Targeting REP:GGTase-II Interaction and Finding New Means to Predict the Protein:Ligand Interactions

DISSERTATION

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Mahesh Kulharia B.Sc., M.Sc. in Biotechnology,

aus Hissar, India

eingereicht bei der
Fakultät Chemie
der Technische Universität Dortmund

Dortmund 2007

PDF-XChange Click to buy NOW! www.docu-track.com

Erstgutachter:	Prof. Dr. Alfons Geiger
Zweitgutachter:	Prof. Dr. Roger Goody
Dritter Prüfer:	Prof. Dr. Martin Englehard

Dedicated to Maa

CONTENTS

Symbols and Abbreviations

Å	Angstrom
CADD	Computer Aided Drug Design
CBP	Carbohydrate Binding Propensity
DPS	Differential Propensity Score
GDP	Guanosine Diphosphate
GGpp	Geranylgeranyl Pyrophosphate
GGTase	Geranylgeranyl Transferase
GTP	Guanosine Triphosphate
HBA	Hydrogen Bond Acceptor
HBD	Hydrogen Bond Donor
IC50	Inhibition Concentration 50%
InCa	Inositol-Carbohydrate
LBDD	Ligand-based Drug Design
M	Molar
nM	nano Molar
NI	Negative Ionisable
PSSBC	Propensity Score of a Site to Bind Carbohydrates
PSSBnC	Propensity Score of a Site to Bind non Carbohydrates
PI	Positive Ionisable
RBP	Rab Binding Platform
REP	Rab Escort Protein
ScoreJE	Score Joint Entropy
SIScoreJE	Solvation Included Joint Entropy
SBDD	Structure-based Drug Design
vdW	Van der Waals
VS	Virtual Screening
ZBG	Zinc Binding Group

Acknowledgements

I would like to take this opportunity to convey my heartfelt thanks to the following people,

Prof. Roger Goody for sharing my problems and extending care and rock-solid support in times of need. His unwavering support has been an immense source of strength and hope.

Prof. Alfons Geiger for listening to my presentations and the valuable advice and guidance he provided during PhD.

Prof. Martin Englehard for taking out time to view my project work and for the kind words of support and encouragement.

Dr. Richard Jackson for burning midnight oil in helping me to realise the objectives we had set.

Dr. Alexey Rak and **Dr. Olena Pylypenko** for the help they extended in the initial stages of my work.

Prof. David Westhead for asking questions and giving good advice when needed.

I am grateful to my father, brother and wife Sarita for being there in times of rough sea.

Thanks Nagaraj and Gurpreet Singh for listening to my daily quota of ideas and giving me valuable advices on feasibility of some of the projects; Amit Sharma and Parbhu Dayal Jakhar for being close to my family in times of need. Harry Mathala for helping me with programming and providing me insights of molecular modelling.

Dr. Tetsuya Kitaguchi, Dr. Anne Adida, Dr. Sergei Mureev (Captain Barbossa) and Melina Terbeck for support and help they provided. Dr. Nicola Gold, Dr. Monika Rella, Dr. Nick Burgoyne, James Dalton and John Fuller for valuable discussions (especially on cricket). Dr. Christoph Schwiteck and Dr. Sean McKillen for Linux software support. Past and present members of the MPI Dortmund and Bioinformatics group in University of Leeds for the friendly help and support.

Special gratitude to **Christa Hornemann** for the magic wand she has for solving all administrative problems.

Chapter 1: Introduction

1.1 General Background

Drugs are single or combinations of small molecules with defined composition and specific pharmacological effect. The process of identification of new drugs is regulated by legal agencies like “Food and drug administration”. This process can be divided in to the phases of drug discovery and drug development. Drug discovery process involves the application of different conceptual strategies to obtain novel protein activity modulators, deduction of the mechanism of these compounds, lead demonstration and optimisation, *in vivo* proof of concept and simultaneous demonstration of a therapeutic index. Drug development begins when the drug molecule is put in phase I clinical trials.

On an average the time from conception of the targeting strategy to the grant of approval by a regulatory authority for a new drug molecule is 10-15 years. It is estimated majority of drug candidates fail along the way. This results in huge loss for consumers (pharmacy companies pass their loss to patients) as the cost of bringing a new drug to market is close to a billion dollars (Dimasi *et al*, 2000). Hence the a number of approaches have been adopted to help distinguish the druggable targets from non-druggable ones. One of the major goals of computational chemistry, or the rational design of compound libraries, is to maximise diversity, to enhance the potential of finding active compounds in the initial rounds of virtual screening programs. Drug discovery has traditionally required testing of hundreds of individually synthesized and characterized chemicals; the new techniques of virtual synthesis in computational chemistry, and virtual screening (VS) offer the possibility of rapidly preparing and examining hundreds of compounds. This increased screening ability dramatically increases the probability of finding a lead compound with the proper balance of activity, specificity, safety, bioavailability, and stability to result in a successful new drug. The process is generally termed as computer aided drug design.

1.2 Computer-Aided Drug Design

Computer-Aided (Assisted) Drug Design (CADD) is a generic term used to address various computer-based drug design strategies. This field can broadly be divided into two categories (1) Ligand-Based Drug Design (LBDD) exploiting information of known actives and (2) Structure-Based Drug Design (SBDD) carried out in the presence of a protein structure. The important background relating to protein-ligand interactions is discussed below. Since the application of computational techniques have the objective of designing the small molecules and this we designed inhibitors for the disruption of REP:GGTase-II interaction, the general aspects of protein-protein interaction followed by concepts in current understanding about kinetic and thermodynamic aspects of protein-ligand binding are discussed. The general methods used in the process of CADD are discussed next i.e. virtual screening (VS), docking, and *de novo* drug design.

1.3 Protein-protein interaction: general aspects

Protein-protein interaction is the fundamental process for the functioning of the huge number of processes in the living cells. Malfunctioning of any part of protein turnover machinery can cause occurrence of non-native interactions that may lead to pathological disorders such as Alzheimer’s disease. The regulation of protein-protein interaction is mediated either through control of external conditions (such as pH and ionic strength) or by the activity of other cellular proteins (example enzymes). An important feature of protein-protein interaction is the variety in their interaction modes. The types of pits, grooves, voids and pockets that can possibly be generated by the arrangement of amino acid side chains are extremely diverse. Current approach for rational drug design involves the targeting the active sites in a protein which leads to broad spectrum of effects. Moreover the targeting of enzyme active sites by this approach is under effective as the mutations in active site coupled with natural selection is able to overcome such inhibition (for example HIV protease). On the other hand targeting of protein-protein interaction interfaces can be more effective because the system needs to alter both interacting surfaces to overcome such inhibition. Some protein-protein interactions are tissue specific and targeting these is potentially more beneficial than targeting the active sites of enzyme.

The potential problem in any such approach is the versatility in protein–protein interactions. Proteins may interact in extremely diverse range of concentrations. More over the cellular process have an inbuilt redundancy and alternative pathways generally can compensate the inhibition. In addition the knowledge about any starting compound for the inhibitor design is not straight forward as unlike the enzymes there is no small molecular substrate.

1.4 Protein-protein contacts: Composition and nature of interactions

Protein-protein interactions typically bury 1600Å^2 of the surface area at the interface (Buckingham, 2004). The interface is potentially rich in arginine, histidine, asparagine, tryptophan, tyrosine and serine (Davies, D.R. et al, 1996). Analysis of secondary structures in the interface areas showed that the random coil comprises 47% of the protein-protein interaction interface; 36% α-helix; 17% β-sheet (Nissinov, R 1997). The interaction forces are van der Waals, hydrophobic and electrostatic in nature. The degree of surface complementarity between interacting interfaces is dependent on the strength of complex. Permanent complexes interfaces have a high surface complementarity whereas temporary complexes have less interfacial complementarity. (Jones S *et al* 1996).

1.5 Thermodynamics and kinetics of protein-ligand interactions

Protein-ligand interactions can be experimentally measured under thermodynamic equilibrium conditions from which the inhibition constant K_i can be obtained (Equation 1.2). The inhibition (or dissociation) constant describes the strength of protein-ligand binding as mole/l. A ligand binds stronger to the receptor when the K_i is small (e.g. nanomolar). If there is less ligand present than the value of $K_{i,}$ then only a small proportion of the protein will be associated with the ligand and a biological effect may be difficult to measure. IC_{50} term gives the ligand concentration at which the enzyme activity decreases to 50%. It is shown that both IC_{50} and K_i characterise protein-ligand interactions in a similar way, so that the easily measurable IC_{50} values can be used to compare ligands with each other (Gohlke and Klebe, 2002).

The binding process is driven by the standard Gibb's free energy of binding ΔG° which is related to K_i (Equation 1.3). At 25˚C, a K_i of 1 nM would be equivalent to -12.2 kcal/mole. Changing the K_i by one order of magnitude will shift ΔG° by -1.4 kcal/mole. Inhibition constants usually take values between 10^{-2} and 10^{-12} M, which are equivalent to -2.4- to -16.7 kcal/mole at 25˚C (Boehm and Klebe, 1996). The binding energy ΔG° comprises enthalpic

(ΔH°) and entropic contributions (TΔS°) which can be measured experimentally by Isothermal Titration Calorimetry (ITC) or van't Hoff analysis (Holdgate and Ward, 2005). These experiments have shown that ΔG° and ΔH° are not directly correlated, thus enthalpy alone is not an adequate measure for binding affinity (Boehm and Klebe, 1996). Receptor [R] and ligand [L] associate and form a non-covalent, reversible receptor-ligand complex [LR] in solution under thermodynamic equilibrium conditions.

$$[R]+[L]\leftrightarrow[RL] \tag{1.1}$$

The experimentally determined inhibition constant (K_i) or dissociation constant (K_D) or reciprocal association constant (K_A) describes the relationship between bound and unbound molecules.

$$K_i = K_D = \frac{1}{K_A} = \frac{[R][L]}{[RL]} \tag{1.2}$$

The Gibb's free energy of binding (ΔG°) comprises an enthalpic (ΔH°) and an entropic term (TΔS°) where T is the temperature in Kelvin and R is the gas constant (1.987 cal /(K mole)).

$$\Delta G^\circ = -RT \ln K_A = RT \ln K_i = \Delta H^\circ - T\Delta S^\circ \tag{1.3}$$

1.6 Molecular mechanics-based scoring functions

Computational methods such as docking are applied to identify the correct orientation of the ligand in the binding site and estimate ligand binding affinities. These docking protocols comprise of an algorithm for searching the conformational space to identify the most probable orientation of a molecule in the binding pocket and a scoring function which is used to quantify the strength of interaction a molecule can make in a particular orientation. The aim of a scoring function is to correctly predict the experimental binding free energy in addition to predicting the most probable conformation of the ligand in concurrence with the crystallographic orientation. Most scoring functions report scores in arbitrary units, some scoring functions were particularly designed to estimate the Gibbs free energy changes of binding and are reportedly able to predict these within 1.7-2.4 kcal/mole (Bissantz *et al.*, 2000). Scoring functions can be classified in three main categories: (1) molecular mechanics or force field methods (e.g. AutoDock, GoldScore), (2) empirical free energy or regression-based functions (e.g. ChemScore, X-Score) or (3) knowledge-based potentials (e.g. PMF, DrugScore). These different types of scoring functions will be reviewed in the following sections.

1.7 Molecular mechanics-based scoring function

Molecular Mechanics (MM)-based scoring functions (also termed force field or first principle based methods) approximate binding affinity by summing individual contributions in a master equation. The terms used for different interaction types are based on physicochemical theory and should not be cross correlated with each other. These terms are often combined with solvation and entropic terms.

An example in terms of docking is the original DOCK 3.0 score (Meng *et al.*, 1992). It is one of the earliest scoring functions and covers the principal contributions to binding: shape and electrostatics accounted for in terms of a van der Waals term and an electrostatic potential term. These separable terms are combined into a grid-based AMBER force-field scoring function which is computed at specific grid points according to the field generated by the receptor. The overall score is then calculated as the sum of ligand atom interactions at the grid points (using a interpolation scheme) assuming additivity of individual terms (Tame, 1999).

In contrast to time-consuming quantum mechanics methods, that describe molecules based on their electron distribution by ab-initio or semi-empirical approaches, force fields or molecular mechanics describe molecules reduced to their atoms and bonds i.e. as charged atom centres, with masses assigned according to atomic weight connected by springs. They usually comprise two energy components, one for the protein-ligand interaction and another for the internal (conformational/strain) energy of the ligand (and sometimes the protein). The protein conformational energy is often left out as usually only a single conformation is considered during docking. MM-based scoring methods most often assume a common functional form; however they derive the parameters in slightly different ways. For example, the CHARMM force field uses an empirical energy function to describe the forces on atoms in a molecule and the molecule's potential energy. This function is the sum of many individual energy terms and comprises bonded and pairwise non-bonded interaction terms listed in Equation 1.4 (Brooks *et al.*, 1983).

$$\text{Potential Energy} = \underbrace{E_{bond} + E_{angle} + E_{dihedral}}_{\text{(bonded)}} + \underbrace{E_{elec} + E_{vdw}}_{\text{(non-bonded)}} \qquad \mathbf{1.4}$$

The total energy of a conformation comprises several energy terms (Brooks *et al.*, 1983).

1.7.1 Bonded energy terms

The bonded energy terms comprise the bond (E_{bond}), bond angle (E_{angle}), dihedral ($E_{dihedral}$) and improper torsional potentials (E_{impr}), all together referred to as the bonded interactions (Equation 1.5). The bond and angle deformations (E_{bond}, E_{angle}) are generally small. As such, deviations from equilibrium bond and angle values are treated with large energy penalties. The dihedral angle is defined by four atoms, with the torsion angle about the axis of the middle pair of atoms. The improper torsion potential is necessary to maintain chirality.

$$\mathbf{E_{bonded}} = \underbrace{\sum k_b(r - r_0)^2}_{\text{bond}} + \underbrace{\sum k_\theta(\theta - \theta_0)^2}_{\text{angle}} + \underbrace{\sum |k_\varphi| - k_\varphi\cos(n\varphi)}_{\text{dihedral}} \qquad 1.5$$

Internal energy terms $k_{b,}$ $k_{\theta,}$ k_φ are constants, r =bond length between two atoms (A, B), θ = bond angle between three atoms (A, B, C), φ = torsion angle between two planes defined by four atoms (A, B, C and B, C, D), n = number of least points at 360˚ rotation of B-C bond, $r_{0,}$ θ_0 are the equilibrium values of these variables.

1.7.2 Non-bonded energy term

Van der Waals energy (E_{elec})

The van der Waals energy calculation is calculated by the Lennard-Jones potential energy function, an approximation also called the “6-12 potential”, where the attractive force is treated as being proportional to $1/r^6$ and the repulsive force as being proportional to $1/r^{12}$ (where r is the distance between two atoms).

Electrostatic energy (E_{vdw})

The electrostatic energy calculation is based on partial atomic charges. It can be calculated by applying Coulombs law. Setting the dielectric constant (ε) proportional to r is a standard procedure to mimic electrostatic shielding by solvent when it is not included explicitly (the calculation of additional solvent is CPU intensive). In the presence of solvent, a dielectric constant of 1 is used (i.e. the relative permittivity of free space). The experimentally derived dielectric constant is a bulk solvent property and depends on the polarisability of solvent molecules. It increases with highly polarisable solvents like water (ε =80), reducing greatly the electrostatic interaction. In protein simulations without explicit solvent it usually takes as a value between 2 and 10, or 4r (known as a distance dependent dielectric).

The calculation of the non-bonded energy terms (Equation 1.6) takes up the majority of computing time for energy evaluation because it is proportional to n^2 and not n, as for other terms in Equation 1.6. It can be decreased by using a non-bonded cut-off radius at which the energy becomes zero. In this case, only atom pairs within the cut-off contribute to the calculated interaction energy. A switching function near the cut off distance is used to avoid discontinuity in the energy function and possible instability of the calculated energy.

$$\mathbf{E}_{\textbf{non-bonded}} = \underbrace{\sum_{excl(i,j)=1} \frac{q_i q_j}{4\pi\varepsilon_r r_{ij}}}_{\text{Electrostatic}} + \underbrace{\sum_{excl(i,j)=1} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right)}_{\text{vdW}} \qquad 1.6$$

Non-bonded energy terms. q_i, q_j = point charges of a non-bonded atom pair, ε_r = distance dependent dielectric constant, r_{ij} = distance between atom pair ij, A, B = adjustable Van der Waals repulsion and attraction parameters for atom pairs ij.

1.8 Empirical scoring functions

The second type of scoring functions are developed more specifically for protein-ligand docking by fitting experimental binding affinities using a training set of protein-ligand complexes and are thus dependent on their training set. The free energy of binding is approximated by summing up individual energy terms, which are often simpler but related to molecular mechanics energy terms. Weights or coefficients for each term are derived by regression analysis. Different functions implement various types of energy terms and can include entropic and desolvation terms (albeit these are still approximations). ChemScore (Eldridge *et al.*, 1997) is given as an example in Equation 1.7. It comprises four simple terms: two contact terms for lipophilic and metal interactions, a hydrogen bonding and a penalty term depending on the number of rotatable bonds. The weights were derived by regression based on a training set of 82 protein-ligand complexes with known binding affinity and their robustness assessed by cross validation. The design concept involved reduction of the total number of terms and exclusion of those that showed inter-correlation. In addition, all terms and coefficients should be physics based and interpretable. The scoring function was later applied to *de novo* designed compounds that were synthesised and tested (Murray *et al.*, 1998). The scoring function was found to be valuable, however, it overestimated binding affinity in several cases and subtle changes between close analogues were not predicted with accuracy.

$$\Delta G_{bind} = \Delta G_{H-bond} \sum_{H-bond} f(\Delta R, \Delta\alpha) + \Delta G_{metal} \sum_{metal} f(\Delta R, \Delta\alpha) + \Delta G_{lipo} \sum_{lipo} f(\Delta R) + \Delta G_{rotor} \sum_{rotor} f(P_{nl}, P'_{nl}) + \Delta G_0 \qquad 1.7$$

Free energy of binding (ΔG_{bind}) for ChemScore H-bond = hydrogen bonding, metal = metal interaction, lipo = lipophilic, rotor = rotational entropy, ΔR = distance term, Δα = angular term, ΔG_0 = regression constant, ΔG = regression coefficients for each term, P_{nl} = penalty (dependent on number rotatable bonds and their environment).

1.9 Knowledge-based scoring functions

Knowledge-based scoring functions are derived by statistical analysis of the frequency distributions within a set of protein-ligand structures from which pairwise atomic interaction potentials are deduced. As such they reproduce observed preferences of functional group binding i.e. experimental structures rather than binding affinities. Like empirical scoring functions, these functions try to overcome the problem of insufficient description of a complex binding event due to the lack of explicit parameters. Well known examples are PMF (Muegge and Martin, 1999) and DrugScore (Gohlke and Klebe, 2001), and their generic functional form is outlined in Equation 1.8. With the increase in available crystal structures (and therefore knowledge) these scoring functions are expected to further improve in the future. The scoring functions differ in respect to their chosen reference distribution (g_{ref}), an important term influencing the distance-dependent pair potentials. PMF sets the cut off at 12Å for sampling atom pair contacts but DrugScore at 6 Å (Gohlke and Klebe, 2001). The larger PMF cut off value was chosen to include implicit solvation effects, whereas specific interactions are considered by DrugScore. Additionally, DrugScore incorporates Solvent Accessible Surface singlet potentials. DrugScore correctly identified the best ligand pose in 75% of cases for 160 complexes (Gohlke *et al.*, 2000).

$$\Delta W_{ij}(r) = -\ln \frac{g_{ij}(r)}{g_{ref}} \tag{1.8}$$

Where, $g_{ij}(r)$ =frequency (probability distribution) of atom pair ij separated by a distance r, g_{ref} = reference distribution. $\Delta W_{ij}(r)$ =pair-(pseudo-) potentials of atom pair ij.

1.10 Treatment of divalent ions (such as zinc) in scoring function

Zinc is essential for the catalytic function of metalloenzymes and coordinated in a number of distinct geometries (Alberts *et al.*, 1998). Zinc binding groups in protein-ligand complexes can be classified according to their coordination geometry such as tetrahedral for thiolates and sulfonamides, distorted trigonal bipyramidal for hydroxamates, carboxylates, phosphonates and phosphinates (Hu *et al.*, 2004). Recreating the correct coordination geometry is essential for successful docking (Hu *et al.*, 2004), however modelling of ligand binding to zinc is challenging due to multiple coordination geometries (Figure 1.1), as well as polarisation, charge-transfer and inadequate force fields (Jain & Jayaram, 2007). Zinc can be modelled in a classic energy function by treating it as either bonded (e.g. GOLD) or non-bonded (e.g. DOCK). The first integrates angle and bond terms in the potential function whereas the latter simply treats it with electrostatic and vdW terms.

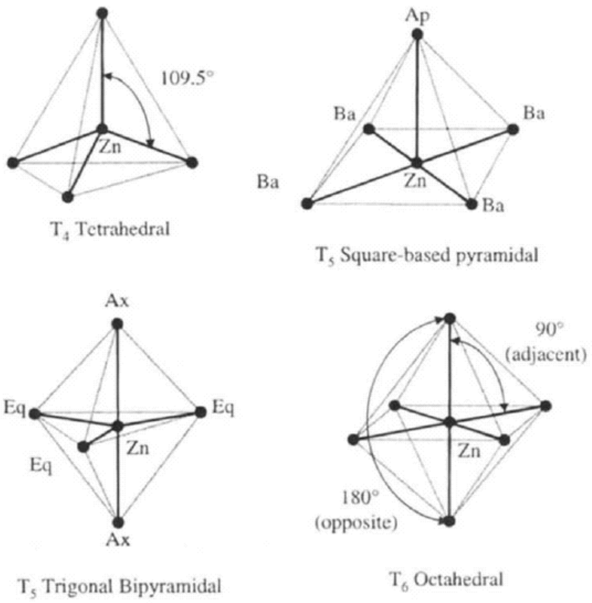


Figure 1.1 Zinc coordination geometries in protein-ligand complexes (Alberts *et al.*, 1998).

1.11 Virtual screening

A widely used application of both structure and ligand based design methods is virtual screening, where large compound libraries are screened *in silico* as opposed to experimental high-throughput screening (HTS) where compounds are screened against a target using a bioassay. Experimental HTS is the standard technique used in the pharmaceutical industry for lead discovery, but a costly approach due to its random nature and expense in screening large numbers of compounds. In virtual screening, structural descriptors are used as filters to retrieve active compounds that can provide new leads. Many different virtual screening methodologies exist, taking into account ligand or protein information ranging from 1D (e.g. molecular weight) to 2D (e.g. topology or substructure) and 3D (e.g. shape similarity, 3D pharmacophore or protein structure) properties. Ligand-based VS approaches have recently been reviewed by Eckert and Bajorath (2007). Pharmacophores represent key interactions between ligand and proteins and can be ligand and/or protein based. The general concept and different pharmacophore generation methods were recently described by Khedkar *et al.* (2007). We will return to more specific detail of the methods used (Chapter 2) and their application to ACE2 (Chapter 5) later.

