, MO, USA).

PLP modification of endo-inulinase

Endo-inulinase modification was performed by using PLP and ascorbate as the modifier and reducing agents, respectively [6]. In brief, endoinulinase at 0.1 mg ml−1 was modified by using PLP at 31.25 μM in 50 mM sodium phosphate buffer (pH 7.5) for at least 30 min then the enzyme–PLP complex was stabilized by reduction with ascorbic acid using 100 μl of freshly prepared solution of ascorbic acid (1 mM) in 50 mM phosphate buffer pH 7.0. To remove the excess ascorbate and PLP, overnight dialysis was performed against 50 mM sodium acetate buffer of pH 6.0 at 4 ˚C.

Enzyme assay

Endoinulinase assay was carried out through the assessment of the librated reducing sugar by using DNS. In brief, the assay mixture (55 μg ml−1 enzyme and 0.36 mg ml−1 inulin in 50 mM sodium acetate buffer of pH 5.5–6) was incubated at 37 ˚C for 15 min. The same concentration of sucrose was used for determining invertase activity instead of inulinase activity of the enzyme species. Next, the reaction was terminated by adding an equal volume of DNS reagent followed by incubating at 90 ˚C for 10 min. The absorbance was measured at 575 nm using a Camspec M550 spectrophotometer in cells with 1 cm path length. One unit inulinase or invertase is defined as the amount of enzyme that liberated 1 μmol fructose (or glucose) per minute under assay conditions [20,21].

Structural thermostability analysis of endoinulinase

Comparative structural and functional thermostability analysis of non-modified and modified endoinulinases was carried out using Differential Scanning Calorimetry **(DSC)** measurements with a Model 6100 Nano II differential scanning calorimeter (Calorimetry Sciences Corp., USA) at a heating rate of 2 ˚C min−1 between 10 and 85 ˚C, under an extra constant pressure of 2 atm. The sample protein was used at 2.0 mg ml−1 in 50 mM phosphate buffer of pH 7.5, after degassing. The same buffer was used as a reference to perform the baseline run. For data collection and estimation of Tm values, the standard CpCalc software package and data acquisition program, DSC Run, were used [22].

Molecular dynamics simulation

The crystal structure of the native endo-inulinase from Aspergillus niger has not yet been elucidated, whereas the crystal structure of exo-inulinase from Aspergillus awamori has been

reported [2]. To determine the identity and similarity between amino acid sequences of exo-inulinase from A. awamori (EC 3.2.1.80, with 537 amino acid residues) and endo-inulinase from A. niger (EC 3.2.1.7 with 494 amino acid residues), binary sequence alignment was performed using the FASTA program version 3 (35.04) that is located on the ExPasy server. The results revealed that the sequence identity between the two proteins was 31.6% and the similarity was 62.9–64.5% over aligned residues.

Because, the similarity value between endo-inulinases and exo-inulinases was greater than 50%, for exo-inulinase, the three-dimensional crystal structure coordinates were used to determine the Molecular Dynamics **(MD)** simulation and docking studies of endoinulinase [6]. MD simulations were performed using the GROMACS program (version 4.0.3) with the GROMACS force field [23]. The starting geometries for the simulation were prepared from the initial coordinates of protein that were extracted from the modeled structure of inulinase in the protein data bank (PDB, entry 1Y9M). The protein consisted of 537 residues. Modified lysine was created as a new residue in the GROMACS program database. For this purpose, a PLP–lysine parameter was created by the PRODRG server [24], which is used to generate topologies for ligand–protein Complexes. The parameters were then transferred to the GROMACS-related libraries. The MD simulation protocol is as follows: The protein was first placed in a simulation cubic box of a suitable size and solvated with a 37,578 Simple Point Charge **(SPC)** water model and 20 Na+ counter ions to neutralize the entire negative charge. The water molecules and ions were subjected to energy minimization, while the inulinase was kept fixed in its initial configuration. Subsequently, the water and ions were allowed to evolve using MD simulation for 50 ps with a step time of 1 fs, again keeping the structure of the inulinase fixed. Next, the entire system was minimized using the steepest descent of 1,000 steps followed by conjugate gradients of 9,000 steps. In order to obtain equilibrium geometry at 300 K and 1 atm, the system was heated at a weak temperature (τ=0.1 ps) and pressure (τ=0.5 ps) coupling by taking advantage of the Berendsen algorithms [25]. The heating time for MD simulation at 100 and 200 K was 100 ps with a nonbounded cutoff of 14 Å. The MD simulation was further carried out for 2 ns at 300 K. We used LINCS to constrain the bond length [26]. MD simulations were carried out by particle mesh Ewald method [27]. The dynamic behavior and structural changes of the protein were analyzed by calculation of the Root Mean Square Deviation **(RMSD).**