The most common approach for structure-based virtual screening, however, is by protein-ligand docking. Success in virtual screening is judged on enrichment - i.e. the retrieval of known actives from a set of inactives. The enrichment factor is calculated as $(A_h/T_h)/(A/T)$, where A_h is the number of active compounds found in a selected subset of the ranked database (T_h), A is the total number of actives and T is the total database size. Many validation studies have been undertaken, comparing ligand and protein structure-based methods with docking for their effectiveness in VS (Chen *et al.*, 2006, Hawkins *et al.*, 2007, McGaughey *et al.*, 2007) and a plethora of comparative docking (enrichment) studies exist (Perola *et al.*, 2004, Chen *et al.*, 2006, Zhou *et al.*, 2007). The performance of specific docking tools is usually dependent on the target involved but also on the preparation of the compound database (Knox *et al.*, 2005). Comparison of different docking programmes is difficult due to non-standardised parameter settings/ligand and protein preparation (Cole *et al.*, 2005). Independent investigators can arrive at conflicting results related to docking success for individual programmes as recently discussed by Chen *et al* (2006). Different implementation of a scoring function can also lead to different results (Wang *et al.*, 2004). In conclusion, there is not a single docking programme that outperforms others in all circumstances.

1.12 Docking and scoring

In docking, a ligand is first placed into the binding site of a protein in various different orientations and conformations (conformational search stage) and each conformation (or pose) is scored by evaluation of the ligand-protein interactions according to a predefined scoring function. The highest ranking pose is assumed to resemble the "correct" binding mode and sometimes an estimate is also made of a ligand’s binding affinity. Docking algorithms can be classified according to their search methodology and the way they treat ligand flexibility. Systematic methods investigate all degrees of freedoms and often use incremental construction to build up ligands in a stepwise manner and use pruning methods to cope with the combinatorial explosion problem. FlexX (Rarey *et al.*, 1996) or DOCK 4.0 (Ewing *et al.*, 2001) are examples of these. Alternatively, in methods such as FLOG (Miller *et al.*, 1994) conformations can be pre-generated and then docked rigidly to the protein receptor. Lastly, stochastic approaches randomly change conformations of a single ligand or whole ligand populations in the receptor binding site. These include Monte Carlo simulations e.g. QXP (Bohacek and McMartin, 1997), genetic algorithms e.g. GOLD (Jones *et al.*, 1997) or AutoDock 3 & 4 (Morris *et al.*, 1998), Monte Carlo simulated annealing e.g. AutoDock 1 (Goodsell and Olson, 1990) or Tabu search e.g. PRO_LEADS (Baxter *et al.*, 1998). Simulation methods such as molecular dynamics or energy minimisation have also been applied alone or alongside other search methods (Brooijmans and Kuntz, 2003, Sousa *et al.*, 2006). Conformational sampling prior to or during docking and the ability to regenerate the bioactive ligand conformation is an essential part of both ligand- and structure-based approaches and has been analysed in a number of studies (Good and Cheney, 2003, Perola and Charifson, 2004, Kirchmair *et al.*, 2005).

Docking is primarily used as a VS tool to identify promising bioactive compounds (or hits), but can also be used later for lead optimisation. In both cases, the docking programme needs to first sample and recognise the bioactive conformation of each ligand and secondly reliably rank the ligands according to their predicted binding affinities. It is the scoring functions, responsible for prioritising compounds, which are the major weakness of current docking programmes rather than the conformational sampling methods (Warren *et al.*, 2006). Of the many scoring functions that have been developed to address this issue, so far, none have consistently proven superior for all protein targets (Wang *et al.*, 2003, Wang *et al.*, 2004). Target dependency is a general issue in docking and scoring as ligand binding can be either dominated by enthalpic or entropic contributions which need to be captured in the scoring functions, however, the latter effect is poorly treated or neglected completely. Consensus scoring is often applied and was shown to reduce the number of false positives in VS (Charifson *et al.*, 1999). In contrast, Wang et al. (2004) found that a number of scoring functions were more correlated to each other than to experimental binding affinities, but that consensus scoring improved the determination of the correct binding mode. This finding was supported by Yang *et al.* (2005) who concluded that consensus scoring enhanced enrichment if individual scoring functions performed well and were distinct.

1.13 Predicting Ligand binding sites

The function of a protein is dependent on the nature of molecules it can interact with. Even though the number of known structures of proteins has grown rapidly in the recent years (Tagari, Tate et al. 2006) a large number of protein-ligand interaction sites remain uncharacterised (Laurie and Jackson 2005) . A number of approaches have been adopted to make predictions about the function of a protein from its structure (Laskowski, Watson et al. 2005). Some methods look for secondary structural arrangement patters such as motifs or domains associated with specific functions (Laskowski, Watson et al. 2005), others tend to look for characteristic arrangement of functionally important or conserved residues (Burgoyne and Jackson 2006). The function of a protein depends upon the nature of ligand it can interact with, hence identification of the ligand binding sites and assignment of the nature of the putative ligand that can interact is important for the prediction of function to the protein structure as well as for rational structure-based drug design.

Carbohydrate binding proteins play an important role in cellular systems. Carbohydrate binding is involved in energy metabolism, intercellular communication and adhesion (Brandley and Schnaar 1986). Ligand binding sites are very diverse in structure and function (Bertozzi and Kiessling 2001). Only a few of them are druggable. Carbohydrate binding sites are increasingly being considered as putative drug targets (Bertozzi and Kiessling 2001) because of their role in intra and inter-cellular communication. Carbohydrate binding sites have been extensively studied (Weis and Drickamer 1996) in the past. However, only a few approaches developed for the prediction of carbohydrate binding sites (Taroni, Jones et al. 2000), (Shionyu-Mitsuyama, Shirai et al. 2003) and (Malik and Ahmad 2007). But these methods have not been very successful.

In the third chapter of the thesis development of a new computational method for predicting carbohydrate binding sites is presented. The overall aim was to develop a new computational method for predicting carbohydrate binding sites with high accuracy. The method differs from the previous carbohydrate binding site prediction methods in two important aspects. Firstly it uses 375 non-covalent protein-carbohydrate complexes for the derivation of amino acid propensity scores. This is more than used in calculation of amino acid propensities in the previous methods. Secondly it uses a two-step procedure to identify sites. In step one; it uses a grid-based approach to identify sites on the protein with a high probability of being a binding site, using the recently proposed method of Laurie and Jackson, 2005. In step two; it uses these sites and amino acid propensity scores to predict the location of carbohydrate binding sites. The ultimate aim of the project was to produce a method that could both locate likely binding sites and then distinguish the nature of the binding site, to ascertain if the site has the ability to preferentially bind a carbohydrate ligand.

1.14 Predicting protein-ligand affinity

The success of *in silico* approaches for SBDD depend on the application of the principles governing the dynamics of ligand-protein interactions (Rauh, Klebe et al. 2004). The current approach of docking involves generating favourable ligand orientations in the protein binding site, by sampling conformational space, followed by scoring these by their predicted interaction energy (Klebe 2006). The limitation in the scoring step stems from the time needed to score each potential solution and the level of accuracy required for the calculation of the interaction energy, or at the very least, the correct discrimination of active from inactive compounds. A number of simplified scoring functions have been developed which are fast and easy to apply but provide only moderate levels of accuracy. Hence continued efforts are needed to improve upon existing scoring functions.

Current, scoring functions used to estimate ligand-protein affinity can be classified into three categories: first-principle methods, knowledge-based methods and finally, regression-based scoring functions (Zentgraf, Steuber et al. 2007). Knowledge-based scoring functions are derived from the quantification of frequencies of interacting atomic pairs observed in protein-ligand complexes (Gohlke and Klebe 2001). The process of atomic-pair-interaction-frequency quantification has been based on a number of mathematical relationships. The earliest example of such a function was in the field of protein folding where Boltzmann's law was used to derive the potential of mean force for interacting residue (Tanaka and Scheraga 1976; Hendlich, Lackner et al. 1990; Sippl 1990). Later, similar functions were developed for scoring ligand-protein interactions. Wallqvist et al. (Wallqvist, Jernigan et al. 1995) studied a dataset of 38 complexes, calculating the frequencies of atomic interactions at the protein-protein interface and converted these into an atom-atom preference score using the ratio of fraction of the total interface area contributed by each pair to the product of the fraction of their respective contributions to the surface of respective protein. For a set of 30 proteases-inhibitor complexes, Verkhivker et al. (Verkhivker, Appelt et al. 1995) used the inverse Boltzmann law to develop distance-dependent pair potentials from interacting atoms in combination with conformational entropic (Pickett and Sternberg 1993) and hydrophobic (Sharp, Nicholls et al. 1991) terms. Using this scoring function they could estimate the affinity of HIV-1 proteases for several different inhibitors. SMoG-Score was developed from 109 crystal structures using statistical mechanics (DeWitte and Shakhnovich 1996). Potentials of mean force were derived by Muegge et al. using the inverse Boltzmann law by converting the distance dependent number density of interacting atom pairs from a dataset of

697 protein-ligand complexes into their respective Helmholtz interaction free energies (Muegge and Martin 1999; Muegge, Martin et al. 1999). Mitchell et al. developed BLEEP using a dataset of 820 protein-ligand complexes with hydrogen atoms added (using HBPlus (McDonald and Thornton 1994)) and used the inverse Boltzmann law (Mitchell et al. 1999). A semi-empirical pair-potential for Ne-Ne was used as a reference state. They further derived BLEEP-II by including interactions of protein and ligand with water molecules (explicitly added using Aquarius2 (Pitt and Goodfellow 1991)). Gohlke et al (Gohlke, Hendlich et al. 2000) derived DrugScore using distance-dependent pair-potentials from a dataset of 6026 protein-ligand complexes and incorporated solvent accessible surface area based solvation potentials from a database of 1376 protein-ligand complexes. Cline et al (Cline, Karplus et al. 2002) used an information theoretic relationship of mutual information to quantify information in amino-acid contact potentials for protein structure prediction. They studied the contribution of amino-acid character in terms of hydropathy, charge, disulphide bonding and residue burial to the mutual information.

The Boltzmann law is very useful for determining the interaction energy values from a database of the observed frequencies of joint occurrences. The variation in temperature factors for the protein-ligand atoms (Finkelstein, Gutin et al. 1995) give rise to heterogeneity in the interaction database which complicates the application of the inverse Boltzmann law. However, even though knowledge-based methods are susceptible to the artefacts in data collection they have performed surprisingly well, often better than force-field based scoring functions (Sternberg, Bates et al. 1999; Wang, Lu et al. 2004).

In the fourth chapter of the thesis the development of a novel knowledge-based scoring function: ScoreJE - derived from the ligand-protein interacting atomic pairs is presented. Our approach differs from the previous scoring functions in two important aspects. Firstly, it uses over 3,000 structurally non-redundant protein-ligand complexes. This is more complexes than used in constructing previous knowledge-based scoring functions, the only exception being DrugScore, which uses a 30% sequence identity cut-off for the creation of the protein non-redundant dataset. Secondly in using the mathematical relationship of joint entropy for deriving the atomic contact preferences it bypasses the problems implicit in the application of the inverse Boltzmann law, eliminating the need for a reference state. These preferences are derived for describing the energetics of short-range atomic interactions. A Single-body Solvation Potential (SSP) is developed using the joint entropy of protein-water atom contact probabilities and is combined with ScoreJE to obtain SIScoreJE (SSP included ScoreJE).

These functions were tested for their ability to predict the binding energies of test datasets containing 100 protein-ligand complexes.

The overall aim was to develop a novel knowledge-based scoring function for predicting protein-ligand interaction energy. The main objective was to calculate a non-redundant set of atomic contact preferences for the protein-ligand and protein-water interactions and to use these to develop a scoring function using information theory. A secondary aim was to evaluate the potential of using information theory and new atom type classification schemes (alongside popular atom-type classification schemes currently in use) to optimally describe protein-ligand interactions.

1.15 Aims and objectives

GGTase-II is important enzyme in the membrane trafficking regulation system. GGTase-II prenylates the small GTPases from Rab family by transferring geranylgeranyl (a 20 carbon atom lipid molecule) from its pyrophosphate form to the C-terminal cysteine residues. This covalent modification allows RabGTPases to localise on the membranes where “Guanine nucleotide exchange factors” interacts and induces the exchange the GDP from Rab-GDP complex by GTP. GTP bound Rabs interact with a plethora of effector protein molecules and mediate vesicular transport. In metastasis cancer protease enzymes are released by exocytosis for dissolution of collagen matrix so that the metastatic ells can invade other tissues. Inhibition of the Rab prenylation reaction could result in shut down the Rab mediated vesicular trafficking hence GGTase-II is an important, target for cancer therapeutics. Our objective was to disrupt REP:GGTase-II interaction.

Even though a number of programs for ligand binding site identification are available the existing methods do not specifically identify carbohydrate or drug-like compound binding sites. A new approach was our objective for the assignment of the character to the ligand binding sites.

Another aspect of computational tools that need improvement is the estimation protocols for fast estimating the binding affinity between the ligand and its cognate protein receptor. Our objective in this regard was to use the information theoretic relationships to create the scoring function for estimation of the binding affinity. The information theoretic approach was considered better than the existing ones like inverse Boltzmann’ s law as there was no assumption in our model.

1.16 Thesis outline

The remainder of this thesis is structured with in chapters, three results chapters and a general conclusions chapter. The results chapters include a chapter on structure based drug design and two methods development chapters. The first results chapter (chapter 2) presents the development of REP-GGTase-II interaction inhibitor. Chapter 3 describes a development of a tool for the identification of ligand binding sites and determination of the nature of the ligand that shall bind the predicted site. Chapter 4 presents the development of information theory based novel scoring function for the estimating the binding affinity between the ligand and its cognate receptor. Finally, general conclusions are drawn regarding this work in chapter 5.

Abbenante, G. and Fairlie, D. P. (2005). "Protease inhibitors in the clinic." Med Chem 1(1): 71-104.
Accelrys "Catalyst 4.9." 9685 Scranton Road, San Diego, USA. http://www.accelrys.com.

Bendtsen, J. D., Nielsen, H., von Heijne, G. and Brunak, S. (2004). "Improved prediction of signal peptides: SignalP 3.0." J Mol Biol 340(4): 783-95.

Bissantz, C., Folkers, G. and Rognan, D. (2000). "Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations." J Med Chem 43(25): 4759-67.

Boehm, H. J. and Klebe, G. (1996). "Was läßt sich aus der molekularen Erkennung in Protein-Ligand-Komplexen für das Design neuer Wirkstoffe lernen?" Angew. Chem 108(22): 2750 - 2778.

Bohacek, R. S. and McMartin, C. (1997). "Modern computational chemistry and drug discovery: structure generating programs." Curr. Opin. Chem. Biol. 1: 157-161.

Brooijmans, N. and Kuntz, I. D. (2003). "Molecular recognition and docking algorithms." Annu Rev Biophys Biomol Struct 32: 335-73.

Charifson, P. S., Corkery, J. J., Murcko, M. A. and Walters, W. P. (1999). "Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins." J Med Chem 42(25): 5100-9.

Chen, H., Lyne, P. D., Giordanetto, F., Lovell, T. and Li, J. (2006). "On evaluating molecular-docking methods for pose prediction and enrichment factors." J Chem Inf Model 46(1): 401-15.

Cole, J. C., Murray, C. W., Nissink, J. W., Taylor, R. D. and Taylor, R. (2005). "Comparing protein-ligand docking programs is difficult." Proteins 60(3): 325-32.

Eckert, H. and Bajorath, J. (2007). "Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches." Drug Discov Today 12(5-6): 225-33.
Ehrlich, P. (1909). Dtsch Chem Ges 42: 17.

Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V. and Mee, R. P. (1997). "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 11(5): 425-45.

Fraternali, F. and Van Gunsteren, W. F. (1996). "An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution." J Mol Biol 256(5): 939-48.

Gastreich, M., Lilienthal, M., Briem, H. and Claussen, H. (2006). "Ultrafast de novo docking combining pharmacophores and combinatorics." J Comput Aided Mol Des 20(12): 717-34.

Gohlke, H. and Klebe, G. (2002). "Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors." Angew Chem Int Ed Engl 41(15): 2644-76.

Gohlke, H., Hendlich, M. and Klebe, G. (2000). "Knowledge-based scoring function to predict protein-ligand interactions." J Mol Biol 295(2): 337-56.

Goldberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. Reading, Mass., London, Addison-Wesley.

Good, A. C. and Cheney, D. L. (2003). "Analysis and optimization of structure-based virtual screening protocols (1): exploration of ligand conformational sampling techniques." J Mol Graph Model 22(1): 23-30.

Goodsell, D. S. and Olson, A. J. (1990). "Automated docking of substrates to proteins by simulated annealing." Proteins 8(3): 195-202.

Holdgate, G. A. and Ward, W. H. (2005). "Measurements of binding thermodynamics in drug discovery." Drug Discov Today 10(22): 1543-50.

Jones, G., Willett, P. and Glen, R. C. (1995). "Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation." J Mol Biol 245(1): 43-53.

Jones, G., Willett, P., Glen, R. C., Leach, A. R. and Taylor, R. (1997). "Development and validation of a genetic algorithm for flexible docking." J Mol Biol 267(3): 727-48.

Khedkar, S. A., Malde, A. K., Coutinho, E. C. and Srivastava, S. (2007). "Pharmacophore modeling in drug discovery and development: an overview." Med Chem 3(2): 187-97.

Kirchmair, J., Laggner, C., Wolber, G. and Langer, T. (2005). "Comparative analysis of protein-bound ligand conformations with respect to catalyst's conformational space subsampling algorithms." J Chem Inf Model 45(2): 422-30.

Kirchmair, J., Wolber, G., Laggner, C. and Langer, T. (2006). "Comparative performance assessment of the conformational model generators omega and catalyst: a large-scale survey on the retrieval of protein-bound ligand conformations." J Chem Inf Model 46(4): 1848-61.

Kitchen, D. B., Decornez, H., Furr, J. R. and Bajorath, J. (2004). "Docking and scoring in virtual screening for drug discovery: methods and applications." Nat Rev Drug Discov 3(11): 935-49.

Klebe, G. (2006). "Virtual ligand screening: strategies, perspectives and limitations." Drug Discov Today 11(13-14): 580-94.

Kramer, J. A., Sagartz, J. E. and Morris, D. L. (2007). "The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates." Nat Rev Drug Discov.

Kroemer, R. T., Vulpetti, A., McDonald, J. J., Rohrer, D. C., Trosset, J. Y., Giordanetto, F., Cotesta, S., McMartin, C., Kihlen, M. and Stouten, P. F. (2004). "Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations." J Chem Inf Comput Sci 44(3): 871-81.

Kurogi, Y. and Guner, O. F. (2001). "Pharmacophore modeling and three-dimensional database searching for drug design using catalyst." Curr Med Chem 8(9): 1035-55.

Kyte, J. and Doolittle, R. F. (1982). "A simple method for displaying the hydropathic character of a protein." J Mol Biol 157(1): 105-32.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. a. and Thornton, J. M. (1993). " PROCHECK - a program to check the stereochemical quality of protein structures." J. Appl. Cryst. 26: 283.

Meng, E. C., Shoichet, B. K. and Kuntz, I. D. (1992). " Automated docking with grid-based energy evaluation." J. Comp. Chem. 13: 505-524.

Muegge, I. and Martin, Y. C. (1999). "A general and fast scoring function for protein-ligand interactions: a simplified potential approach." J Med Chem 42(5): 791-804.

Murray, C. W., Auton, T. R. and Eldridge, M. D. (1998). "Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model." J Comput Aided Mol Des 12(5): 503-19.

Rella, M., Elliot, J. L., Ballard, S., Lanfear, J., Phelan, A., Jackson, R. M., Turner, A. J. and Hooper, N. M. (2007). "Identification and characterisation of the angiotensin converting enzyme-3 (ACE3) gene: a novel mammalian homologue of ACE." BMC Genomics 8 (1): 194.

Rella, M., Rushworth, C. A., Guy, J. L., Turner, A. J., Langer, T. and Jackson, R. M. (2006). "Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors." J Chem Inf Model 46(2): 708-16.

Rousseau, A., Michaud, A., Chauvet, M. T., Lenfant, M. and Corvol, P. (1995). "The hemoregulatory peptide N-acetyl-Ser-Asp-Lys-Pro is a natural and specific substrate of the N-terminal active site of human angiotensin-converting enzyme." J Biol Chem 270(8): 3656-61.

Rychlewski, L. and Fischer, D. (2005). "LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction." Protein Sci 14(1): 240-5.

Sadowski, J. and Schwab, C. H. (2006). "3D structure generator CORINA 3.4: Generation of high quality three-dimensional molecular models (program manual). http://www.molecular-networks.com/software/corina."

Sansom, C. E., Hoang, M. V. and Turner, A. J. (1998). "Molecular modelling and site-directed mutagenesis of the active site of endothelin-converting enzyme." Protein Eng 11(12): 1235-41.

Schneider, G. and Fechner, U. (2005). "Computer-based de novo design of drug-like molecules." Nat Rev Drug Discov 4(8): 649-63.

Schwab, C. H. and Gasteiger, J. (2002). "ROTATE 1.1: Conformer generator for acyclic molecules and fragments (program manual). http://www.molecular-networks.com/software/rotate."

Schwede, T., Kopp, J., Guex, N. and Peitsch, M. C. (2003). "SWISS-MODEL: An automated protein homology-modeling server." Nucleic Acids Res 31(13): 3381-5.

Wermuth, C. G., Ganellin, C. R., Lindberg, P. and Mitscher, L. A. (1998). "Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998)." Pure Appl. Chem. 70: 1129-1143.

Zhang, Z. and Gerstein, M. (2004). "Large-scale analysis of pseudogenes in the human genome." Curr Opin Genet Dev 14(4): 328-35.

Zhang, Z., Carriero, N. and Gerstein, M. (2004). "Comparative analysis of processed pseudogenes in the mouse and human genomes." Trends Genet 20(2): 62-7.

Zhou, Z., Felts, A. K., Friesner, R. A. and Levy, R. M. (2007). "Comparative Performance of Several Flexible Docking Programs and Scoring Functions: Enrichment Studies for a Diverse Set of Pharmaceutically Relevant Targets." J Chem Inf Model.