Docking analysis

The intramolecular residues interacting with PLP were examined by using the LIGPLOT program. The program generates Lys–PLP interactions from three-dimensional coordinates in a 1Y9M PDB file (PDB ID, 1Y9M). The LIGPLOT algorithm was described by Wallace et al. and Turnay et al. in 2002 [28]. Simulations were performed for both the nonmodified and modified enzyme species.

RESULTS

Structural Thermostability Analysis

We examined the thermodynamic and molecular simulation approaches to explain the structural, storage, and functional thermostabilization of endo-inulinase after the application of PLP modification followed by ascorbate reduction. According to our previous report [6], Figure 1 presents the heat capacity scans of the non-modified and PLP modified without reducing (En-PLP), and PLP modified with reducing by ascorbic acid (En-PLP-As). The results revealed that the thermal denaturation of endo-inulinase was an irreversible process. The Tm of modified enzyme (En-PLP-As) was 72.2 ˚C, whereas it was 64.1 ˚C for the non-modified and 64.9 ˚C for PLP-treated (without reducing) sample (En-PLP). Accordingly, the Tm of the enzyme is shown to have an 8.1 ˚C increase upon modification and emphasizes on the necessity of the reducing step. Therefore, the structural thermo-stabilization of the endo-inulinase has been achieved byusing this modification strategy consisting of PLP treatment and ascorbate reduction.

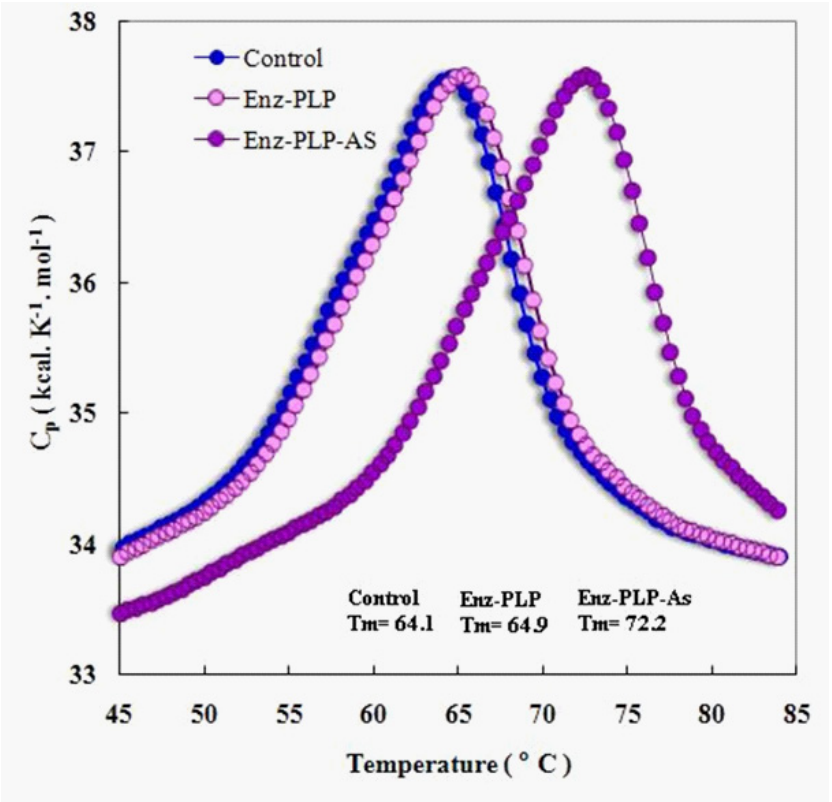


Figure 1: Heat capacity scans of non-modified endo-inulinase (control) and modified without (Enz-PLP) and with reduction (Enz-PLP-As) by ascorbic acid. The standard CpCalc software package and data acquisition program, DSC Run, were employed for data analysis and evaluation of Tm values, (n=3).