Chapter 2: Structure-based pharmacophore design and targeting REP-GGTase-II interaction interface

2.1 Abstract

A structure-based approach was applied to identify novel inhibitors for inhibiting the GGTase-II (Geranylgeranyltransferase-II) and Rab escort protein (REP) interaction. REP and GGTase-II interaction is bimodal and limited to an area of 650Å^2. Structure-based inhibitor design approaches were used to model molecules for targeting the hydrophobic interactions in REP-GGTase-II interaction interface site. These molecules were screened by docking to the targeted site followed by evaluation using consensus scoring. The virtual molecules thus modelled were used to create pharmacophore hypothesis for virtual screening. Volume exclusion features were added to the ligand derived pharmacophore hypothesis from the structure of targeted site. Using catalyst the ZINC database was screened using the modified pharmacophore hypothesis. The molecules were evaluated by docking and consensus scoring. Out of 27 top hits 9 molecules (which were available) were tested. A novel inhibitor was identified with IC50 values in the range of 7.0. The binding mode of inhibitor molecule and its probable inhibitory mechanism were analysed via retrospective docking.

2.2 Introduction

Protein-protein interactions are fundamental to the functioning of biological systems - from cell division to programmed cell death - and therefore represent a large and important class for human therapeutics (Martin, 1998; Arkin, 2004). Protein-protein interactions can be of obligate and non-obligate nature. Proteins forming non-obligate complexes can fold and exist independently. The formation of these transient, non-obligate protein-protein complexes can be driven by concentration (e.g. Sperm Lysin protein dimmer formation) or covalent modification (e.g. Phosphorylation of cyclins drives its complexation with cyclin dependent kinases) or change in effector molecule structure (e.g. upon GTP hydrolysis in Gα proteins Gβγ bind to it). These transient complexes are important targets for human therapeutics.

However targeting protein-protein interaction interface (PPII) in tricky business. Often the starting point for the inhibitor design is missing. The interaction interfaces comprise of mostly planar surface which is very difficult to target. Very few examples of naturally occurring compounds that target protein-protein interaction interfaces are known. The apparent surface complementarity in PPII involves significant conformational changes making it harder to identify the transient small molecule binding sites. In spite of the difficulties success has been achieved in some cases (Arkin and Wells 2004). One approach for targeting the PPII include mapping of the epitope structure of the interacting proteins on the small peptide surface (Arkin, Randal et al. 2003). Random screening for compounds has also yielded molecules that can target PPII such as certain alkaloids which affect the polymerisation of tubulin (Nooren and Thornton 2003). In the absence of larger libraries of known protein-protein interaction inhibitors the research has remained focused on structure, virtual screening and fragment-based discovery. In this project, strategy of structure based rational inhibitor design was used to target protein-protein interaction interface of geranylgeranyltransferase-II (GGTase-II) and Rab escort protein (REP).

2.3 Biological perspective

Vesicular trafficking is a very tightly controlled process of transport of proteins and membrane components from the site of synthesis/modification to the site of functionality. A number of proteins interact to keep the process tightly regulated. The regulators of vesicle trafficking: select cargo proteins during vesicle assembly, control vesicle formation at donor membrane, direct transport direction, brings about the anchorage of vesicle near the acceptor membrane compartment, initiate and drive the fusion of the vesicle with the acceptor membrane. Any abnormality in the components of vesicle trafficking regulatory machinery leads to pathological state. Rab proteins, which are central regulators of the vesicular trafficking, are known to cause diseases, when defective. Mutations in Rab27a are known to cause type II Griscelli syndrome in humans. Griscelli syndrome is autosomal recessive condition characterized by hypo-pigmentation of skin. People suffering from this disorder also develop haemophagocytic syndrome characterized by uncontrolled T lymphocyte and macrophage activation (Rak, Pylypenko et al. 2004). Over expression of Rab25 is known to occur in cancers of ovary and prostate. Its expression is also upregulated in invasive breast cell tumor, and transitional cell carcinoma. Rab5a and Rab7 are found to be over expressed in thyroid-associated adenomas. Cancerous cells are thought to have increased vesicle trafficking as compared to normal cells as these invasive cells need to secrete proteolytic enzymes to escape the physical barrier of tissue structure. Increased expression of these vesicle trafficking regulators is considered to be the part of overall upregulation of the entire trafficking machinery.

Hence, Rab proteins are lucrative targets for the disruption of the vesicle trafficking. Functionality of Rab proteins is dependent upon its prenylation which is carried out by an enzyme called GGTase-II and is mediated through another protein REP. Targeting the GGTase-II enzyme should halt the vesiclular trafficking as Rab proteins shall not be able to localize on the membrane in the absence of prenyl moiety on it C-terminus tail.

2.4 Understanding the enzyme system of Rab prenylation

2.4.1 Rab proteins

Rab proteins are membrane anchored, small GTPases (of molecular weight 23-26 kDa) that are central to the regulation of vesicular transport. Rab proteins are membrane anchored by virtue of a 20 carbon atom (Figure 2.1), tetra-unsaturated lipid molecule, covalently attached to the cysteine in C-terminus tail via thioester bond. There are over 60 Rab proteins in human genome which exist in GTP and GDP bound states. In GTP bound form, Rab proteins interact and recruit a number of effectors which trigger a chain of events including change in curvature of the membrane of Rab location, packaging of cargo, pinching-off of the vesicle, transport towards a specific target membrane, loose tethering and finally docking of the cargo packed vesicle to the target membrane (Figure 2.2). The hydrolysis of the GTP changes the profile of Rab interacting partners. During the fusion of the vesicle with the acceptor compartment, membrane bound "GTPase activating proteins" (GAPs) interact with RabGTPases and increase the rate of GTP hydrolysis to GDP. After the fusion of vesicle the GDP bound Rab is extracted from the membrane by another protein called GDP dissociation inhibitor (GDI). GDI delivers the Rab to the source membrane where membrane bound GDP exchange factors (called GEFs) catalyse the exchange of GDP for GTP (Itzen, Pylypenko et al. 2006). This triggers the chain of events as outlined above and results in continuous packaging and delivery of cargo proteins and lipids.

The anchorage of Rabs on the membranes is critical to their functioning. Unsaturated, aliphatic geranylgeranyl isoprenoids molecules are post-translationally attached to conserved cysteine residues in the hypervariable C-terminus tail of Rab proteins by GGTase-II (also referred as RabGGTase in literature).

PDF-XChange Click to buy NOW! www.docu-track.com

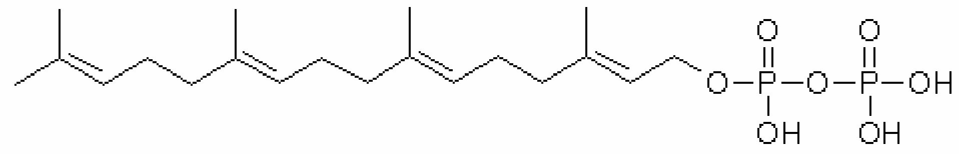


Figure 2.1 Geranylgeranylpyrophosphate.

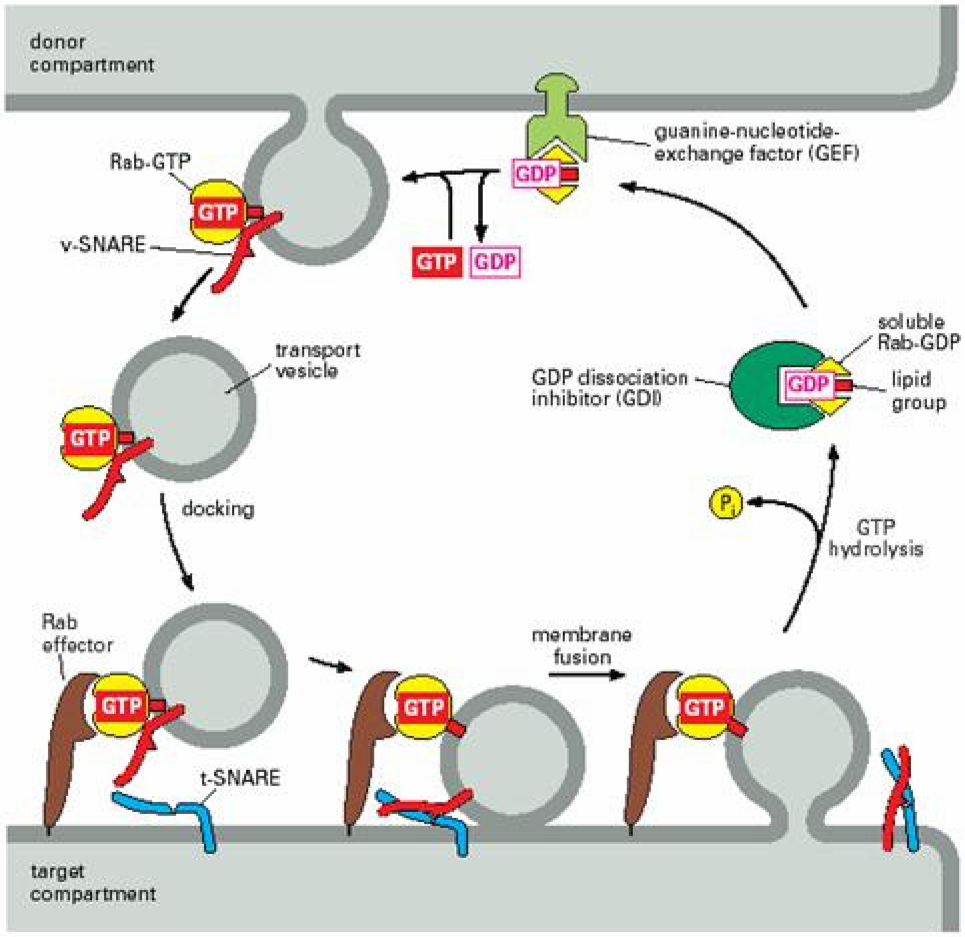


Figure 2.2 The functional cycle of Rab proteins

2.4.2 GGTase-II

GGTase-II is a heterodimer comprising of two subunits (α and β). The molecular mass of α and β subunits is ca. 60 and 40 kDa respectively. It belongs to the family of prenyltransferases. Members of this family include Farnesyltransferase (FTase) and GGTase-I (Figure 2.3a and Figure 2.3b). While FTase transfers 15 carbon-atom, unsaturated hydrocarbon (called farnesyl) to the C-terminus cysteine of RasGTPases/Lamins/transducin-γ subunit, GGTase-I transfers 20 carbon atom, geranylgeranyl moiety on the C-terminus cysteine of Rac/RhoGTPAses/trimericGα. FTase and GGTase-I are functionally similar as they recognise CaaX motif in C-terminus as substrate for prenylation. CaaX stands for prenylatable cysteine residue (C), followed by two aliphatic residues (a) followed by an "enzyme-determining" residue X. The carboxyl-terminal amino acid (X) discriminates FTase targets from those of the GGTase-I, as FTase can transfer sequences that have X = Gln, Met, Ser, Ala whereas for geranylgeranylation by GGTase-I X could be either leucine or phenylalanine (Ohkanda, Lockman et al. 2001). The C-terminus Rab sequences recognised by GGTase-II as prenylation substrate are more diverse and cysteine residues in CC, CXC, CCX, CCXX, CCXXX sequences can be prenylated (Pylypenko, 2003).

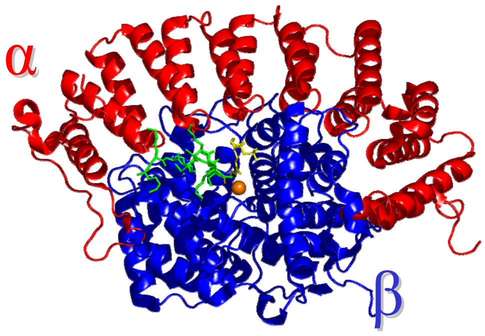


Figure 2.3a Farnesyl transferase (1qbq)

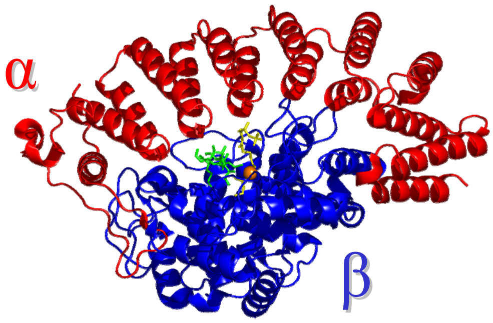


Figure 2.3b GGTase-I (1tnu)

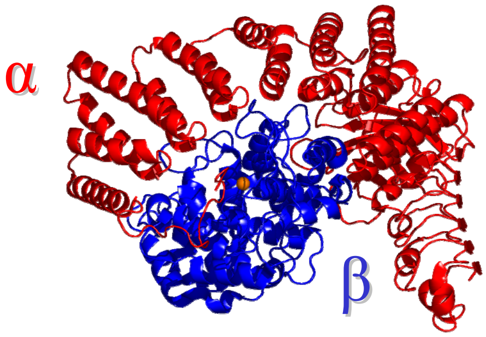


Figure 2.4 GGTase-II.

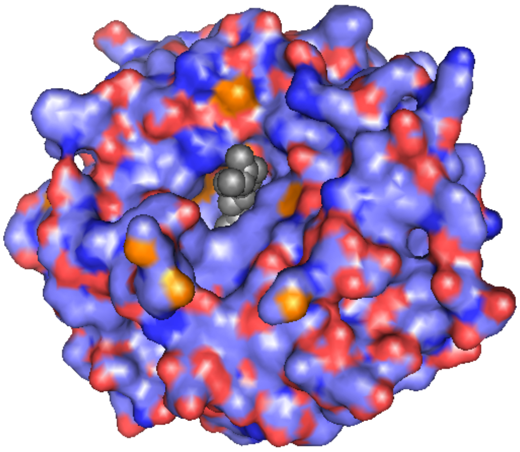


Figure 2.5a GGTase-II Chain B with Lipid in the binding pocket (1ltx)

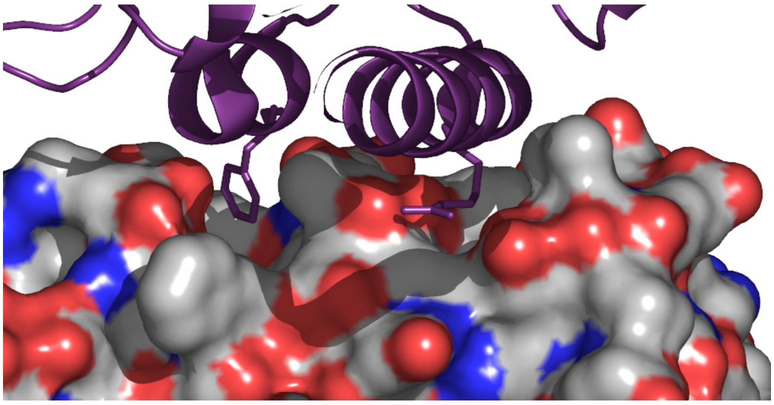


Figure 2.5a GGTase-II Chain A (Solid Surface) interacting with REP (Ribbon) (1ltx)

GGTase-II prenylates Rab GTPases by transferring the geranylgeranyl (GG) group from its pyrophosphate conjugate to the C-terminus cysteine residues. Unlike FTase and GGTase-I, GGTase-II interact indirectly (via Rab Escort Protein) with the protein substrate (RabGTPase) (Figure 2.4). Chain B of GGTase-II harbours a lipid binding pocket of ca. 490 Å3 wherein binds a single molecule of GGpp (Figure 2.5a). Chain A of GGTase-II has REP binding site (Figure 2.5b). GGTase-II and REP interaction interface is ca. 650 Å2 and involves bimodal interaction patches. GGTase-II also has a small hydrophobic patch (consisting of Ser249, Ala252, Phe254) which is probably involved in anchoring the C-terminus tail of Rab proteins thus increasing the activity (effective concentration) of prenylatable cysteine residues near the active site.

2.4.3 Rab Escort Protein

Rab escort protein or REP is 75 kDa protein organised in two domains: larger domain-I consists of 4 β-sheets and 6 α-helices and a smaller domain-II comprising of 5 α-helices (Figure 2.6). It can form transient complex with RabGTPases and GGTase-II-GGpp. The function of REP is to present the RabGTPases for prenylation followed by delivering it to the membrane.

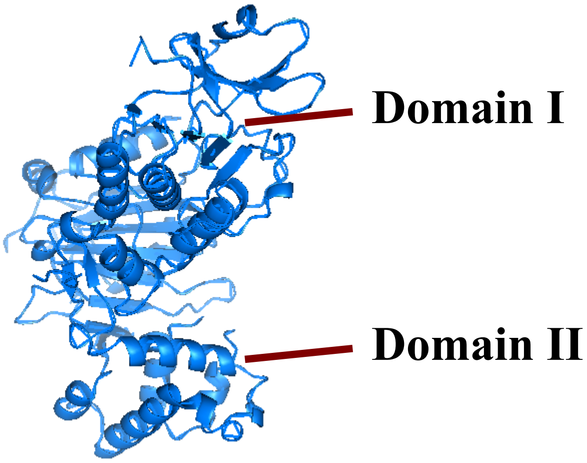


Figure 2.6 Domain arrangement of Rab Escort Protein (1ltx)

2.4.4 Rab-REP interface

REP interacts with Rab proteins via domain-I (Figure 2.7a). The REP surface involved in interaction with RabGTPases is called as Rab binding platform (RBP). The Rab-REP interaction interface is modular with patches of hydrogen bond making residues interspersed with hydrophobic patches and is ca 1075 Å^2 in size. It is quite unique in the absence of any major hydrophobic groove or pocket.

As the Rab proteins show considerable sequence diversity, interactions of Rab7 with REP are discussed. The interaction interface consists of Arg79 of Rab7 which makes a number of hydrogen bonds with the Asn225 and Glu379 of RBP. Asp44 and Asp63 residues of Rab7 also form hydrogen bonds with Arg386 of RBP. A number of hydrophobic residues in the switch II region of Rab7 interact with hydrophobic residues of RBP. However these hydrophobic residues are present either in shallow sites or on protein surface.

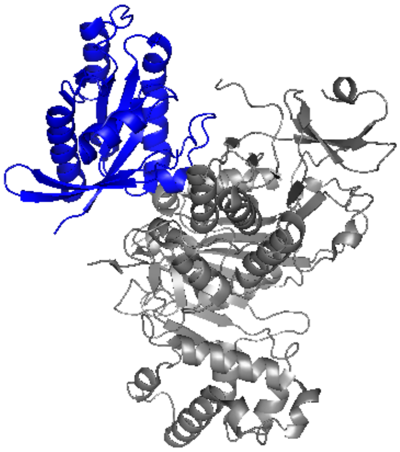


Figure 2.7a Rab7 (blue) interacting with REP (Black) (1vg9)

2.4.5 REP-GGTase-II interface

GGTase-II interaction with REP happens via the α-subunit of GGTase-II and domain II of REP (Figure 2.8a). The REP-GGTase-II interaction interface is very small ca. 690 Å^{2}. The interaction interface can be divided into a hydrophobic pocket and a hydrophilic patch. The hydrophilic patch on the surface of GGTase-II interacts with Arg290 of REP and a hydrophobic groove on the surface of GGTase-II harbours the side chain of Phe279 of REP near its opening (Figure 2.8b). This groove is not present on the surface of apo-GGTase-II structure (1dce) indicating that binding of GGpp in the lipid binding site in β-subunit of GGTase-II triggers its formation (Figure 2.8c) (Pylypenko, 2006).

The structure of GGTase-II alone (in the absence of lipid molecule) differs slightly from its structure in “GGpp bound GGTase-II”-REP complex. In the α-subunit, the differences are limited to the arrangement of residues of α-helices (8 and 10) which, along with helices (10 and 12) form the REP interacting interface, and in β-subunit conformational states of residues Tyr241, Trp244 and His190 differ in two states. The residues (Tyr241, Trp244 and His190) in the β-subunit form the lipid binding pocket in the GGTase-II. Hence the conformational change could be considered as the effect of the approach and binding of GGpp. However the shift in the position of α-helices (8) in α-subunit facing REP is considered to be necessary for the generation of deep hydrophobic pocket which interacts with Phe279 of REP by forming CH/pi interaction.