Molecular Dynamics Simulation and Docking Analysis

Molecular dynamics has enabled us to simulate the PLP-modified species of the enzyme to determine that modification created additional intramolecular contacts, which are thought be responsible for improving the thermostability and therefore the resistance as it relates to the thermoinactivation of the endoinulinase. Figure 2 shows the simulated structure and focuses on lysine residue (Lys381) at the linker closer to the C-domain, which is prone to modification.

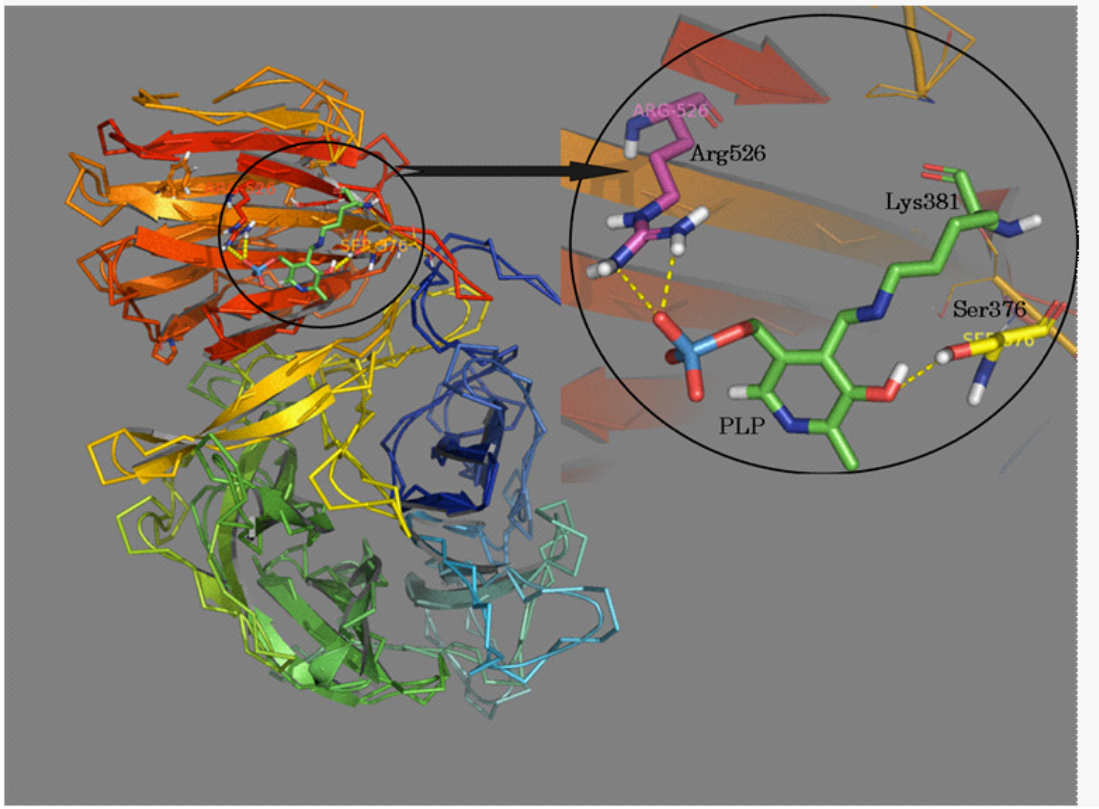


Figure 2: The simulated structure of PLP-modified endoinulinase on Lys^{381} at the link closer to the C-domain. Molecular dynamics simulations were performed by using the GROMACS program (version 4.0.3). For more details, please see the "Materials and Methods" section. In the modified enzyme, additional intramolecular contacts between covalently attached PLP-Lys^{381} with guanidinium group of Arg^{526} and at the hydroxyl group of Ser^{376} are formed.

The RMSD value as a factor of the equilibrated modified lysine inulinase and non-modified lysine is shown in Figure 3. It is evident that after 300 ps, both the structures reached an equilibrated state.

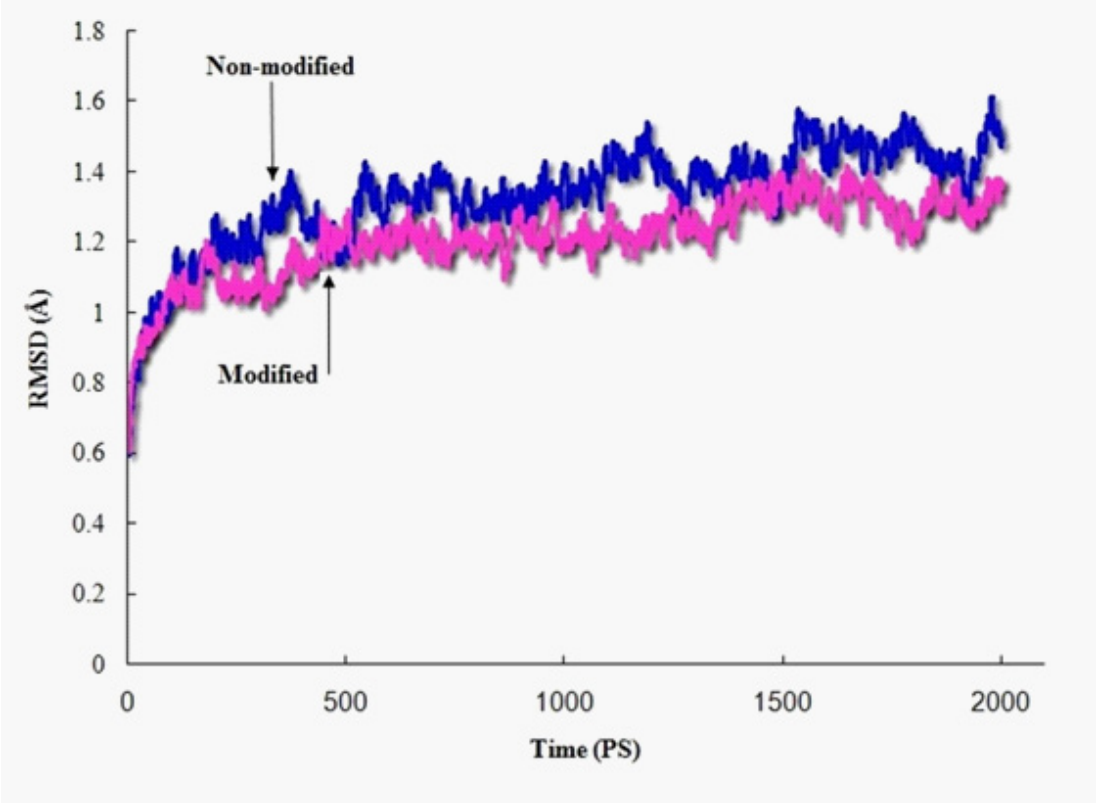


Figure 3: Presents the RMSD versus time plot for the non-modified and PLP-modified endoinulinase during 2 ns of MD simulation.

The modified molecule was analyzed against the non-modified one by using the LIGPLOT program to examine the newly established intramolecular hydrogen bonds after the PLP modification process at the portion of Lys381 (Figure 4). Modification made it possible to create additional intramolecular contacts between covalently attached PLP–Lys381 with Arg526 (two possible hydrogen bonds between the phosphate of PLP and the guanidinium group of Arg526) and with Ser376, as illustrated in Figures 2 and 4. Also, additional intramolecular hydrophobic interactions are formed due to the modification process, in which Pro383 is involved. It seems that the modification process involved an interaction that enhanced the rigidity of the enzyme molecule.