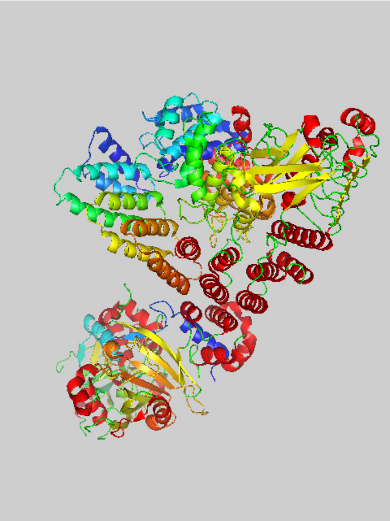


Figure 2.8a GGTase-II:REP

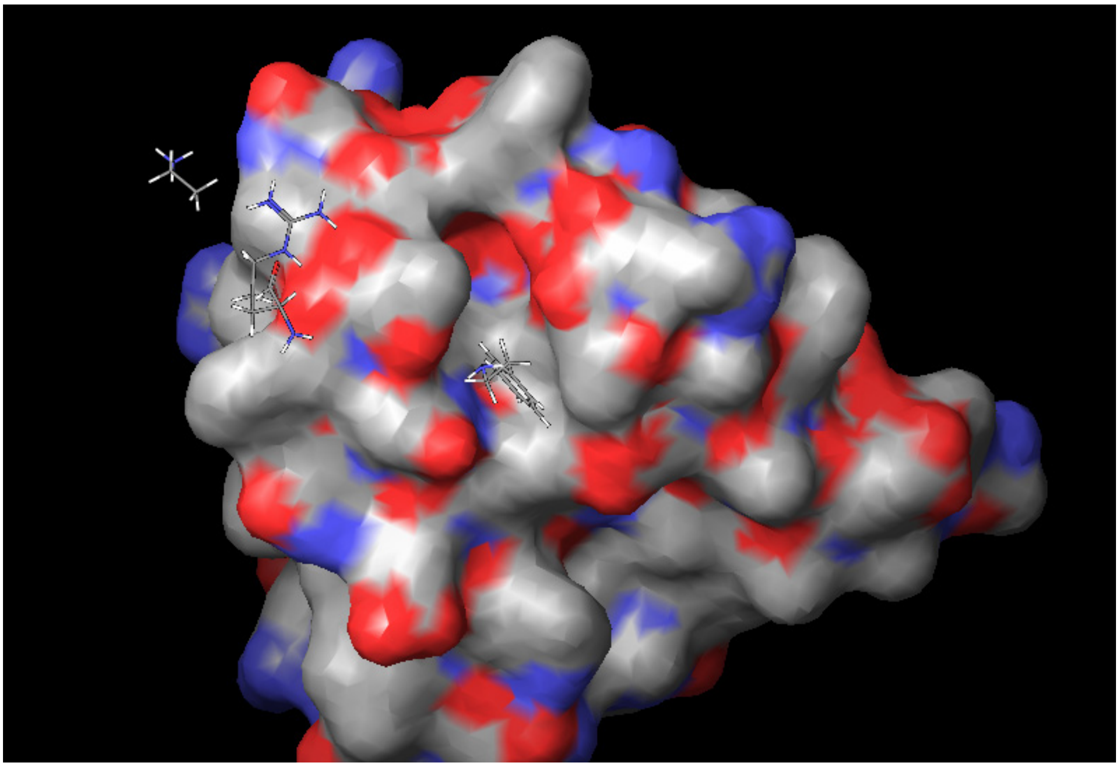


Figure 2.8b REP interaction interface of GGTase-II in complex with GGpp. (REP Side chains that interact with GGTase-II are represented by sticks and the GGTase-II interface is solid surface)

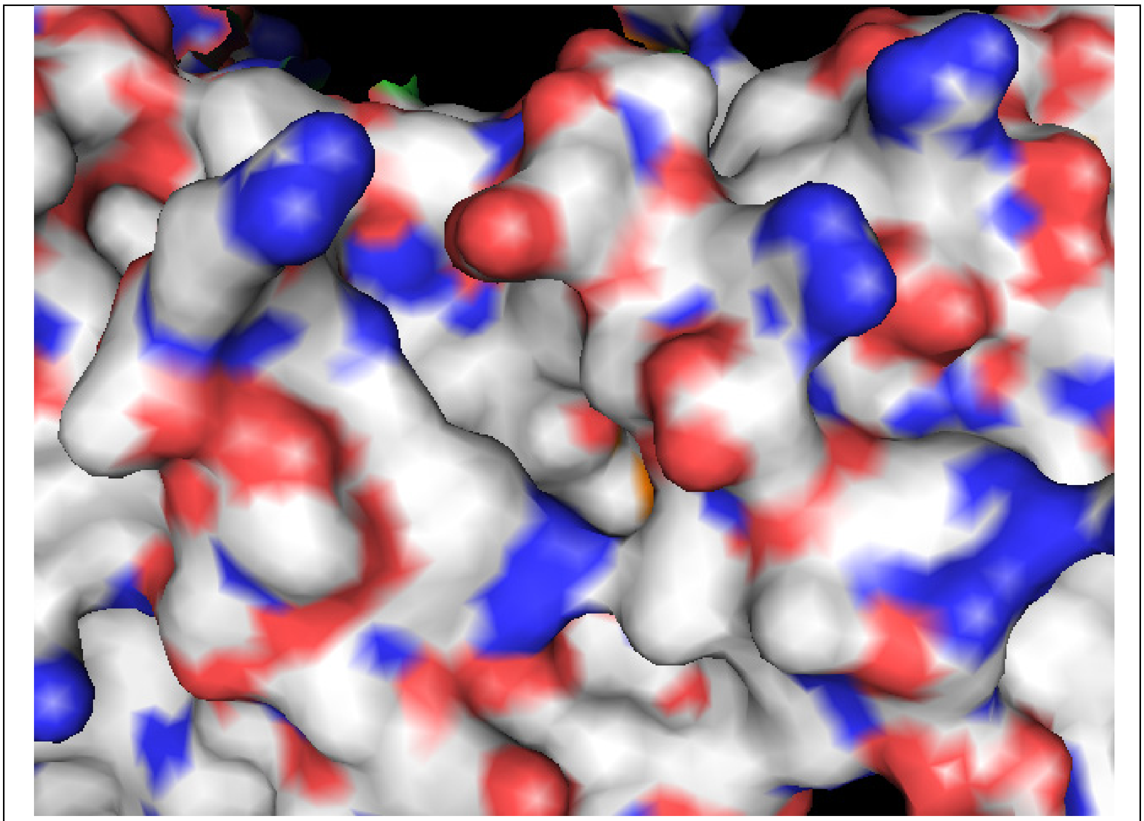


Figure 2.8c REP interacting interface of apoGGTase-II (1dce)

2.5 Targeting the enzyme system

2.5.1 Putative targets in the enzyme system

2.5.1.1 Ligand binding sites

The RabGTPase-REP-GGTase-II enzyme system interacts with lipid molecule. The lipid binding site in GGTase-II could be targeted and was being investigated by another group in the institute.

2.5.1.2 Selection of target site

Targeting REP-Rab interaction: Prenylation of Rab proteins is dependent on its interaction with REP; hence, the inhibition of REP-Rab protein interaction could disrupt the process. More over the RBP on the surface of REP seems to be conformationally stable and does not appear to undergo any major changes during its interaction with Rab hence targeting a hydrophobic pocket on RBP will not have to contend with any drastic conformational changes. Unfortunately, the RBP does not have any major hydrophobic pocket or groove and this eliminates the possibility for choosing REP-Rab interaction interface as potential targeting candidate.

Targeting REP-GGTase-II interaction: GGTase-II-REP interaction interface appears more suitable for targeting because even though the interface is small (690Å2), the affinity of REP for GGTase-II-GGpp binary complex is ca. 10nM (Rak, Pylypenko et al. 2004). Majority of interactions is mediated through Arg290 of REP which forms a number of hydrogen bonds with the GGTase-II. Inspite of presence of this Arg290-interacting hydrophilic patch on the surface of the apo-GGTase-II surface the affinity of apo-GGTase-II for REP is ca. 2 orders of magnitude less than the affinity between REP and GGTase-II:GGpp. Apparently the hydrophobic groove on the surface of GGTase-II:GGpp which interacts with Phe279 of REP is responsible for the higher affinity. Detailed analysis of the hydrophobic pocket on the REP-GGTase-II interaction interface revealed some of interesting structural features. The hydrophobic pocket on its REP-proximal end is mostly hydrophobic. The various methyl groups present in this part make CH/pi interactions with the Phe279 of REP. The REP-distal end of the hydrophobic pocket is rich in hydrogen bond acceptors. However, the absence of high resolution structure (crystal structure of REP-GGTase-II complex structure (1ltx) is of 2.75 Å resolution whereas the conventional accepted practice is to consider the structures with resolution better than 2.5Å) and the absence of any starting compound for targeting the site makes it a difficult choice.

2.6 Methods

2.6.1 Protocol used for epitope-linking using LigBuilder

LINK module of LigBuilder (See Appenndix-2.10.2) was used for interlinking the REP molecule's GGTase-II interacting epitopes. A MOL2 file containing seed structure comprising of the side chains of Phe279, Arg290 and Lys325 of the REP molecule was prepared using AstexViewer2.0. Addition of hydrogen atoms and assignment of ionisation states was done using molcharge program of OpenEye software. Terminal hydrogen atoms of these side chains were marked for linking process. The "population size" and "number of generations" for the linking the epitopes was fixed at 1000 members and 30,000 cycles respectively. The default values of growing probability, linking probability, and mutation probability (1.0, 1.0, 0.5) were not altered. The Tripos force field parameters were used for the linker generation. The program was run on a desktop computer having RAM - 1GB and processor clock time of 2.2 GHz.

1000 different molecules were produced by interlinking the epitopes (see section 2.6.1). From this library top 250 compounds were docked in the targeted site and the results were subjected to the RMSD filter.

2.6.2 Protocol for growing molecule using LigBuilder

In the hydrophobic groove of GGTase-II which is part of REP interaction interface novel ligand molecules were grown on the docked structure of guanidine using GROW module of LigBuilder. The MOL2 file of the seed structure of guanidine in the site was prepared using AstexViewer2.0. Hydrogen atoms were added and ionisation states were assigned using molcharge program of OpenEye. The "population size" and "number of generations" for the growing the hydrophobic tail of guanidine was fixed at 3000 members and 20 cycles respectively. The number of parents and similarity cut-off were fixed at 200 and 0.90. The values of growing probability, linking probability, and mutation probability (1.0, 1.0, 0.60) were used. The Tripos force field parameters were used for the growing the molecules. The program was run on a desktop computer having RAM - 1GB and processor clock time of 2.2 GHz. 3000 molecules were obtained after one growth event. These were docked and the top 300 molecules were subjected to the RMSD filter. The process was repeated 32 times and in the end the library of selected top grown molecules consisted of 9600 molecules. After the docking solutions for each of these molecules was subjected to RMSD filter.

2.6.3 Protocol for diversification of the hydrophobic part of lead_molecule_1

The cyclo-hexane part of lead_molecule_1 was marked for mutation and subjected to repeated growth cycles. The values of growth probability, linking probability and mutation probability was fixed at 0.5, 0.5 and 0.95 respectively. The fragment library from which LigBuilder selects molecular fragment for incremental construction was reorganized by retaining only hydrophobic and aromatic ring structures (for example benzene, anthracene etc). This reorganized fragment library contained around 100 fragments. After 30 cycles a diversification library of 9000 top structures was created. These structures were docked using GOLD in the targeted site and subjected to RMSD filter.

2.6.4 Pharmacophore generation and virtual screening

Selected molecules from diversified library passed through the RMSD filter and were used in generation of pharmacophore hypothesis using catalyst (see Appendix-2.10.3). Using CONFIRM program from the catalyst package conformation for the members of ZINC database of drug-like chemically available structures was created. This conformational database was then screened using the pharmacophore hypothesis. The selected compounds were subjected to the docking analysis. Compounds that passed RMSD filter with both GoldScore and ChemScore were purchased and assayed.

2.6.5 Drug-likeness Filter

The compounds designed during the process of epitope linking and guanidine tail growth were subjected to a drug-likeness filter. The filter comprised of a set of ranges for molecular weight, number of heavy atoms, lipophilicity, number of hydrogen bond donors and acceptors.

Molecular weight from 160 to 480

Number of heavy atoms from 20 to 70

Lipophilicity from 40 to 130

Number of hydrogen bond donors from 4 to 7

Number of hydrogen bond acceptors from 8 to 12

Only those molecules that conformed to the above mentioned criteria were selected for evaluation by docking.

2.6.6 GOLD docking protocol

All of the docking runs were carried out using the default parameters of GOLD program. For each molecule docking runs were carried out twice, once using GoldScore and second using ChemScore. Only top 10 docking solutions were considered.

2.6.7 RMSD stability Filter

For the top 10 docking solutions for each molecule average RMSD was calculated. Mathematically,

$$\text{AveRMSD} = (\Sigma_{ij}\surd((x_{ij} - X)^2 + (y_{ij} - Y)^2 + (z_{ij} - Z)^2))/(10N) \qquad 2.1$$

Where, *i* subscript range for all of the 10 poses and *j* subscript ranges for all of the atoms in the molecule. The X, Y and Z are the average values for X, Y and Z coordinates for the top ranked docking solution. N is the number of atoms in the molecule. The molecules having average RMSD less than 2.0Å were selected as being stable in the target site.

2.7 Results and discussion

2.7.1 Linking epitope

None of the 250 compounds had average RMSD below 2.0Å. Visual examination of the individual structures revealed presence of more than 10 single bonds in each of the structures. One of the examples of the docked structure is shown in Figure 2.9.

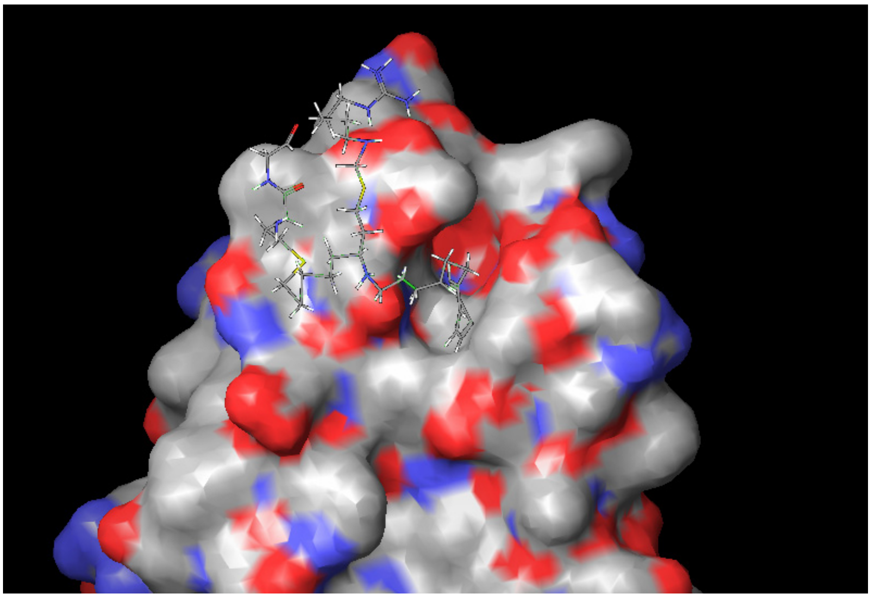


Figure 2.9 Side chains of Phe279 and Arg290 were linked using LINK module of LigBuilder and docked in the Phe279 interacting groove on GGTase-II surface

2.7.2 Docking based identification of stable “anchor” fragment for growing molecules

Ammonia, benzene, guanidine, methanol and methanoic acid were docked using GOLD program in the targeted site. The docking solution for each of small molecule fragment was subjected to RMSD filter. Except for guanidine docking solutions of the rest of the molecules had average RMSD above 2.0Å. The highest scoring docking pose of guanidine in the targeted site is shown in Figure 2.10.

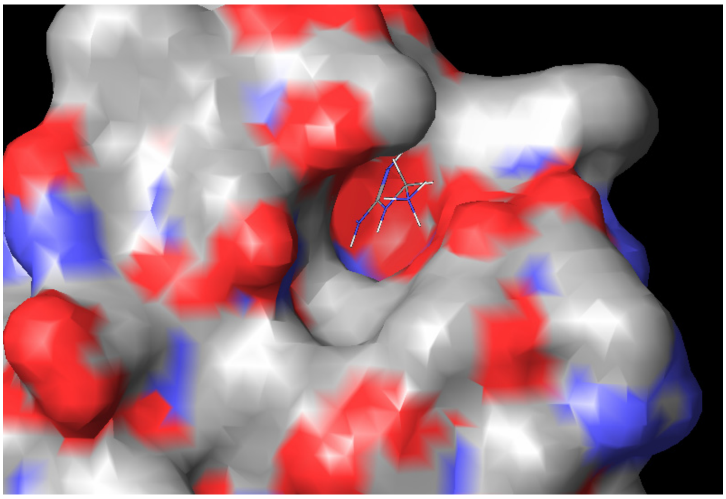


Figure 2.10 Docking solutions for Guanidine molecule in the site were stable in terms of RMSD and formed multiple hydrogen bonds with the hydrogen bond acceptors at the base of the cavity

2.7.3 Molecules generated by growing the guanidine tail

Only 23 molecules in the library of 9600 grown molecules had average RMSD less than 2.0Å. After visual analysis of the docking results only one was selected. As seen in Figure 2.11 the guanidine head makes multiple hydrogen bonds with the oxygen bond acceptors with the side chains of residues Ser227, Asp225, and main chain carbonyl group of Asn174 at the base of the targeted site. The hydrophobic cyclo-hexane part makes hydrophobic interactions with the hydrophobic side chains of Ile171, Ala218 and Leu214. This molecule was named as lead_molecule_1.

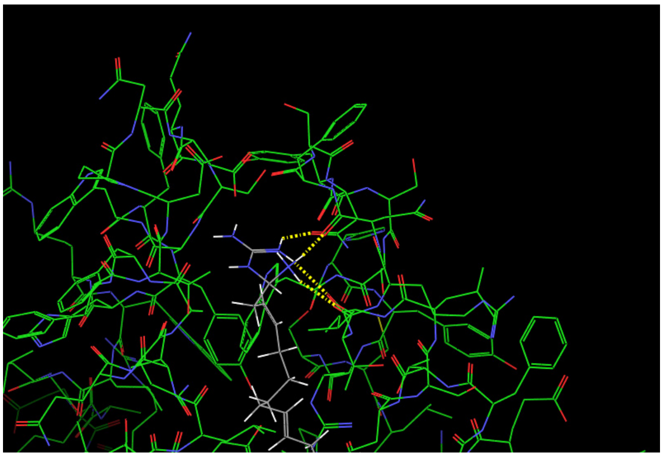


Figure 2.11 Virtual lead compound.

2.7.4 Diversification of the hydrophobic part of lead_molecule_1; Pharmacophore generation; Virtual screening and Assay results

Out of 9000 members of diversified library only 73 could pass the RMSD filter. A pharmacophore query was built from these 73 compounds as shown in Figure 2.12.

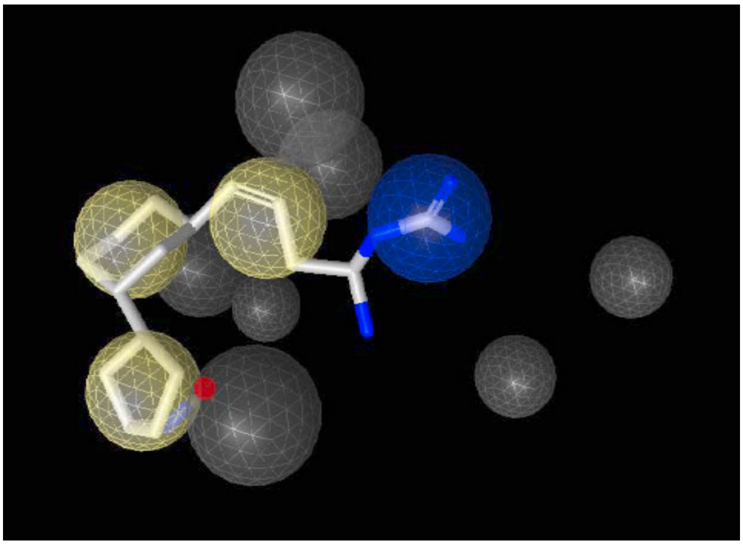


Figure 2.12 Pharmacophore has 3 Hydrophobic regions (cream spheres) and one positively charged feature (blue sphere) The grey spheres are the exclusion volume region.

9883 compounds conformed to the pharmacophore hypothesis. These molecules were visually inspected to remove molecules with possible steric clashes near the hydrogen bond donor fragment. Only 41 compounds had no clashes. After the docking analysis of these 41 compounds only 27 that could pass the RMSD filter were ordered Figure 2.13(a-d). 9 compounds (Figure 2.14) could be purchased and assayed. The assay was carried out by Yao-Wen Wu in the Department of Physical Biochemistry, Max Planck Institute for Molecular Physiology. Compound named as MK_INH_X21985 was found to have IC50 value of 7.0µM. The compound is non-competitive for the lipid substrate as per the results of competitive assay.

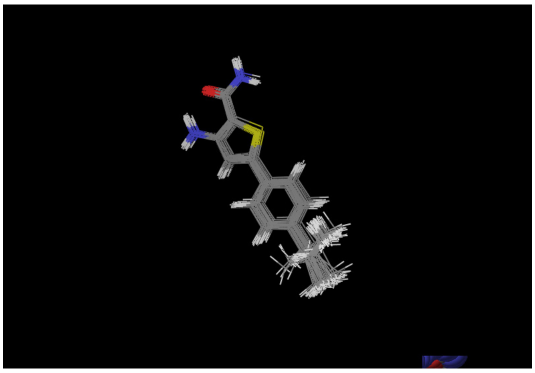
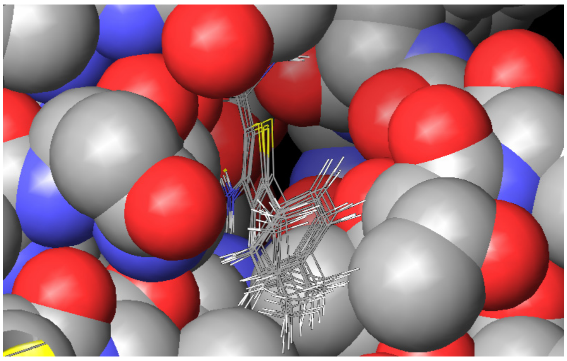
	

Figure 2.13a Docked poses of Compound MK_INH_X16156	Figure 2.13b Docked poses of Compound MK_INH_X16156 in the target site

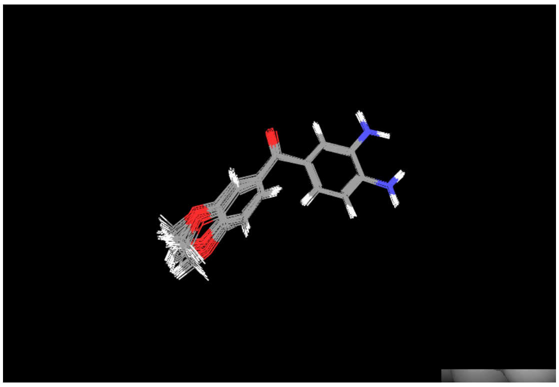
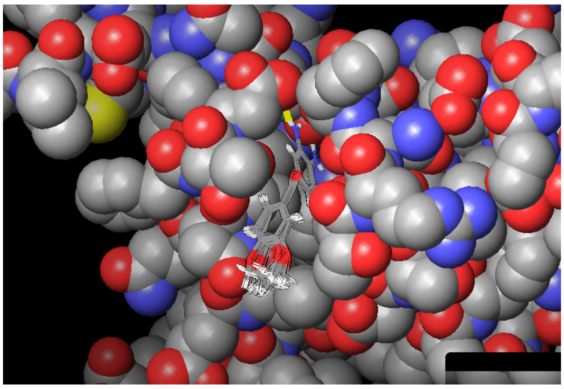

	

Figure 2.13c Docked poses of Compound MK_INH_X16188	Figure 2.13d Docked poses of Compound MK_INH_X16188 in the target site

2.8 Conclusions

The IC50 of the compound MK_INH_X21986 is below 10µM and hence by convention it is classified as active. This compound was assayed earlier for inhibition of RNAseH and was found to be inactive hence the compound is not a chelator of divalent ions or a non specific protein poison. The compound is noncompetitive with the lipid substrate and this excludes the possibility of its binding in the lipid binding pocket. The docking analysis of the compound for the peptide binding site shows poor stability. The docking analysis of the compound in the sites of GGTase-II along with available wet lab results indicates that the compound is targeting REP interaction site. However, in the absence of conclusive assay the mechanism of inhibition is still an open question.

Rational drug design has been the ultimate aim of the structural bioinformatics. The targeting of protein-protein interaction on the basis of knowledge about the receptor structure is difficult. However the successful targeting of GGTase-II activity by MK_INH_X21986 once again underlines the possibilities in this field. Here, it must be noted that later the compound was also found to inhibit the homologs of GGTase-II. Careful docking analysis predicted FTase inhibition due to the competition of peptide substrate with the MK_INH_X21986. This was indeed found to be the case in wet lab experiments. The cross reactivity question was not addressed when the molecule was being designed as it was assumed that the shape of REP interacting hydrophobic groove on GGTase-II surface is unlikely to find anything similar in the homologous structures. This was a mistake. The strategy for targeting protein-protein interaction interface must be such that only unique druggable sites are chosen in the first place. Alternatively targeting a site which may have close resemblances on the surface of homologous proteins can be done through substractive docking. This strategy was indeed used in the later process and a virtual library of molecules was created that targeted only the GGTase-II hydrophobic groove.The validity of the model is yet to be tested.