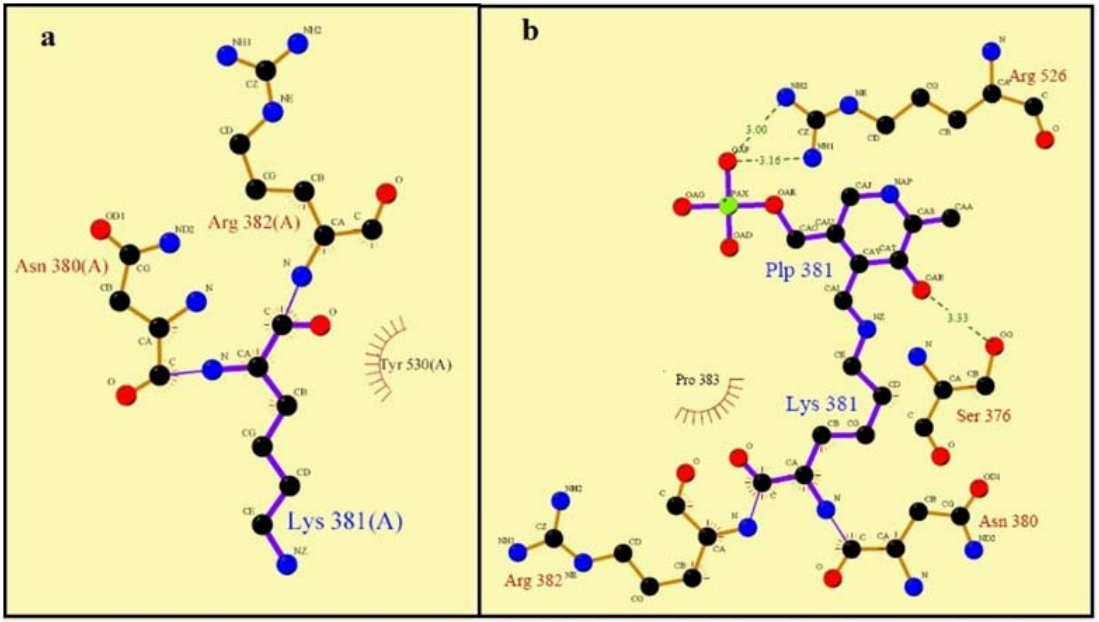


Figure 4: Docking analysis of intramolecular interactions in a two-dimensional representation of simulated endoinulinase structure [based on *A. awamori* exoinulinase (PDB ID, 1Y9M)], before (a) and after (b) PLP-modification at the Lys^{381} using the LIGPLOT program. New hydrogen bonds are formed between the phosphate group of PLP and Arg^{526} and Ser^{376} . The *dashed lines* indicate hydrogen bonds and the values indicate hydrogen bond lengths (in angstrom).

DISCUSSION

We are reporting the chemical modification of endo-inulinase to improve the thermostability of the enzyme through covalent attachment of the PLP molecules to the accessible lysine residues followed by ascorbate reduction [6]. Although most reports on the PLP modification of enzymes have resulted in enzyme inactivation due to the engagement of essential residues at the active site or non-desired structural alterations [14,18], our observations have demonstrated that the PLP modification of endoinulinase not only result in no activity loss but also brings about thermostability.

Modified endoinulinase has an industrial potential for inulo-oligosaccharides and HFS production for use in food and pharmaceutical products. Thus, the industrial process for HFS and fructo-oligosaccharide production from inulin could be carried out at the higher temperatures necessary to achieve the appropriate hydrolysis rate. The DSC results (Figure.1) revealed that PLP modification of endoinulinase increases the Tm of the enzyme from 64.1 to 72.2 ˚C (8.1 ˚C increase in Tm). The thermal unfolding processes of the enzyme samples were achieved to be irreversible, as no transition peaks were obtained in the second run of heating. Therefore, under

such conditions, the thermodynamic parameters can be calculated using Ea (in) [13,21,28,29]. The results of the DSC analysis showed good agreement with the observed 4% increase in α-helical contents revealed by circular dichroism spectropolarimetry (details not shown) [6]. Protein thermostability has been reported to correlate with the larger fraction of residues in the α-helical conformation (Table 1) [29].

Table 1: The results of the secondary structure analysis of endoinulinase, before and after modification [6].

		Scondary structure		
	α- Helix (%)	β- Sheet (%)	Turns (%)	Random (%)
Control	13.60	42.60	14.50	29.30
En-PLP	13.80	44.50	13.10	28.60
Endo-PLP-As	17.60	31.80	20.20	30.40

Molecular dynamics and simulation approaches were used to elucidate the newly established interactions that explain the thermostabilization of the engineered enzyme. The simulated structure of the PLP-modified enzyme at the linker closer to the C domain, which is prone to modification (Figure.2) and the constructed LIGPLOT (Figure.4), presents newly established intramolecular interactions after modification at Lys381. This modification results in two possible hydrogen bonds between phosphates covalently attached to PLP–Lys381 and the guanidinium group of Arg526 residue as well as between the hydroxyl group of PLP with Ser376. The hydrophobic groups of amino acid residues become closer through the formation of these intramolecular interactions. In conclusion, the structural, storage, and functional thermostabilization of PLP-modified/ascorbate-reduced endoinulinase has been documented. In conclusion, the substrate specificity, structural, and storage/functional thermostability of the enzyme were efficiently improved upon PLP modification strategy, which has a potential application in the inulinase-based technology for the production of HFS and fructooligosaccharides. The mechanism of thermostabilization was assumed to be involved with the establishment of intramolecular interactions between covalently attached PLP–Lys381 and both the Arg526 and Ser376 residues as representative modification originated intramolecular contacts in the modified enzyme.

ACKNOWLEDGMENTS

We are grateful for the support from the research council of the University of Tehran.

References

1. Coutinho PM, Henrissat B. Recent advances in carbohydrate bioengineering. London: The Royal Society of Chemistry. 1999.
2. Nagem RAP, Rojas AL, Golubev OS, Korneeva EV, Eneyskaya AA, Kulminskaya KN, et al. Crystal structure of exo-inulinase from Aspergillus awamori: the enzyme fold and structural determinants of substrate recognition. J Mol Biol. 2004; 344: 471-480.
3. Edelman J, Jefford TG. The metabolism of fructose polymers in plants. 4. Beta-fructofuranosidases of tubers of Helianthus tuberosus L. Biochem J. 1964; 93: 148-161.
4. Shapiro S, Enser M, Pugh E, Horecker BL. The effect of pyridoxal phosphate on rabbit muscle aldolase. Arch Biochem Biophys. 1968; 128: 554-562.

5. Gill PK, Manhas RK, Singh P. Hydrolysis of inulin by immobilized thermostable extracellular exoinulinase from Aspergillus fumigates. J Food Eng. 2006; 76: 369-375.

6. Torabizadeh H, Habibi-Rezaei M, Safari M, Moosavi-Movahedi AA, Razavi H. Semi-rational chemical modification of endoinulinase by pyridoxal 5´-phosphate and ascorbic acid. J Mol Cat. B: Enz. 2010; 62: 257-264.

7. Jones CW, Priest DG. Interaction of Pyridoxal 5-phosphate with Apo-Serine Hydroxymethyltransferase. Biochem Biophys Acta. 1978; 526: 369-374.

8. Paech C, Tolbert NE. Active site studies of ribulose-1, 5-bisphosphate carboxylase/oxygenase with pyridoxal 5'-phosphate. J Biol Chem. 1978; 253: 7864-7873.

9. Moldoon T G, Cidlowski JA. Specific modifications of rat uterine estrogen receptor by pyridoxal-5´-phosphate. J Biol Chem. 1980; 55: 3100-3107.

10. Gould KG, Engel PC. Modification of lactate dehydrogenase by pyridoxal phosphate and adenosine polyphosphopyridoxal. Biochem J. 1980; 191:365-371.

11. Ogawa H, Fujioka M. The reaction of pyridoxal 5'-phosphate with an essential lysine residue of saccharopine dehydrogenase (L-lysine-forming). J Biol Chem. 1980; 255: 7420-7425.

12. Ohsawa H, Gualerzi C. Structure-function relationship in *Escherichia coli* factors. Identification of a lysine residue in the ribosomal binding site of initiation factor by site-specific chemical modification with pyridoxal phosphate, J Biol Chem. 1981; 256: 4905.

13. Sánchez-Ruiz JM, López-Lacomba JL, Cortijo M, Mateo PL. Differential scanning calorimetry of the irreversible thermal denaturation of thermolysin. Biochemistry. 1988; 27: 1648-1652.

14. Chen CH1, Wu SJ, Martin DL. Structural characteristics of brain glutamate decarboxylase in relation to its interaction and activation. Arch Biochem Biophys. 1998; 349: 175-182.