PDF-XChange Click to buy NOW! www.docu-track.com

2.9 References

Arkin, M. R., M. Randal, et al. (2003). "Binding of small molecules to an adaptive protein-protein interface." Proceedings of the National Academy of Sciences of the United States of America **100**(4): 1603-1608.

Arkin, M. R. and J. A. Wells (2004). "Small-molecule inhibitors of protein-protein interactions: Progressing towards the dream." Nature Reviews Drug Discovery **3**(4): 301-317.

Itzen, A., O. Pylypenko, et al. (2006). "Nucleotide exchange via local protein unfolding - structure of Rab8 in complex with MSS4." Embo Journal **25**(7): 1445-1455.

Barnum, D., J. Greene, et al. (1996). "Identification of common functional configurations among molecules." J Chem Inf Comput Sci **36**(3): 563-71.

Eldridge, M. D., C. W. Murray, et al. (1997). "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 **11**(5): 425-45.

Goldberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. Reading, Mass., London, Addison-Wesley.

Kirchmair, J., G. Wolber, et al. (2006). "Comparative performance assessment of the conformational model generators omega and catalyst: a large-scale survey on the retrieval of protein-bound ligand conformations." J Chem Inf Model **46**(4): 1848-61.

Kurogi, Y. and O. F. Guner (2001). "Pharmacophore modeling and three-dimensional database searching for drug design using catalyst." Curr Med Chem **8**(9): 1035-55.

Li, H., J. Sutter, et al. (2000). HypoGen: An automated system for generating 3D predictive pharmacophore models. LaJolla, CA International University Line.

Ohkanda, J., J. W. Lockman, et al. (2001). "Peptidomimetic inhibitors of protein farnesyltransferase show potent antimalarial activity." Bioorganic & Medicinal Chemistry Letters **11**(6): 761-764.

Olsen, L., I. Pettersson, et al. (2004). "Docking and scoring of metallo-beta-lactamases inhibitors." J Comput Aided Mol Des **18**(4): 287-302.

Smellie, A., S. L. Teig, et al. (1995). "Poling: Promoting Conformational Variation." Journal of Computational Chemistry **16**(2): 171-187.

Toba, S., J. Srinivasan, et al. (2006). "Using pharmacophore models to gain insight into structural binding and virtual screening: an application study with CDK2 and human DHFR." J Chem Inf Model **46**(2): 728-35.

Verdonk, M. L., J. C. Cole, et al. (2003). "Improved protein-ligand docking using GOLD." Proteins **52**(4): 609-23.

Nooren, I. M. A. and J. M. Thornton (2003). "Diversity of protein-protein interactions." Embo Journal **22**(14): 3486-3492.

Ohkanda, J., J. W. Lockman, et al. (2001). "Peptidomimetic inhibitors of protein farnesyltransferase show potent antimalarial activity." Bioorganic & Medicinal Chemistry Letters **11**(6): 761-764.

Rak, A., O. Pylypenko, et al. (2004). "Structure of the Rab7 : REP-1 complex: Insights into the mechanism of rab prenylation and choroideremia disease." Cell **117**(6): 749-760.

2.10 Appendices

2.10.1 GOLD

2.10.1.1 Algorithm overview

GOLD (Genetic Optimisation for Ligand Docking) uses a genetic algorithm (GA) to explore the conformational search space and a molecular mechanics like scoring function (see section 2.9.1.3) to evaluate and rank generated docking solutions(Eldridge et al, 1997). Genetic algorithms are widely used as search algorithms for optimisation problems. During optimisation they use evolution as model and adapt and improve the solution by using the strategy of mutation and selection. The existing solutions are changed randomly and then selected by filtering out less fit solutions based on a scoring function (Goldberg et al , 1989). It does this by manipulating so called chromosomes, which are represented as strings that can undergo reproduction, crossover and mutation. Just as the total set of chromosomes makes up the genotype of a species, a collection of strings or in the most simple case just one string are termed the structure of an artificial system. This structure encodes a set of parameters or points in solution space, similar to the genotype encoding the phenotype of an organism. Chromosomes are a collection of genes, each represented by an allele and location, whereas strings comprise features, each associated with a value and location.

In terms of GOLD docking, each chromosome represents a possible solution to the ligand-docking problem. Chromosomes are treated as individuals and as part of a population (of fixed population size) where each member is evaluated for fitness. Parent chromosomes are then randomly chosen but biased towards fitness and subjected to reproduction operators, producing child chromosomes. Their fitness is evaluated and if novel, they replace the least fit individual in the population. The whole process including operator and parent selection is repeated unless an acceptable solution is found. In the extended version of the algorithm, populations are split into sub-populations and additional genetic operator migration is introduced, allowing individuals to move across sub-populations. Crossover recombines two parent chromosomes whereas mutation changes a value at random. “Survival of the fittest” is achieved over time, moving the population to the best solution for the docking problem. The fitness function plays an essential role in the selection process and determines how accurately it can predict the binding conformation. Starting ligand poses (chromosomes) are generated at random. Each chromosome contains protein-ligand mappings of interaction

PDF-XChange Click to buy NOW! www.docu-track.com

points (hydrogen bonds, hydrophobic points, conformation around rotatable bonds) and is given a fitness score according to evaluation of the scoring function.

Irrespective of the stochastic (random or probabilistic) nature of the algorithm, it is generally highly reproducible but with some targets it is target-dependent, requiring more or longer GA runs to obtain a match to the crystallographic binding mode (Kirchmair et al, 2006).

2.10.1.2 Handling metal ions in GOLD

GOLD uses coordination geometry templates which are mapped onto the metal coordinating protein residues in order to establish its coordination geometry. Fitting points are then created for unoccupied sites which are used in docking for ligand acceptor matching. The metal-ligand interaction is treated as pseudo-hydrogen bonding, where ligands are acceptors and the metal competes with hydrogen bond donors for ligand binding. The significance of metal treatment for this study is discussed in section 1.2.2.4.

2.10.1.3 Scoring functions in GOLD

Two scoring functions are implemented in GOLD, the force-field based GoldScore and the regression-based ChemScore. In addition, there is the option to use a user-specific scoring function. The GOLD implementation of ChemScore (ChemScoreGOLD) differs from the original scoring function (Kurogi et al 2001), since a clash penalty and an internal torsion term are added to the final score to penalise bad contacts and poor conformations. The GOLD scoring function GoldScore uses a set of empirical parameters from a modifiable parameter file. Some correlation was found with experimental binding affinities (Li et al 2000), although it was originally designed for optimal pose selection (Okhanda et al 2001). The function uses a 6-12 potential Lennard-Jones function for the intra-molecular (internal) vdW score and a “soft” 4-8 for the intermolecular (external) score.

2.10.2 LIGBUILDER

LigBuilder (wang et al 2001) is structure -based drug design software. On the basis of the 3-dimensional structure of the target protein, it builds ligand molecules within the binding pocket. The program identifies the key interaction sites by analyzing the binding pocket of the target protein. For the identified sites a pharmacophore model is built which could be applied to 3-dimensional database search for finding novel ligand molecules. Molecules are constructed using incremental construction and the minimization of conformation is performed during the building-up procedure. While the target protein is kept rigid, flexibility of the ligand molecules is considered. Molecules can be built in the site by growing or linking strategy. Built molecules are evolved by Genetic Algorithm. The fitness score of a molecule is evaluated by considering its chemical viability as well as binding affinity. Chemical rules are adopted for evaluating "drug-likeness" of the resultant molecules. Chemical stability, synthesis feasibility, and toxicity can also be taken into account by defining "forbidden structure" libraries.

2.10.2.1 Growing strategy and linking strategy

The central function of LigBuilder is constructing ligand molecules within the constraints of the target protein. LigBuilder supports two strategies to do this, i.e. growing strategy and linking strategy. To apply the growing strategy, you need to provide a pre-placed "seed" structure inside the binding pocket and LigBuilder will subsequently add fragments onto it to build molecules. This strategy may be helpful when you have got an interesting lead compound and want to develop its derivatives to improve the bioactivity (lead optimization). To apply the linking strategy, you also need to provide a starting structure, which consists of several separated chemical fragments. These fragments should be pre-placed inside the binding pocket and would better to form favorable interactions with the target protein. Then LigBuilder will try to build molecular frameworks to link those fragments into integrated molecules. This strategy may be helpful when you try to find novel lead compounds (lead discovery).

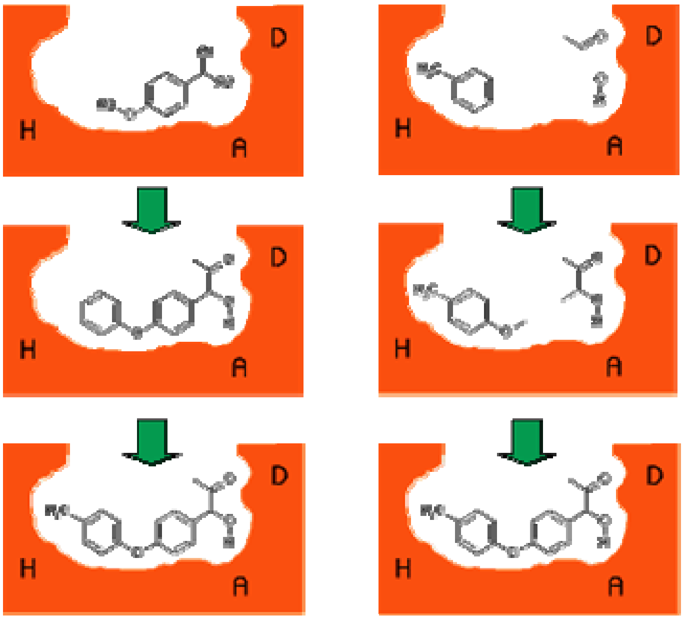
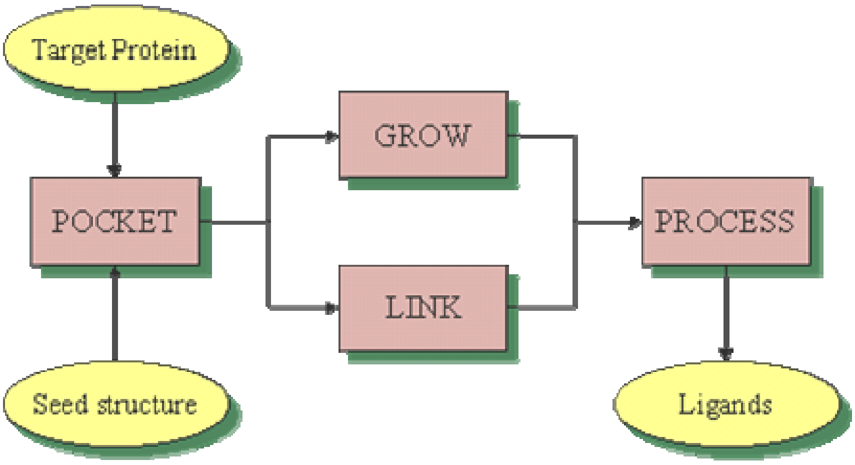


Illustration of growing strategy and linking strategy

2.10.2.2 Overall structure of LigBuilder

LigBuilder has four main functional modules, i.e. POCKET, GROW, LINK, and PROCESS. POCKET is designed to analyze the binding pocket of the given protein and prepare the data necessary to run GROW or LINK. GROW is designed for performing growing strategy while LINK for linking strategy. The ligands generated by GROW or LINK will be collected in a data file and read by PROCESS to give the final viewable results.



2.10.3 Catalyst

Catalyst provides a platform for pharmacophore generation and database searching (Olsen et al 2001). The pharmacophore is an important concept in medicinal chemistry and aids interpretation of structure activity relationships of a series of compounds for the identification of novel ligands. Pharmacophore models (also called hypotheses) comprise chemical features mapped to a coordinate point with a given tolerance sphere. These features represent hydrogen bond donors or acceptors, aliphatic or aromatic groups and positive or negative ionisable groups among others. The model can be made more specific by including exclusion volume spheres as steric constraints (representing regions in the protein) that must not be occupied by a ligand atom. In addition, a shape pharmacophore model can be constructed based on a ligand's shape and used on its own or merged with the chemical feature model. The pharmacophore model can then be used to screen a compound database for compounds with matching features (Smellie et al, 1995).

2.10.3.1 Pharmacophore model generation

Feature-based pharmacophore models can be constructed manually based on a known bioactive conformation or in an automated fashion. Two algorithms are provided for quantitative and qualitative model generation. In the first case, the HypoGen algorithm requires a set of 15-25 diverse molecules with activities spanning at least 4 orders of magnitudes, which were determined under comparable assay conditions (Toba et al, 2006). These models can be used to predict affinity of screening hits. The HipHop algorithm (Verdonk et al, 2003) generates a common feature-based pharmacophore model based on two or more highly active and structurally diverse compounds. The model allows distinguishing between active and inactive compounds. A number of chemical features can be specified for automated hypothesis generation or new, customised features generated such as a zinc binding feature. Selection of an adequate training set is critical for successful predictions. Conformational models of the training set compounds can be created using a Monte Carlo approach to exhaustively cover conformational space by providing a diverse set of low energy conformers. Two methods exist, the fast method which is ideally used for database generation and the best method which implements the poling algorithm and should be used for hypothesis building. Poling promotes the generation of dissimilar conformers and removes redundancy by penalising close conformations. It can be combined with any conformational analysis method that seeks local minimisation of a penalty or energy function. It is implemented by adding an extra term to the energy function. Conformational variation can be enforced for a set of interesting features or dissimilar parts in structurally related compounds. The number of generated conformers is molecule dependent, but a maximum of 50 conformers per ligand was suggested for database generation using the fast method.

2.10.3.2 Database searching

A number of 1D filters such as activity keywords or Molecular Weight can be specified in addition to the pharmacophore hypothesis for database screening. A pre-generated multi-conformer compound database is mapped against the features specified in the pharmacophore hypothesis. Resulting hit compounds are evaluated by calculating a geometric “best fit” score which analyses the quality of matching features. Two algorithms can be applied for the search, either the fast flexible or best flexible methods. The fast algorithm uses only pre-generated conformers, whereas the best algorithm can slightly modify the conformers in order to achieve a fit.

It handles conformational flexibility by pre-generating a representative set of diverse and low energy conformations with the poling algorithm and storing those conformations in the database. This multi-conformer database can be searched rigidly or flexibly, indicated by the fast or best search option. The fast algorithm only considers existing conformers and interrupts a search as soon as a pharmacophore matching conformation is found, whereas the best algorithm additionally ‘tweaks’ bond distances, angles and dihedral angles of pre-generated conformers on the fly to achieve the best matches. Hit molecules can be ranked by their geometric fit values which indicate how well the chemical substructures were mapped onto the hypothesis feature location constraints and their distance deviation from the feature centres. High fit values indicate good matches with the maximum fit value set by the original ligand used to create the pharmacophore.

Chapter 3: InCa-SiteFinder: A method for structure based prediction of carbohydrate binding sites on proteins

3.1 Abstract

We present the development and optimisation of a new method called InCa-SiteFinder to predict inositol and carbohydrate binding sites on the surface of the protein structures. It uses the van der Waals energy of a protein-probe interaction and amino acid propensities to predict carbohydrate binding sites. The protein surface is searched for regions that correspond to a favourable energy for a protein-probe interaction using a grid approach. These regions of favourable interaction energy are subsequently analysed to demarcate regions of high cumulative propensity for binding a carbohydrate moiety based on calculated amino acid propensity scores. These scores were obtained for carbohydrate binding sites using a Non-Redundant Dataset (NRD) of 375 protein structures. In order to optimise the predictive capacity of the method an independent training set of 50 protein-carbohydrate complexes was retained to optimise thresholds values for the protein-probe energy and amino acid propensity. The optimised InCa-SiteFinder was tested on a test set of 80 protein-ligand complexes. It efficiently identifies carbohydrate binding sites with high specificity and sensitivity. It was also tested on a second test set of 80 members containing 40 known carbohydrate binders (having 40 carbohydrate binding sites) and 40 known drug-like compound binder proteins (having 58 known drug-like compound binding sites) for the prediction of the location of the carbohydrate binding sites and to distinguish these from the drug-like compound binding sites. At 72.5% sensitivity the method showed 98.3% specificity. Almost all of the carbohydrate and drug-like compound binding sites were correctly identified with an overall error rate of 12.2%.

3.2 Introduction

The functionality of a protein is closely controlled by the nature of molecules it can interact with. Even though the number of known structures of proteins has grown rapidly in the recent years (Tagari, Tate et al. 2006) a large number of protein-ligand interaction sites remain uncharacterised (Laurie and Jackson 2005) . A number of approaches have been developed to make predictions about the function of a protein from its structure (Laskowski, Watson et al. 2005). Some methods look for motifs or domains associated with specific functions (Laskowski, Watson et al. 2005), others tend to look for characteristic arrangement of functionally important or conserved residues (Burgoyne and Jackson 2006). The function of a protein depends upon the nature of ligand it can interact with, hence demarcation of the ligand binding sites and identification of the type of ligand that can interact is important for the assignment of function to the protein structure as well as for rational structure-based drug design.

Carbohydrate binding proteins play an important role in cellular systems. Carbohydrate binding is involved in energy flow, cellular recognition and adhesion (Brandley and Schnaar 1986). Carbohydrate binding proteins are very diverse in structure and function (Bertozzi and Kiessling 2001). They are increasingly being considered as putative drug targets (Bertozzi and Kiessling 2001) because of their role in intra and inter-cellular communication. Carbohydrate binding sites have been extensively studied (Weis and Drickamer 1996) in the past. However, only a few approaches developed for the prediction of carbohydrate binding sites (Taroni, Jones et al. 2000), (Shionyu-Mitsuyama, Shirai et al. 2003) and (Malik and Ahmad 2007). Taroni et al used amino acid propensity for carbohydrate binding and identified the patches on the protein surface having an average propensity score above a specific threshold. Shionyu-Mitsuyama et al developed a set of rules from a dataset of 80 protein-carbohydrate binding sites that depicted, on a 3-dimensional grid the probable positions of carbohydrate-interacting protein atoms. Using a set of only 10 atom types a 3-dimensional probability density map was created. Each point on this map represented the probability of occurrence of a protein atom which could interact with a carbohydrate. Using these interaction maps they predicted the carbohydrate binding sites. Malik et al trained a neural network using amino acid propensities for the prediction of carbohydrate binding sites. The training set comprised of 40 protein-carbohydrate complexes and the level of redundancy was reduced by removing protein sequences with more than 50% sequence identity.

PDF-XChange Click to buy NOW! www.docu-track.com

Here the development of a new computational method for predicting carbohydrate binding sites is presented. The overall aim was to develop a new computational method for predicting carbohydrate binding sites with high accuracy. The method differs from the previous carbohydrate binding site prediction methods in two important aspects. Firstly it uses 375 non-covalent protein-carbohydrate complexes for the derivation of amino acid propensity scores. This is more than used in calculation of amino acid propensities in the previous methods. Secondly it uses a two-step procedure to identify sites. In step one; it uses a grid-based approach to identify sites on the protein with a high probability of being a binding site, using the recently proposed method of Laurie and Jackson, 2005. In step two; it uses these sites and amino acid propensity scores to predict the location of carbohydrate binding sites. The ultimate aim of developing InCa-SiteFinder was to produce a method that could both locate likely binding sites and then distinguish the nature of the binding site, to ascertain if the site has the ability to preferentially bind a carbohydrate ligand.

3.3 Methods

3.3.1 Construction of dataset for propensity calculation

Nearly 30,000 protein-ligand complexes present in PDBSUM (Laskowski, Hutchinson et al. 1997; Laskowski 2001; Laskowski, Chistyakov et al. 2005) with structural information were extracted from the PDB (Berman, Westbrook et al. 2000). From these only protein-carbohydrate complexes having experimentally determined x-ray crystal structures with a resolution greater than 2.5Å were retained. In addition, complexes having either, a covalently bound ligand or involving a drug-like compound ligand or metallic ions or not having a classification in SCOP (version 1.69) (Murzin, Brenner et al. 1995) were further removed. A ligand was classified as non-covalently bound to the protein if none of its atom was within the covalent interaction distance. The covalent interaction distance for a specific pair of protein and ligand atom was the sum of their atomic radii plus a 10% tolerance limit.

Only the best resolution complex with a unique carbohydrate name and unique SCOP superfamily are further retained. These comprised a non-redundant dataset (NRD) with only one carbohydrate representative for each SCOP superfamily. Hydrogen atoms were added to these protein-carbohydrate complexes using the QuacPac software (OpenEye). The definition of an atomic contact was taken from DrugScore (Gohlke, Hendlich et al. 2000) in which two atoms are considered to be in contact if the intervening distance between the atoms is less than the sum of their van der Waals radii plus 1Å. The cut-off therefore includes only short-range interactions.

3.3.2 Calculation of amino acid propensities

For a non-redundant database of over 375 protein-carbohydrate complexes propensities for a given amino acid to occur in a carbohydrate binding sites were calculated as the ratio of its likelihood to contribute to the carbohydrate binding site surface to its likelihood to contribute to the complete protein surface. The area occupied by an amino acid i in the carbohydrate binding site was considered as the difference in solvent accessible surface area when calculated in the presence and absence of carbohydrate. The Propensity of an amino acid, i, to occur in carbohydrate binding site is given by:

$$\mathrm{Prop}_i = ((\Delta \mathrm{CBS\_SASA}_i/\Sigma_{j=1}^{20}\Delta \mathrm{CBS\_SASA}_j)/(\mathrm{SASA}_i/\Sigma_{j=1}^{20}\mathrm{SASA}_j)) \qquad (3.1)$$

Where, $\Delta \mathrm{CBS\_SASA}_i$ is the solvent accessible surface area buried during the binding of carbohydrate molecule to the cleft for a specific amino acid *i*. $\Sigma_{j=1}^{20}\Delta \mathrm{CBS\_SASA}_j$ is the total solvent accessible surface area of all amino acids buried. SASA_i is the solvent accessible surface area contributed by a specific amino acid *i* on the protein surface. $\Sigma_{j=1}^{20}\mathrm{SASA}_j$ is the total solvent accessible surface area of all amino acids of the protein. For comparison the amino acid propensities of drug-like compound binding sites were also determined in the same way. These were calculated from a non redundant database of 358 complexes of protein-drug-like compounds. The ligands were considered as drug-like if they conformed to Lipinski's rule of 5(Ghose, Viswanadhan et al. 1999; Viswanadhan, Ghose et al. 1999; Viswanadhan, Ghose et al. 1999; Lipinski, Lombardo et al. 2001) and did not contain a carbohydrate moiety.