15. Vojtechová M, Rodríguez-Sotres R, Muñoz-Clares RA. Dehydrogenase from amaranth leaves by pyridoxal 5?-phosphate. Plant Sci. 1999; 143: 9-17.

16. Strucksberg KH, Rosenkranz T, Fitter J. Reversible and irreversible unfolding of multi-domain proteins. Biochim Biophys Acta. 2007; 1774: 1591-1603.

17. Lundblad RL, Noyes CM. Chemical Reagents for Protein Modification. Florida: CRC press. 1985.

18. Costa B, Giusti L, Martini C, Lucacchini A. Chemical modification of the dihydropyridines binding sites by lysine reagent, pyridoxal 5'-phosphate. Neurochem Int. 1998; 32: 361-364.

19. Gao Z, Keeling P, Shibles R, Guan H. Involvement of lysine-193 of the conserved "K-T-G-G" motif in the catalysis of maize starch synthase IIa. Arch Biochem Biophys. 2004; 427: 1-7.

20. Miller GL. Use of dinitrosalicylic acid reagent for determination of reducing sugar, Anal Chem. 1959; 31: 426-428.

21. Vogl T, Jatzke C, Hinz HJ, Benz J, Huber R. Thermodynamic stability of annexin V E17G: equilibrium parameters from an irreversible unfolding reaction. Biochemistry. 1997; 36: 1657-1668.

22. Rosengarth A, Rösgen J, Hinz HJ, Gerke V. A comparison of the energetics of annexin I and annexin V. J Mol Biol. 1999; 288: 1013-1025.

23. Lindhal E, Hess B, Van der Sopel D. GROMACS 3.0: A package for molecular simulation and trajectory analysis, J Mol Mod. 2001; 7: 306-317.

24. Schüttelkopf AW, van Aalten DM. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. Acta Crystallogr D Biol Crystallogr. 2004; 60: 1355-1363.

25. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath, J Chem Phys. 1984; 81: 3584-3590.

26. Hess B, Bekker J, Berendsen HJC, Fraaije JGEM. LINCS: a linear constraint solver for molecular simulations, J Comp Chem. 1997; 18: 1463-1472.

27. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method, J Chem Phys. 1995; 103: 8577-8593.

28. Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng. 1995; 8: 127-134.

29. Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostability. Protein Eng. 2000; 13: 179-191.

TECHNOLOGICAL SPONSOR

Venture Pharmaceuticals Ltd is an international company, dealing globally at intersection of pharmaceutical Discovery, Investment, and Marketing.

www.venture-pharmaceuticals.com

CONTACTS: Venture Pharmaceuticals Ltd,

P.O. Box: 364, Second Floor, 21 Regent Street, Belize City, Belize; Tel/Fax: +44-203-695-8752; E-mail: info@venture-pharmaceuticals.com

The Mission of Venture Pharmaceuticals Ltd – is to be driver of pharmaceutical discoveries. We are connecting various elements of drug discovery pipeline into one by own self. We provide such services to global pharmaceutical corporations as drug discovery research, drug marketing, and drug investment. We are sure our incredible efforts will let Humanity enjoy new powerful and amazing medicines.

SPECIALTIES

Venture Pharmaceuticals (start-up medicines), Drug Investment (Pharmaceutical Investment) Services, Drug Discovery (Pharmaceutical Discovery) Services, Drug Marketing (Pharmaceutical Marketing) Services.

Table 1: Services For Scientists, Research Centers and Pharma Industry.