3.3.3 InCa-SiteFinder

The process of calculating the protein-probe van der Waals interaction energy is described in detail in Laurie and Jackson, 2005. Briefly, the protein atom coordinates are extracted from a PDB file and hydrogen atoms are added to these using the method of Jackson et al (1998). The protein atoms are placed in a 3-dimensional box, which is divided into a cubic grid of resolution 0.9 Å. Using the program Liggrid the van der Waals energy of interaction is calculated between the protein and a methylene (-CH3) probe placed at each grid point. The energy is calculated using the GRID forcefield parameters as described by Jackson 2002. Grid points with a "protein-probe interaction" energy more favourable (negative) than a predetermined threshold are retained (Figure 3.1).

For these grid points, carbohydrate binding propensity (CBP) and drug-like compound binding propensity (nCBP) scores are calculated by considering the identity of amino acid residue whose atoms fall under the interaction zone. An amino acid is considered to be interacting with a grid point if at least one of its atoms is within 1.6 Å of the grid point. The overall carbohydrate binding propensity (CBP) and drug-like compound binding propensity (nCBP) scores of the grid point, *k*, are defined as:

$$\mathrm{CBP\ Score}_k = (\Sigma_{i=1}^{20} n_i \mathrm{CBP}_i/N) \qquad (3.2)$$

$$\mathrm{nCBP\ Score}k = (\Sigma_{i=1}^{20}(n_i)*(\mathrm{nCBP}_i)/N) \qquad (3.3)$$

Where, CBP $Score_k$ and nCBP $Score_k$ are the carbohydrate and drug-like compound binding propensity scores of grid point *k*; n_i is the number of atoms of a specific amino acid (*i*) within 1.6 Å of the grid point. CBP_i and $nCBP_i$ are the propensities of the amino acid (*i*) to occur in the carbohydrate or drug-like compound binding sites respectively. N is the total number of atoms interacting with the grid point *k*.

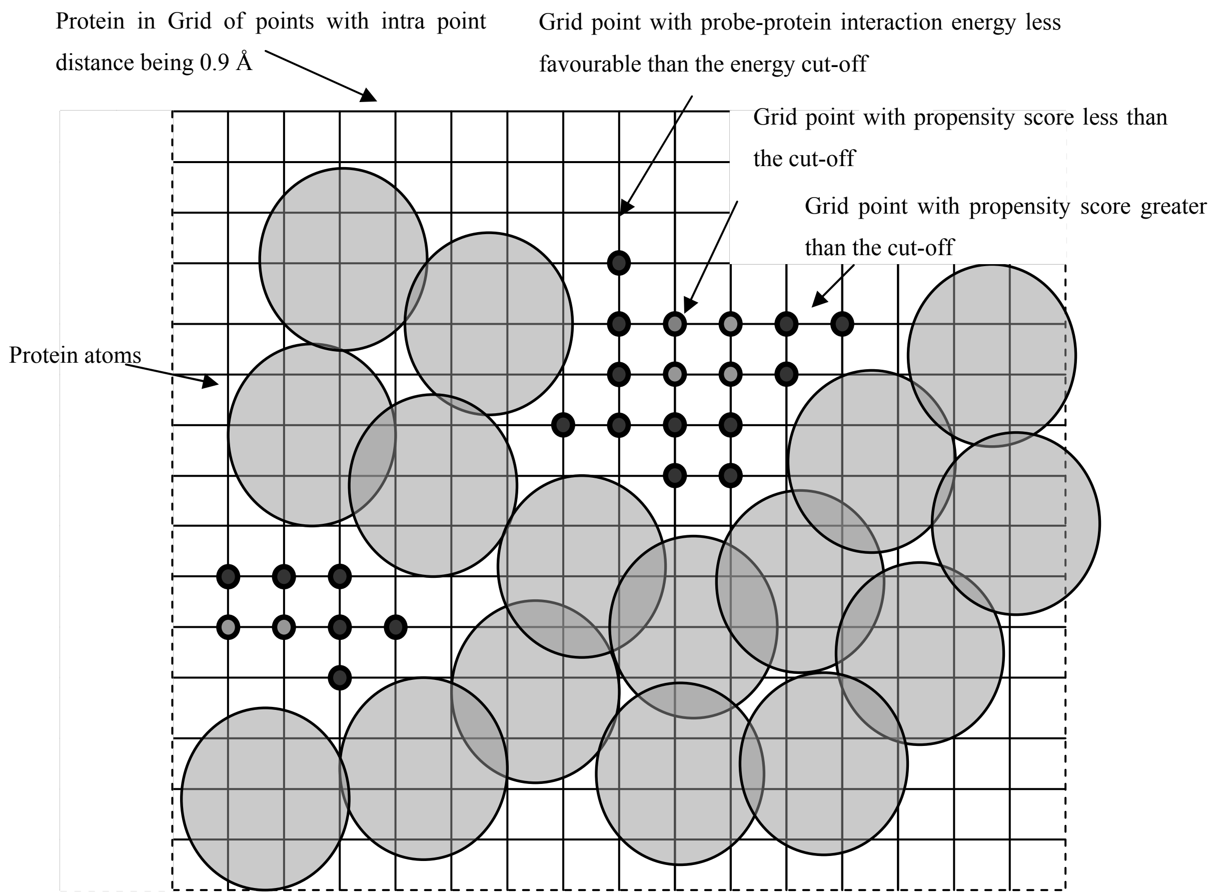

Protein in Grid of points with intra point distance being 0.9 Å
Grid point with probe-protein interaction energy less favourable than the energy cut-off
Grid point with propensity score less than the cut-off
Grid point with propensity score greater than the cut-off
Protein atoms

Figure 3.1: An initial van der Waals energy cut-off is used to retain grid points in energetically favourable binding regions (small filled circles). A carbohydrate binding site occurrence propensity score cut-off is used to remove grid points in regions of low CBP score (small grey circles).

The grid points having a propensity score below a predetermined threshold values are removed (Figure 3.1). The remaining grid points are clustered on the basis of their spatial proximity. A cluster is defined as the group of grid points wherein none of the grid points has its centre farther than 1.0Å from the centre of the nearest grid point. For each of the clusters a sum of CBP and nCBP scores of the grid points are calculated according to;

$$PSSBC_i = \Sigma_{j=1}{}^{n}CBP\ Score_j \tag{3.4}$$

$$\mathrm{PSSBnC}_i = \Sigma_{j=1}{}^{n}\mathrm{nCBP\ Score}_j \qquad (3.5)$$

Where, PSSBC_i and PSSBnC_i are the propensity scores of the site, *i*, to bind carbohydrates and drug-like compounds respectively; n is the total number of the grid points in the cluster. The sites are then subjected to a threshold, whereby the sites having scores less than a given cut-off are removed. The remaining sites are ranked in order of their PSSBC. For each site a differential propensity score (DPS) is calculated as the difference in PSSBC and PSSBnC. DPS represents the overall preference of the predicted site for carbohydrate over drug-like compound ligands. DPS is defined as:

$$\mathrm{DPS}_i = (\mathrm{PSSBC}_i - \mathrm{PSSBnC}_i) \qquad (3.6)$$

Where, DPS_i is the differential propensity of the site *i*. A cut-off for the DPS is applied on the predicted sites in order to optimally identify carbohydrate binding sites over drug-like compound binding sites.

3.3.4 Definition of a true carbohydrate binding site

Since InCa-SiteFinder predicts a number of potential ligand binding sites, a filter is considered whereby a site is identified as a true carbohydrate binding site if it is at least partially occupied by a carbohydrate ligand. Occupancy is defined as the percentage of the space of the predicted site occupied by the ligand atoms. A site was assumed to be a true carbohydrate binding site if the occupancy was greater than 25%. For estimating the volume of the predicted binding site it is placed in a cubic grid of resolution 0.5 Å. The grid points within 2.0 Å of the binding site points are counted and multiplied by 0.125 Å3. The method is based on a program called PDBVolume which has been used to estimate the volume of proteins within a standard deviation 3.3% of the actual volume. (Laurie and Jackson 2005)

3.3.5 Optimisation of InCa-SiteFinder performance

For the assessment of the performance of the InCa-SiteFinder a number of parameters were calculated. Precision is a measure of the correspondence of the predicted site and actual ligand volume. This parameter was calculated during the optimisation of Q-SiteFinder (Laurie and Jackson 2005). It was calculated by taking the percentage of the volume of the predicted binding site that was occupied by the ligand atoms. The second parameter calculated was coverage. It was calculated by taking the percentage of the ligand atoms that are covered by the predicted site. Precision and coverage alone do not depict the actual success of the prediction method. Hence, to give a single parameter for performance assessment precision was multiplied by coverage (PxC) to obtain a single parameter, tau (Γ).

The performance of InCa-SiteFinder to predict the carbohydrate binding site was optimised and evaluated on a training dataset of 50 protein-carbohydrate complexes (none of these belonged to the SCOP superfamilies of the dataset used to derive the amino acid propensities) by 10 fold cross validation. The optimisation of InCa-SiteFinder involved finding optimal cut-off values for the van der Waals energy of interaction and probe propensity score. Members of the training set were divided into 10 groups. For each group the members were classified into two subsets.

First subset had 45 members and was used as an optimisation dataset. During the optimisation process the putative ligand binding sites were predicted for the members of the optimisation set using a range of van der Waals energy cut-offs (from -0.8 to -1.7 Kcal/mol) and a range of probe propensity score cut-offs (0.125 to 1.25). The incremental step for van der Waals energy was -0.1 Kcal/mol and for the propensity score 0.125. For each of the putative ligand binding sites predicted during optimisation process, Γ was calculated for all combinations of van der Waals energy and probe-propensity score cut-off.

The second subset had 5 members and was used as an evaluation dataset for the evaluation of the performance of InCa-SiteFinder under the optimised values of the van der Waals energy and propensity score cut-off derived from the optimisation dataset. The cut-offs giving the best Γ values in the optimisation set were used to predict 99 putative ligand binding sites for each the member of the evaluation set. These were ranked according to their PSSBC score. For evaluation purposes, sites which were occupied by at least 25% of ligand volume were considered as true carbohydrate binding sites.

This process was repeated for all 10 groups, thus predicting the putative ligand binding sites for each evaluation set member once. The results were processed to obtain "Receiver operating characteristic" plots by plotting Sensitivity versus 1-Specificity.

3.3.6 Calculation of sensitivity and specificity

The sensitivity was calculated by dividing the number of correctly predicted sites (true positives, TP) as defined in section 3.2.4, in the top 'm' predicted sites by the total number of predicted carbohydrate binding sites (true positives plus false positives, (TP + FP)), where the value of 'm' is incremented by a factor of 1 each time. False positive rate or one minus specificity is calculated by dividing the number of predicted sites in the top 'm' that have less than 25% occupancy of carbohydrate like ligand (see section 3.2.4). A random prediction would predict 50% true positives in the top x number of sites. The true positive rate and false positive rate are defined as:

$$\text{Sensitivity} = \text{TPR} = (\text{TP})/(\text{TP} + \text{FN}) \qquad (3.7)$$

$$1\text{- Specificity} = \text{FPR} = (\text{FP})/(\text{FP} + \text{TN}) \qquad (3.8)$$

The plotting of true positive rate against false positive rate produces a “Receiver Operating Characteristic” (ROC) curve. The efficiency of a scoring function for ranking the predicted ligand binding sites is reflected in the ROC curve. Area under the curve (AUC) is a measure of accuracy of prediction. AUC for a random predictor is 0.5. If the correlation between the real ligand binding sites and ranked predicted sites based on a given scoring function is positive, the value of AUC is greater than 0.5. If the correlation is negative the AUC value is less than 0.5. For perfect prediction the AUC value approaches the value of 1.0.

3.3.7 Optimisation of InCa-SiteFinder

3.3.7.1 Determination of PSSBC cut-off for classifying a region as a site

A test set of 45 protein-carbohydrate complexes having 45 carbohydrate binding sites was created (Appendix-III) to determine the cut-off value of PSSBC. These complexes had no overlap with the 50 complexes used for optimising the energy and amino acid propensity cut-offs. Using InCa-SiteFinder the top 30 putative ligand binding sites were predicted for each member of the dataset. These sites were ranked in decreasing order of their PSSBC values. The overall success rate for the *j*th ranked prediction was calculated by dividing the total number of correctly predicted true carbohydrate binding sites (NTCBS), as defined in section 3.2.4, at the *j*th rank, by the total number of true carbohydrate binding sites present in the entire database (NTBS). The value of ‘j’ is incremented by a factor of 1 each time to get a series of success rates for all of the 30 ranks. The success rate for rank, *j*, was defined as:

$$\text{Success rate}_j = (\Sigma_{i=1}{}^{n}\text{N}_i\text{TCBS /NTBS})\text{x } 100 \qquad (3.9)$$

Where, N_iTCBS is the number of true carbohydrate binding site correctly predicted for test set member *i*, in rank, *j*. n is the total number of protein-carbohydrate complexes and the NTBS is the number of total true carbohydrate binding sites in the dataset. In addition, an average site score for each rank, *j*, was similarly calculated as:

$$\text{Average Site Score}_j = (\Sigma_{i=1}{}^{n}\text{PSSBC}_i\ )/\text{n} \tag{3.10}$$

Where, PSSBC_i is the propensity score of the i^{th} site to bind carbohydrates and n is total number of complexes in the dataset. A cut-off for the PSSBC_i was determined (from the plots of the average site score and success rate of predictions versus site ranking, see Figure 3.6) such that none of the true carbohydrate binding sites scored less than the cut-off.

3.3.7.2 Determination of differential propensity score cut-off for carbohydrate binding site

A second test set of 40 protein-carbohydrate complexes and 40 complexes of protein-drug like compounds was created (Appendix-IV). This dataset did not overlap with the training dataset (section-4.2.5) or the first test (section-4.3.7.1). The values for the van der Waals energy of probe-protein interaction and the probe propensity score cut-offs of the training set (section-4.2.5) were used to predict the top 30 sites for each member of the dataset. For each of the predicted sites a differential propensity score (DPS) was calculated (section-4.2.3). The success rate of the method was calculated according to the equation 3.9. Average score for the ranked sites were calculated as:

$$\text{Average Site Score}_j = (\Sigma_{i=1}{}^{n}\text{DPS}_i\ )/\text{n} \tag{3.11}$$

This was used to determine an effective cut-off value for DPS which allows differentiation of carbohydrate and drug-like compound binding sites (section 3.4.4).

3.3.8 Dataset for evaluation of the ability of the InCa-SiteFinder to distinguish between the carbohydrate binding sites and drug-like compound binding sites

A third non-overlapping test dataset (Appendix-V) was prepared for the evaluation of the ability of the method to classify the carbohydrate and drug-like compound binding sites. It comprised of 40 protein-carbohydrate complexes and 40 protein-drug-like compound complexes. In order to be included in this dataset, members were not permitted to have SCOP superfamily representatives in the training or previously used test sets (used in sections 3.2.5, 3.2.7.1 and 3.2.7.2). Also no two members of this dataset were permitted to belong to the same SCOP superfamily. This retains two sets; 1) 40 drug-like compound binders which had 58 drug-like compound binding sites and 2) 40 carbohydrate binders which had 40 carbohydrate binding sites. For each of the protein-ligand complexes the top 30 sites were predicted, scored for PSSBC, PSSBnC and DPS. On the basis of DPS the sites were predicted to be either carbohydrate binding or drug-like compound binding. The predictive capacity of InCa-SiteFinder was evaluated in terms of specificity and sensitivity calculated according to equations 3.7 and 3.12.

$$\text{Specificity} = (\text{TN})/(\text{FP} + \text{TN}) \qquad (3.12)$$

3.4 Results and discussion

3.4.1 Amino acid propensity to interact with carbohydrate molecule

A number of statistical analyses were carried out to identify a property that has maximum potential for differentiating various types of ligand binding site (see Appendix I). The profile of amino acid propensities for occurrence in the carbohydrate binding site calculated as a function of solvent accessible surface area (see section-4.2.2) was very different from drug-like compound binding sites (Figure 3.2).

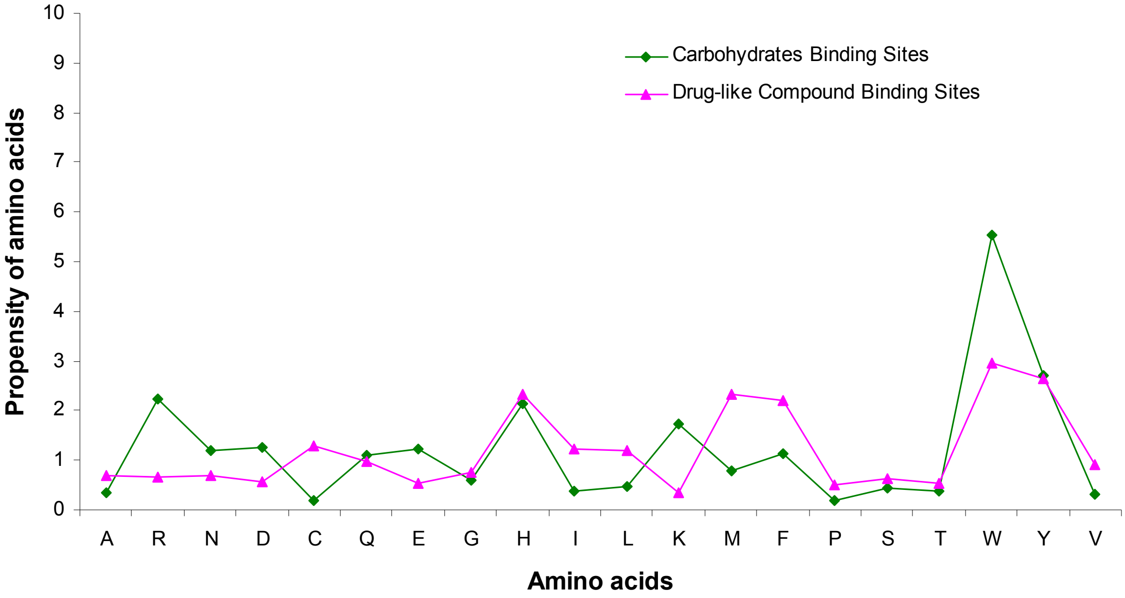


Figure 3.2: The propensity of amino acids to occur in the binding site of carbohydrates and drug-like compounds is shown.

The propensities of arginine, aspartic acid, cystine, glutamic acid, isoleucine, leucine, lysine, methionine, showed maximum differentiation for carbohydrate and drug-like compound binding sites. For carbohydrate binding sites and drug-like binding sites tryptophan has the highest propensity. The propensity profiles of carbohydrate binding sites and drug-like binding site are very different. In a number of studies tryptophan has been identified as an important residue for the ligand binding (Gao, An et al. 2005). However, for carbohydrate binding sites a high occurrence of arginine, lysine, glumatic and asparatic acid, along with the reduced presence of isoleucine, leucine and cystine is seen to be indicative (Figure 3.2).

3.4.2 10-fold cross-validation and optimisation of InCa-SiteFinder

The performance of InCa-SiteFinder was optimised by 10 fold cross validation (see methods). The results are summarised in a 2-dimensional matrix with varying protein-probe van der Waals interaction energy and probe's carbohydrate binding site propensity score cut-off values. The optimal cut-off values are tabulated in Table-3.1. The cut-off values obtained in the optimisation set were used to predict the carbohydrate binding sites for the evaluation set members. The results are plotted in a receiver operating characteristic curves (Figure 3.3).

Set No.	Probe's Propensity Score	Protein-probe interaction Energy (kcal/mol)
1	0.25	-1.0
2	0.25	-1.0
3	0.25	-1.1
4	0.375	-1.0
5	0.125	-1.0
6	0.25	-1.1
7	0.25	-1.0
8	0.25	-1.0
9	0.25	-1.0
10	0.25	-1.0

Table-3.1: The pairs of cut-offs for probe's propensity score and protein-probe vdW interaction which produced best Γ values during 10 fold cross validation.

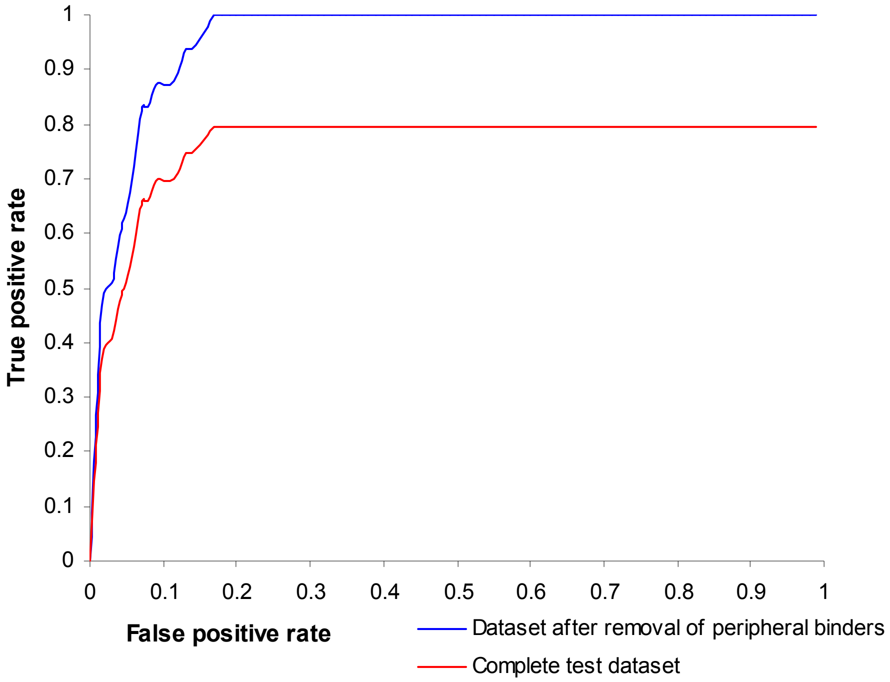


Figure 3.3: Receiver operating characteristic curve illustrating the success of the carbohydrate binding site prediction. The red curve represents the performance of the

method over the entire test dataset. Blue curve represents the performance of the method after peripheral carbohydrate binders were removed from the dataset. These peripheral protein-carbohydrates had only marginal atomic interactions.

The area under the red-curve (AUC) is 81.73%. Even though InCa-SiteFinder reaches a sensitivity of 86% with the specificity of 81%, 7 out of 50 carbohydrate binding sites could not be predicted. Visual examination of the structures revealed that in all of the 7 cases the carbohydrate interaction with the protein receptor was limited to just a few atoms of the ligand molecules which occupied peripheral regions on the protein surface. One example is depicted in Figure 3.4. Such sites are difficult to predict and are of limited interest due to their small size. Functionally important sites generally occur in deep pockets with considerable coverage of the ligand molecule. If these proteins are excluded from the test dataset the performance of InCa-SiteFinder improves. The AUC increased from 81.73% to 95.8%. The method showed absolute sensitivity at 83% specificity. The ROC curve for the reduced test set is shown (blue curve) in Figure 3.3.