	FOR SCIENTISTS	FOR RESEARCH CENTERS	FOR PHARMA INDUSTRY
Free and Custom Consulting	How to attract funds, how to prepare grant proposal, where to submit it, which are the key points in attracting money. Use of molecular modelling methods, software in scientific research – workshops and trainings.	Learn how to use molecular modelling in scientific studies. Get trainings on software packages for molecular modelling. Get training on picture preparation for molecular biologists and biochemists. Get training on grant preparation, search, and submission. Learn how to attract commercial collaborators and partners and deal with them.	How to get more customers, how to get higher profit? How to involve molecular modelling and computational chemistry to increase revenue from your business? How to increase effectiveness of your web-site? How to utilize outsourcing marketing and sales services to get cheaper and higher effectivity result?
Molecular Modelling	Performing molecular modelling and computational chemistry works for primarily biologists, synthetic chemists research groups (for direct payment; participation in grants; and paper co-authorship)	To supplement biochemistry, chemistry, molecular biology works in research center; to guide design of compounds for synthesis; to analyze with QSAR the in-house compounds synthesized; picture preparation service for molecular biologists and biochemists.	Make higher quality decisions on what to synthesize or extract. Cut spends and utilize our outsourcing services. Get our assistance in designing of compounds for synthesis, in analyzing your internal structural and bioactivity data, get guidance to increase directness of your synthetic and experimental works. Use our modelling services in preparation of better marketing materials; get high quality pictures for your scientific reports and publications.
Software Development	Developing and improving/ modifying software which research group uses for research (for direct payment; participation in grants; and paper co-authorship)	Design and development of custom software for scientific research; building beautiful and attracting web-site for your institution and its administration; developing, administration of databases.	Get better results using your custom software without spending resources and funds for internal development. Scientific software, databases, web-applications, marketing and sales automation solutions; website design, development, and administration.
Business Development	We assist in preparation of a grant proposal and its submission to funding bodies.	Outsourced business development department: searching for suppliers, partners, and customers for research center. Assisting in grant proposal preparation, submission, and administration; planning for funds attracting for the research center.	Optimize your business effectiveness, without spending much money and resources for that. Get cheaper and higher quality sources suppliers. Increase number of distributors for your products and services.

E-mail: info@venture-pharmaceuticals.com

Table 2: International Distribution, Sales and Production Services.

Import-Export Activity	International Logistics and Supply of Medical Products to Distribution Chains, Local Pharmaceutical Distributors. Focus on specialized patented medicines for specific human diseases.
International Supply and Distribution of Pharmaceutical Products	Servicing of patients affected with dangerous diseases, with population from 30 million – HIV, Tuberculosis, Hepatitis, etc. Work with Government Bodies, WHO, Non-Commercial Organizations, and Charities
Servicing for the large human populations	Organizing of Regular Testing for People on Special and Dangerous Diseases. Collaborating with Government Bodies, WHO, Non-Commercial Organizations, and Charities.
Internet Services	Paid Services for Healthy/Diseased People: Insurance, Discounted Medicines, Information Supply
Production for Customer Products	Unique Customer Products: Pharmaceuticals, Cosmetics
Invitation for Potential Agent Companies to be Our Local Country Agents	Performing inside-country services for us and for our customers in your country, while being our Agent

E-mail: info@venture-pharmaceuticals.com

Table 3: Discovery, Marketing and Investment Services.

Discovery	Computational Services	
	Computational Bio- and Chemo- informatics: Drug Target search Hit Compounds Search Lead Compounds Search Compounds Optimization QSAR Studies Pharmacophore Modeling Homology Modeling Design and Structure Optimization of Chemicals and Biologicals Molecular Dynamics simulations and analysis	Software training: We provide workshops and trainings on a wide selection of both free-available and commercial software packages for molecular modelling, molecular dynamics simulations, virtual screening, docking, homology modelling, bio-molecule visualization, and others
		Software Sales, Hardware Consulting, Installation, Administration: wide selection of software packages for drug design and drug discovery, molecular modelling, molecular dynamics simulations, virtual screening, computer-aided docking, homology modelling, bio-molecule visualization, and others
	Training and Consulting: All fields of structure-based and ligand-based computer-aided design	Software Design and Developing: Data Bases Design, Developing, and Administration Website Design , Developing, and Administration
	Laboratory Services	
	Synthetic Chemistry: Custom Synthesis Scheme Development Scheme Optimization From µg to kgs scale	Biochemistry: Any type of biochemical studies
Marketing	Marketing and advertising materials: Web-Design and Developing: web-pages, web-sites, e-shops Logo Design, Booklet Design	Business Development: Planning Networking Sales and Sourcing Channels Development
	Marketing and Advertising Activity Plans, Promo-Actions: Design Organizing	
Investment	For Seeker Company – Search for investor and representing them investment project: Angel Investors Venture Investors Investment Institutions	For Investor Company – Search for Investment Projects, pre-selection, and representing them to the investment body: Start-Ups Young Companies Matured Companies

E-mail: info@venture-pharmaceuticals.com