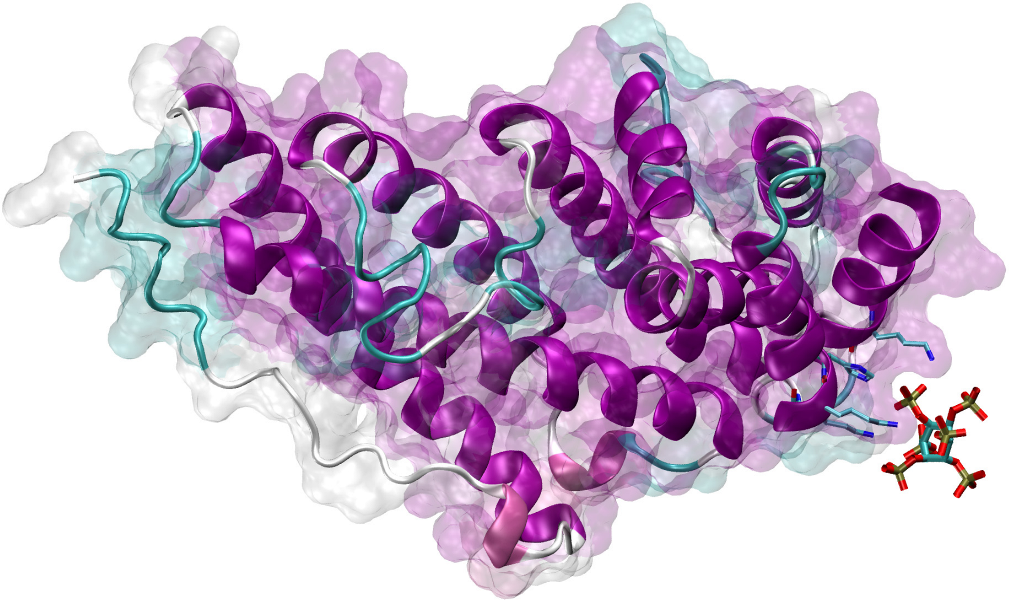


Figure 3.4: Clathrin assembly protein in complex with Inositol hexakisphosphate.

3.4.3 Validation of optimised InCa-SiteFinder

The optimisation matrices calculated for identification of the best possible combination of protein-probe van der Waals interaction energy and the probe's carbohydrate binding propensity, are very similar (see Table-3.1 and Appendix II). An average Γ-matrix was calculated by taking the average of 10 Γ-matrices obtained during 10-fold cross-validation (Figure 3.5). The combination of van der Waals energy cut-off of -1.0 Kcal/mol and propensity score cut-off of 0.25 produced the best result (as seen in the maximum value for Γ (where Γ = P x C) Figure 3.5). These cut-off values were used for the prediction of 45 ligand binding sites for the 45 members of the second validation set (see Section 3.3.7.1). The predicted ligand binding sites were ranked in the decreasing order of their PSSBC (calculated according to equation 3.4) and these values were used as the basis for determining the PSSBC cut-off.

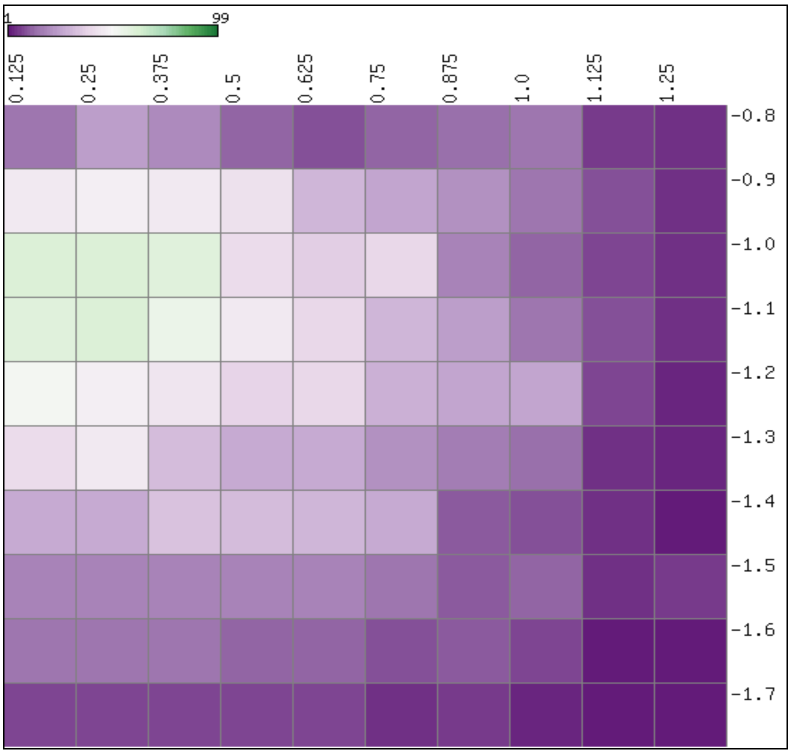


Figure 3.5: Heat map of the average Γ value. The rows represent the variation in the van der Waals energy cut-off and the columns represent the propensity cut-off for amino acids to occur in the carbohydrate binding site.

3.4.4 Determination of PSSBC cut-off value

The ranked predicted ligand binding sites success rate and the average score for each rank were calculated using equation 3.9 and 3.10 respectively. A plot of these values for each rank is given in Figure 3.6. The PSSBC cut-off was determined such that all of the true carbohydrate binding sites scored more than the cut-off. The success rate reaches the value of 100% when the average score of the predicted site is around 60. However, as the test dataset is small, a conservative cut-off value of 30 was chosen, to prevent losing any low PSSBC scoring carbohydrate binding sites.

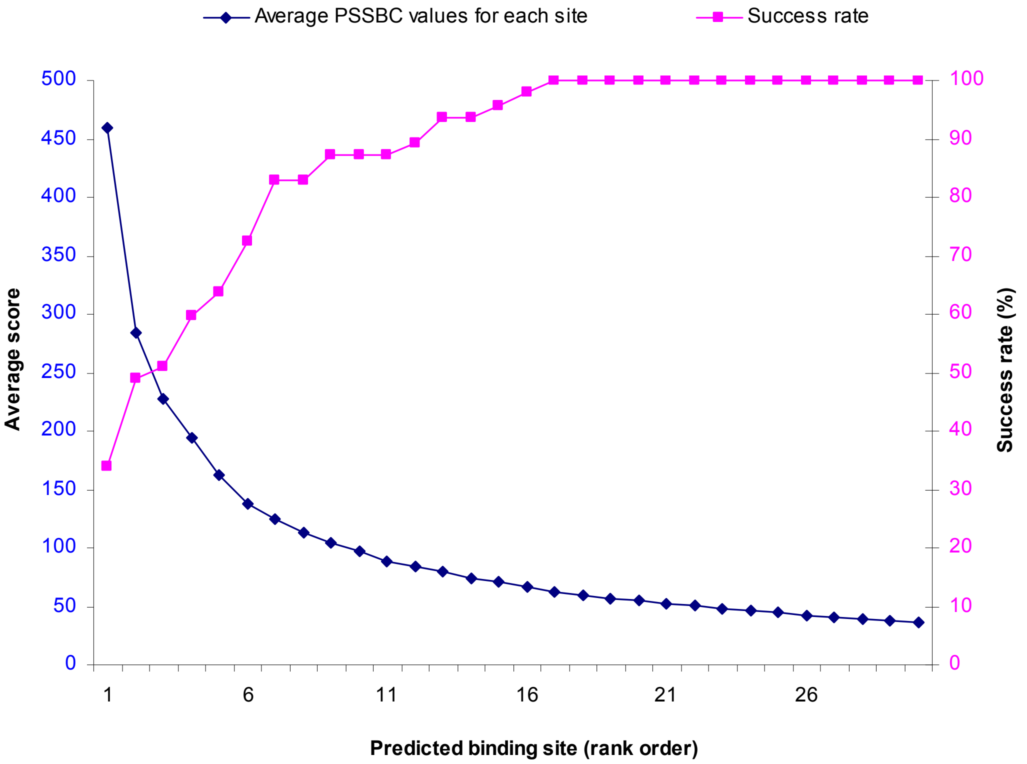


Figure 3.6 Success rate (%) plotted along with average score (average PSSBC values) versus predicted binding site (ranked order).

3.4.5 The importance of differential propensity score for the recognition of carbohydrate binding sites

The success rate (calculated according to equation 3.9) in identifying the potential carbohydrate binding sites was plotted along with the average DPS values (calculated according to equation 3.11) for each of the top 30 sites (Figure 3.7) ranked according to

PSSBC the propensity score of a site to bind carbohydrates. The carbohydrate binding sites have more positive DPS scores and majority of the carbohydrate binding sites are concentrated in the top 4 ranks. Drug-like compound binding sites have more negative DP scores and are concentrated in the last 4 ranks. In the middle region of the DPS range both carbohydrate and drug-like compound binding sites were present.

From the Figure 3.7a and Figure 3.7b two cut-off values were determined. The sites having DPS of more than 10 was considered to be purely carbohydrate binding site and a site was considered to be drug-like compound binding site if its DPS value was less than -20. The sites having DPS values between 10 and -20 were considered to be of dual nature and able to interact with both type of ligands.

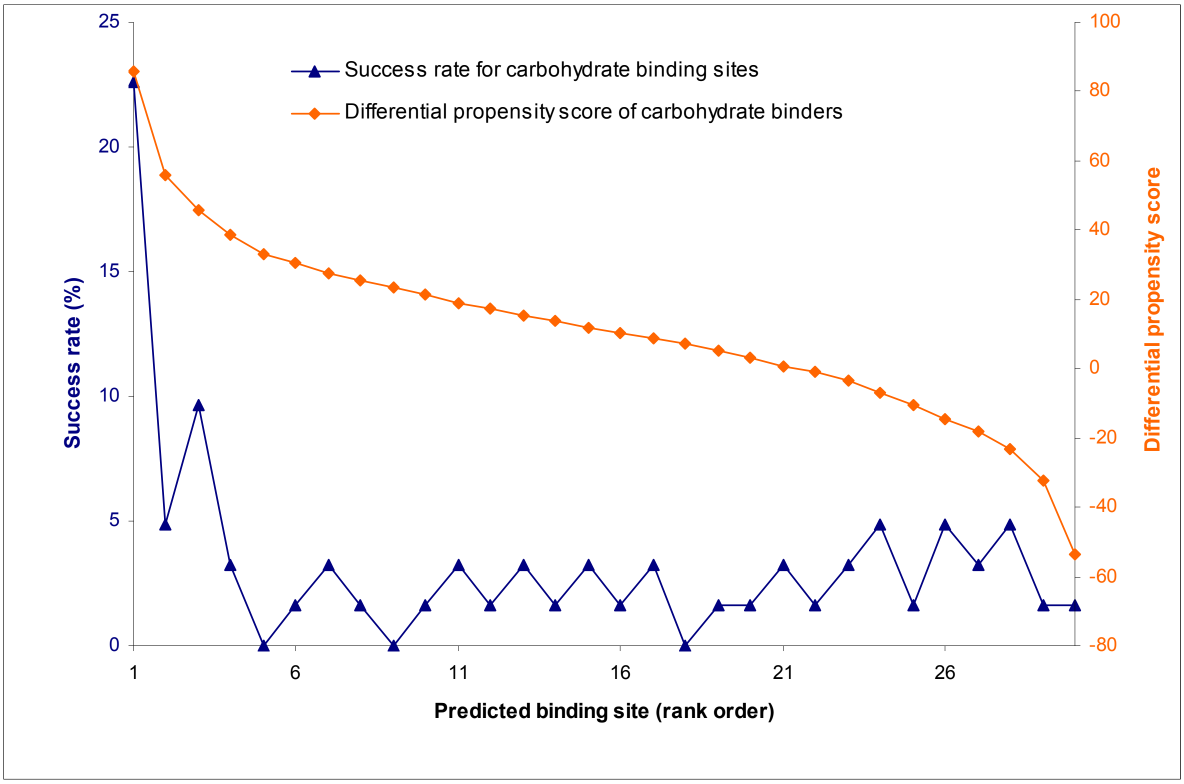


Figure 3.7a Success rate (%) plotted along with DPS for the determination of its cut-off values.

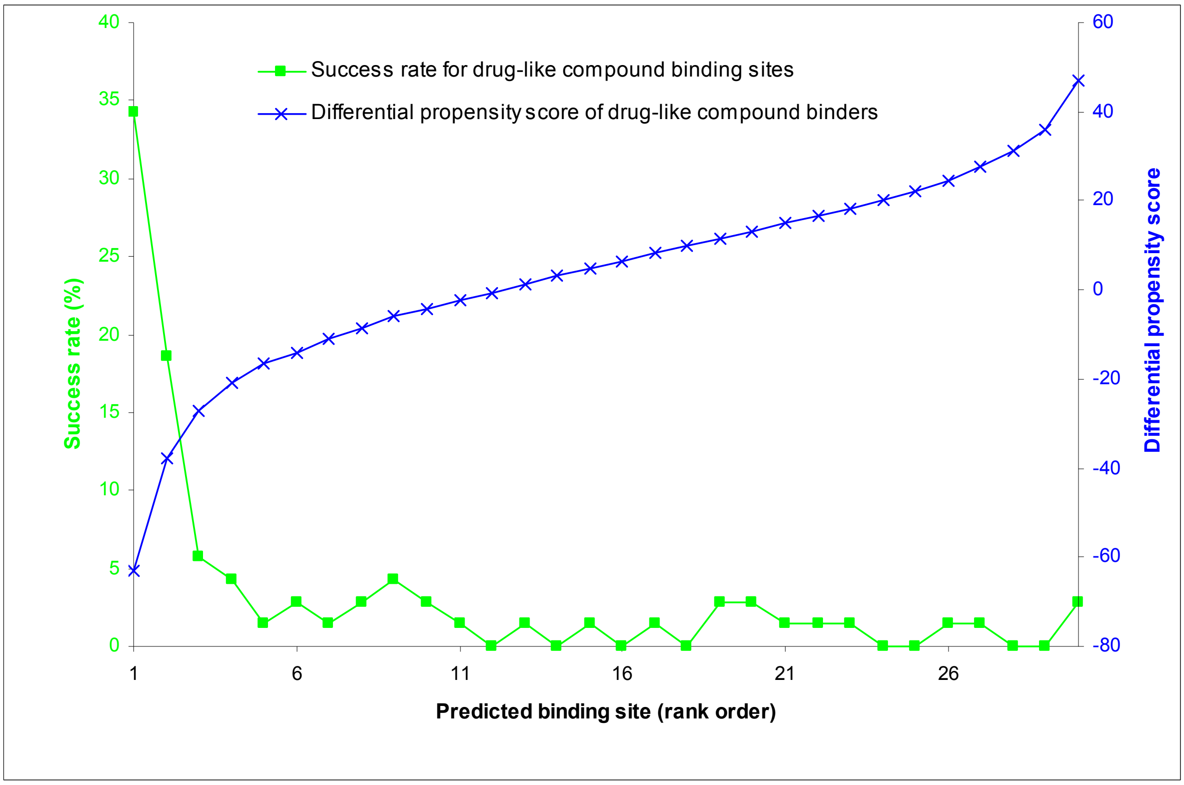


Figure 3.7b Success rate (%) plotted along with DPS for the determination of its cut-off values.

3.4.6 Evaluation of DPS and threshold values

Among the members of the DPS validation dataset (with 58 drug-like and 40 carbohydrate binding sites) out of 98 ligand binding sites 30 are identified as carbohydrate binding sites and 64 site are classified as drug-like compound binding sites. For the remaining 4 proteins (all carbohydrate binders) no site could be identified as true carbohydrate binding sites or drug-like compound binding site. Among 30 sites predicted to be carbohydrate binding sites 29 were true carbohydrate binding sites. The remaining 1 was actually a drug-like compound binding site wrongly classified as a carbohydrate binding site. Among the sites predicted to be drug-like compound binding sites 57 out of 64 predicted sites were true drug-like compound binding sites and 7 were actually carbohydrate binding sites wrongly predicted to be drug-like compound binding sites. The overall specificity of the method was calculated (according to equation 3.12) to be 0.983 and the sensitivity (calculated according to the equation 3.7) was 0.725.

3.5 Some examples of site prediction

Two examples of the correct predictions are shown in Figure 3.8. These sites show the variation in prediction. The precision of predicted carbohydrate binding sites depends upon the ligand and binding site-character. Some of the predicted sites have high precision and high coverage (Figure 3.8a) whereas in other cases the site may have higher coverage of ligand but with less precision (Figure 3.8b). Sites with high precision and low coverage are generally smaller sites that occupy only part of ligand binding site. Such sites are not considered in this method as ligand binding sites because they are removed due to the cut-off of PSSBC, the propensity score of a site to bind carbohydrates.

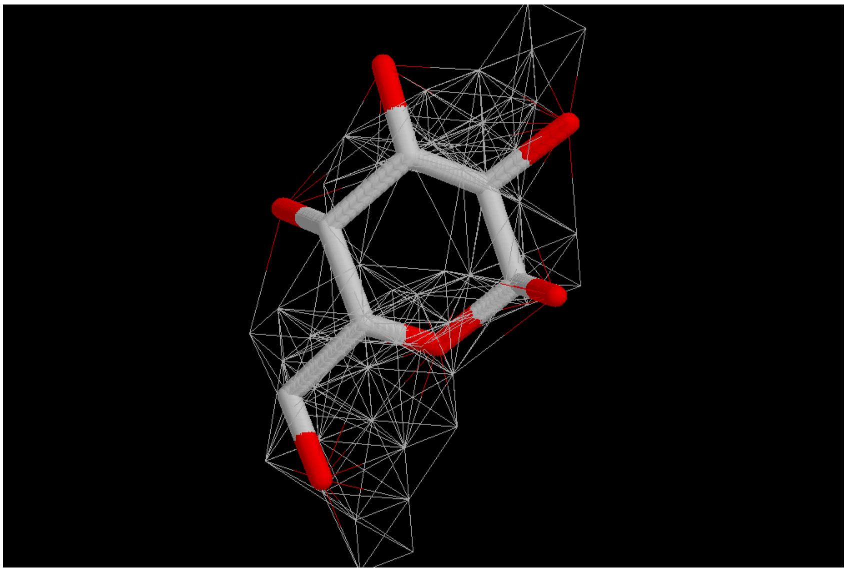


Figure 3.11: (a) Galactose molecule from 5abp covered by the predicted site.

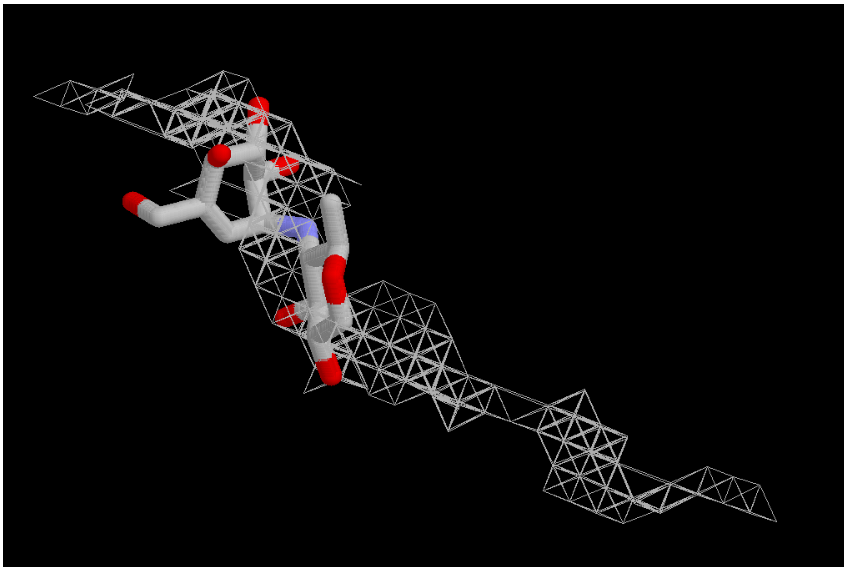


Figure 3.11: (b) Dideoxy-4-amino glucopyranoside from 1hx0 inside the predicted site.

The average volume of all of the sites predicted for the carbohydrates on the surface of the test set was 142.6 Å3. The average volume of the correctly predicted sites is 920 Å3. The average tau value for the correctly predicted carbohydrate binding sites is 324.32.

3.6 Conclusions

We have presented a method, InCa-SiteFinder, for the identification of carbohydrate binding sites by first locating the energetically favourable pockets followed by identifying the regions with high cumulative propensity for binding carbohydrates. It is able to correctly predict the carbohydrate binding site as seen in the ROC plots (Figure 3.3). This ability can be attributed to the application of the combination of using a van der Waals energy of protein-probe interaction developed in the Q-SiteFinder method (Laurie & Jackson, 2005) with the propensity for binding carbohydrate. The carbohydrate binding sites have been shown to be rich in aromatic residues like tryptophan (Gao, An et al. 2005). Presence of such residues in a site increases the potential for van der Waals energy of interaction due to the increase in planar surface area. These residues are also thought to form CH/pi interactions with the carbohydrates by orienting their planar surface for the stacking arrangement (Petrokova, Vondrackova et al. 2005). Also an increased occurrence of residues like arginine, lysine, and glutamic acid coupled with the relative reduced presence of residues like glycine, leucine and isoleucine will give the site a greater potential for making hydrogen bonds. Van der Waals energy alone cannot discriminate between different types of ligands, hence, the use of the propensity scores in combination with protein-probe interaction energetic criteria yields better results. The method is also able to distinguish the carbohydrate binding sites from drug-like compound binding sites with very high specificity (0.983) and sensitivity (0.725).

The value of the differential propensity (DPS) score to distinguish the preference of the predicted site for carbohydrate over drug-like compound ligands is a key factor in the success of InCa-SiteFinder. Sites with high positive values are almost always carbohydrate binder. On the other hand the sites with greater negative values are mostly drug-like compound binding sites. This is valuable information as the carbohydrate binding sites that do not bind drugs-like molecules are easily identified. More remarkable still is the identification of drug-like binding sites with greater success than the identification of carbohydrate binding sites. Though our aim was the prediction of carbohydrate binding sites we have noted the potential use of InCa-SiteFinder in identification of drug-like binding sites, although we have not attempted to optimise the method for this purpose. The sites with dual propensity to bind carbohydrate and drug-like compounds have the DPS values between 10 and -20. It is interesting to speculate that the tool developed here may form the basis for a method that could not only discriminate between different types of functional site, but also facilitate the process of structure-based drug design. In the later case an ability to characterise sites that are amenable to binding drug-like molecules are of great interest for medicinal applications, including blocking protein-protein interactions and for design of competitive inhibitors.

3.7 References

Berman, H. M., J. Westbrook, et al. (2000). "The Protein Data Bank." Nucleic Acids Res 28(1): 235-42.

Bertozzi, C. R. and L. L. Kiessling (2001). "Chemical glycobiology." Science 291(5512): 2357-2364.

Brandley, B. K. and R. L. Schnaar (1986). "Cell-Surface Carbohydrates in Cell Recognition and Response." Journal of Leukocyte Biology 40(1): 97-111.

Gao, S., J. An, et al. (2005). "Effect of amino acid residue and oligosaccharide chain chemical modifications on spectral and hemagglutinating activity of Millettia dielsiana Harms. ex Diels. lectin." Acta Biochim Biophys Sin (Shanghai) 37(1): 47-54.

Ghose, A. K., V. N. Viswanadhan, et al. (1999). "A knowledge based approach in designing combinatorial and medicinal chemistry libraries for drug discovery: 1. Qualitative and quantitative definitions of a drug like molecule." Abstracts of Papers of the American Chemical Society 217: U708-U708.

Gohlke, H., M. Hendlich, et al. (2000). "Knowledge-based scoring function to predict protein-ligand interactions." J Mol Biol 295(2): 337-56.

Laskowski, R. A. (2001). "PDBsum: summaries and analyses of PDB structures." Nucleic Acids Res 29(1): 221-2.

Laskowski, R. A., V. V. Chistyakov, et al. (2005). "PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids." Nucleic Acids Res 33(Database issue): D266-8.

Laskowski, R. A., E. G. Hutchinson, et al. (1997). "PDBsum: a Web-based database of summaries and analyses of all PDB structures." Trends Biochem Sci 22(12): 488-90.

Laskowski, R. A., J. D. Watson, et al. (2005). "ProFunc: a server for predicting protein function from 3D structure." Nucleic Acids Res 33(Web Server issue): W89-93.

Laurie, A. T. R. and R. M. Jackson (2005). "Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites." Bioinformatics 21(9): 1908-1916.

Lipinski, C. A., F. Lombardo, et al. (2001). "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings." Adv Drug Deliv Rev 46(1-3): 3-26.

Malik, A. and S. Ahmad (2007). "Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network." Bmc Structural Biology 7: -.

Murzin, A. G., S. E. Brenner, et al. (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." J Mol Biol 247(4): 536-40.

Petrokova, H., E. Vondrackova, et al. (2005). "Crystallization and preliminary X-ray diffraction analysis of cold-active beta-galactosidase from Arthrobacter sp C2-2." Collection of Czechoslovak Chemical Communications 70(1): 124-132.

Shionyu-Mitsuyama, C., T. Shirai, et al. (2003). "An empirical approach for structure-based prediction of carbohydrate-binding sites on proteins." Protein Engineering 16(7): 467-478.

Tagari, M., J. Tate, et al. (2006). "E-MSD: improving data deposition and structure quality." Nucleic Acids Research 34: D287-D290.

Taroni, C., S. Jones, et al. (2000). "Analysis and prediction of carbohydrate binding sites." Protein Eng 13(2): 89-98.

Viswanadhan, V. N., A. K. Ghose, et al. (1999). "A knowledge-based approach in designing combinatorial and medicinal chemistry libraries for drug discovery: 2. The design of a drug-like library." Abstracts of Papers of the American Chemical Society 217: U708-U708.

Viswanadhan, V. N., A. K. Ghose, et al. (1999). "Prediction of solvation free energies of small organic molecules: Additive-constitutive models based on molecular fingerprints and atomic constants." Journal of Chemical Information and Computer Sciences 39(2): 405-412.

Weis, W. I. and K. Drickamer (1996). "Structural basis of lectin-carbohydrate recognition." Annu Rev Biochem 65: 441-73.

3.8 Appendices

Appendix I: Statistical analysis of Ligand binding sites.

Secondary structure composition of ligand binding sites.

Protein-carbohydrate complexes		
Helix	Beta	Coil
26.12	19.97	53.9

Protein-drug like compound complexes		
Helix	Beta	Coil
30.73	22.83	46.43

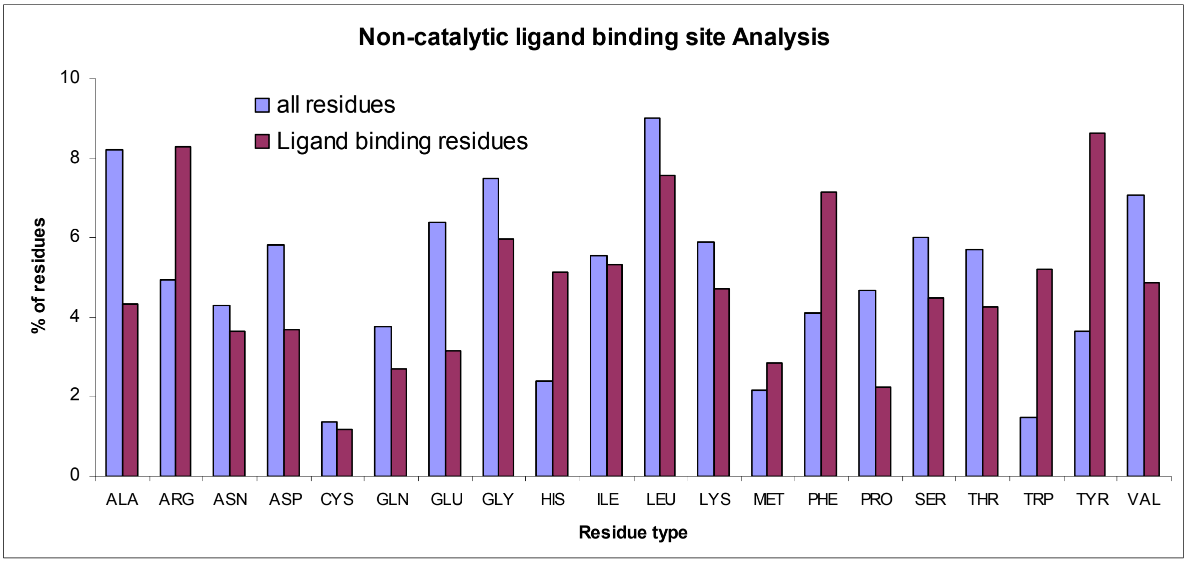
Electrochemically classified amino acid composition of ligand binding sites.

Protein-carbohydrate complexes		
Charged	Polar	Hydrophobic
31.13	21.6	47.26

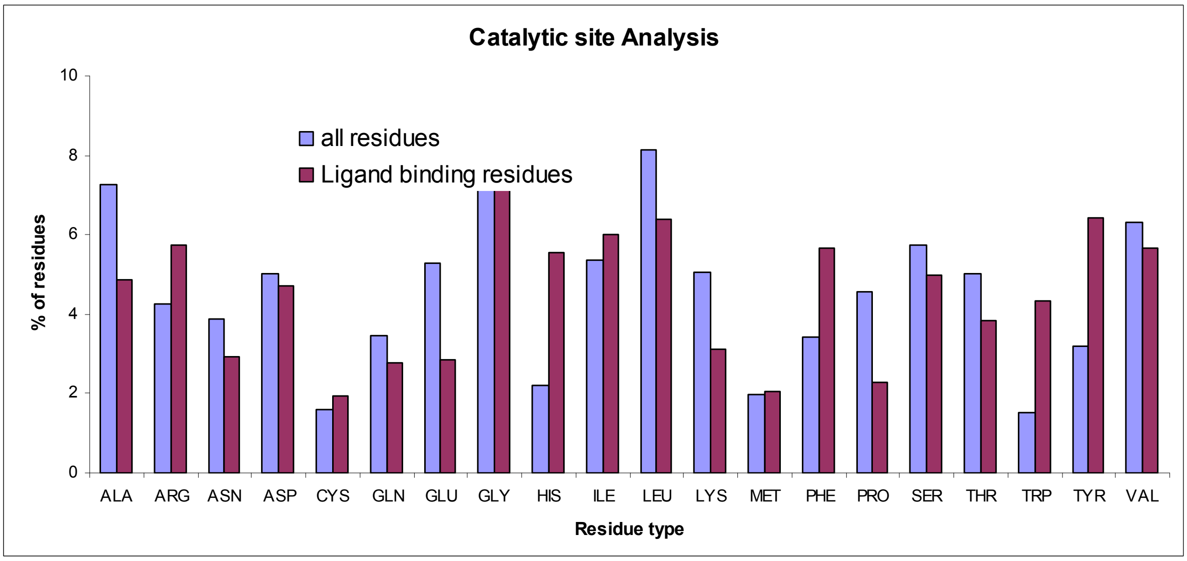
Protein-drug like compound complexes		
Charged	Polar	Hydrophobic
21.94	22.59	55.46

Amino acid frequencies

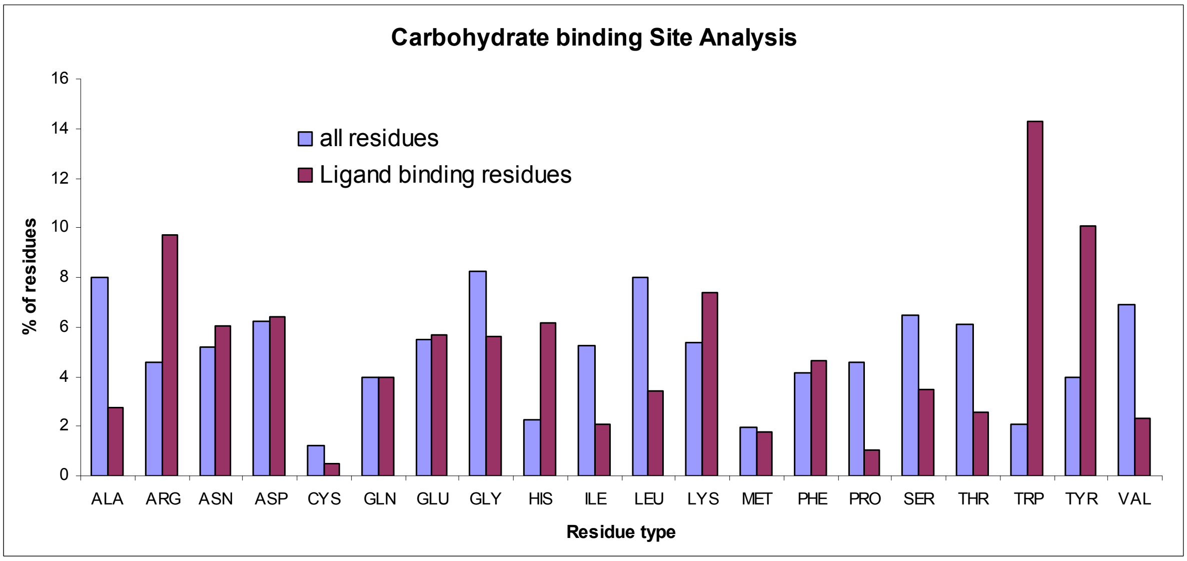
Non-Catalytic ligand binding sites.



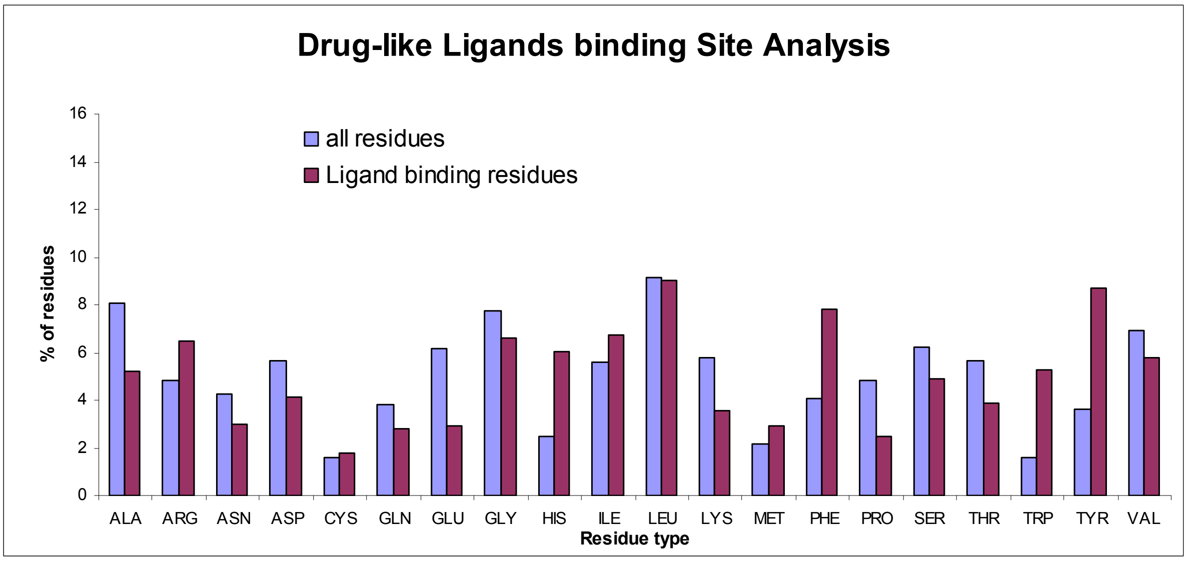
Catalytic ligand binding sites



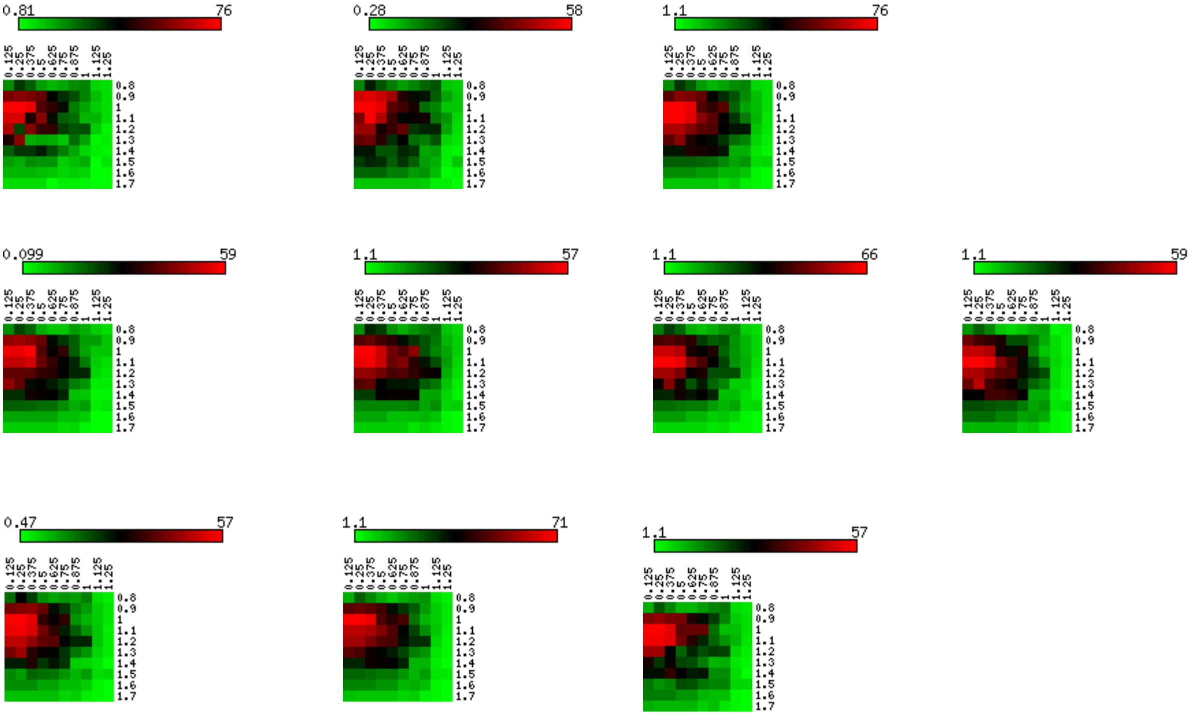
Carbohydrate binding sites.



Drug-like compound binding sites.



Appendix II: Pictures of 10 fold crossvalidation.



Appendix-III

1awb, 1b55, 1bdg, 1bq3, 1btn, 1bwn, 1dbo, 1djx, 1djy, 1dkp, 1dkq, 1e3z, 1ece, 1fao, 1fgy, 1fhw, 1fhx, 1g0h, 1ga2, 1gca, 1gjw, 1gr0, 1gzq, 1h0a, 1h10, 1hfa, 1hg2, 1hx0, 1i82, 1i9z, 1ima, 1imb, 1j8v, 1kwf, 1lbx, 1mai, 1nu2, 1p1h, 1q6c, 1ua7, 1unq, 1w2c, 2bqp, 2nlr, 5abp.

Appendix-IV: Protein-carbohydrate and protein-drug-like compound complex dataset for the determination of DPS cut-off value

Carbohydrate binders:

1a78, 1af6, 1b4d, 1b8o, 1b9z, 1bb5, 1bb6, 1bb7, 1bwn, 1bwv, 1byd,
1byk, 1c39, 1c7s, 1d0k, 1djx, 1dkp, 1dmb, 1e55, 1e8v, 1eou, 1eu8,
1exa, 1f0p, 1f8c, 1f8d, 1f8r, 1fa2, 1fbh, 1fd7, 1fhw, 1fpd, 1g0c,
1g97, 1gjw, 1gpe, 1gs4, 1gup, 1gx4, 1gz9

Drug-like compound binders:

13gs, 19gs, 1a0j, 1a28, 1a4k, 1a9u, 1aax, 1aqb, 1auj, 1az8, 1b11,
1b3d, 1b8y, 1b9s, 1b9v, 1bh6, 1bj0, 1bju, 1bjv, 1bl6, 1bmk, 1bn1
1bn3, 1bn4, 1bnn, 1bnt, 1bnu, 1bnv, 1br6, 1bu5, 1bzc, 1bzj, 1bzs
1c3b, 1c3r, 1c84, 1cbq, 1cbs, 1cet, 1cgk.

Appendix-V: Test set for determining the performance of DPS cut-off value

Carbohydrate binders:

1gi6, 1gii, 1gij, 1gp6, 1gwq, 1h01, 1h60, 1h62, 1h83, 1hxc, 1hy7,
1i7g, 1i8z, 1i91, 1i9n, 1i9o, 1i9q, 1ia2, 1ia3, 1ia4, 1ie9, 1if7,
1if8, 1if9, 1iwh, 1j3k, 1j4h, 1j4i, 1j78, 1j96, 1jcs, 1jgu, 1jho,
1jhq, 1jio, 1jk3, 1jnq, 1jqe, 1jtv, 1k3t.

Drug-like compound binders:

1jr0, 1jvy, 1khz, 1l5k, 1l9n, 1lbf, 1llr, 1lr5, 1lwj, 1m26, 1m6p,
1ms1, 1ms9, 1n9b, 1nb5, 1nmu, 1nnc, 1o45, 1pk9, 1pr4, 1pwb, 1q23,
1qho, 1qi3, 1qnr, 1qw8, 1r82, 1rk2, 1rq5, 1rv0, 1s0j, 1tw3, 1u30,
1ua3, 1ugy, 1umz, 1ur8, 1urd, 1uz8, 1v3c.

PDF-XChange Click to buy NOW! www.docu-track.com

Chapter 4: An Information theory-based Scoring Function for the Structure-based Prediction of Protein-Ligand Binding Affinity

4.1 Abstract

The development and validation of a new knowledge based scoring function (SIScoreJE) to predict binding energy between proteins and ligands is presented. SIScoreJE efficiently predicts the binding energy between a small molecule and its protein receptor. Protein-ligand atomic contact information was derived from a Non-Redundant Dataset (NRD) of over 3,000 x-ray crystal structures of protein-ligand complexes. This information was classified for individual "atom contact pairs" (ACP) which are used to calculate the atomic contact preferences. In addition to two schemes generated in this study we have used a large number of other atom-type classification schemes. The preferences were calculated using an information theoretic relationship of joint entropy. Among 18 different atom-type classification schemes "ScoreJE atom type set2" (SATs2) was found to be most suitable for our approach. To test the sensitivity of the method to the inclusion of solvent, Single-body Solvation Potentials (SSP) were also derived from the atomic contacts between the protein atom types and water molecules modelled using AQUARIUS2. Validation was carried out using an evaluation dataset of 100 protein-ligand complexes with known binding energies to test the ability of the scoring functions to reproduce known binding affinities. A combined SSP/ScoreJE (SIScoreJE) performed significantly better than ScoreJE alone. Also SIScoreJE and ScoreJE performed better than GOLD::GoldScore, GOLD::ChemScore, and XScore.

4.2 Introduction

The success of in silico approaches for structure-based drug design depend on the timely application of the principles governing the dynamics of ligand-protein interactions (Rauh, Klebe et al. 2004). The current approach of docking involves generating favourable ligand orientations in the protein binding site, by sampling conformational space, followed by scoring these by their predicted interaction energy (Klebe 2006). The limitation in the scoring step stems from the time needed to score each potential solution and the level of accuracy required for the calculation of the interaction energy, or at the very least, the correct discrimination of active from inactive compounds. A number of simplified scoring functions have been developed which are fast and easy to apply but provide only moderate levels of accuracy. Hence continued efforts are needed to improve upon existing scoring functions.

Current, scoring functions used to estimate ligand-protein affinity can be classified into three categories: first-principle methods, knowledge-based methods and finally, regression-based scoring functions (Zentgraf, Steuber et al. 2007). Knowledge-based scoring functions are derived from the quantification of frequencies of interacting atomic pairs observed in protein-ligand complexes (Gohlke and Klebe 2001). The process of atomic-pair-interaction-frequency quantification has been based on a number of mathematical relationships. The earliest example of such a function was in the field of protein folding where Boltzmann's law was used to derive the potential of mean force for interacting residue (Tanaka and Scheraga 1976; Hendlich, Lackner et al. 1990; Sippl 1990). Later, similar functions were developed for scoring ligand-protein interactions. Wallqvist et al. (Wallqvist, Jernigan et al. 1995) studied a dataset of 38 complexes, calculating the frequencies of atomic interactions at the protein-protein interface and converted these into an atom-atom preference score using the ratio of fraction of the total interface area contributed by each pair to the product of the fraction of their respective contributions to the surface of respective protein. For a set of 30 proteases-inhibitor complexes, Verkhivker et al. (Verkhivker, Appelt et al. 1995) used the inverse Boltzmann law to develop distance-dependent pair potentials from interacting atoms in combination with conformational entropic (Pickett and Sternberg 1993) and hydrophobic (Sharp, Nicholls et al. 1991) terms. Using this scoring function they could estimate the affinity of HIV-1 proteases for several different inhibitors. SMoG-Score was developed from 109 crystal structures using statistical mechanics (DeWitte and Shakhnovich 1996). Potentials of mean force were derived by Muegge et al. using the inverse Boltzmann law by converting the distance dependent number density of interacting atom pairs from a dataset of 697 protein-ligand complexes into their respective Helmholtz interaction free energies (Muegge and Martin 1999; Muegge, Martin et al. 1999). Mitchell et al. developed BLEEP using a dataset of 820 protein-ligand complexes with hydrogen atoms added (using HBPlus (McDonald and Thornton 1994)) and used the inverse Boltzmann law (Mitchell et al. 1999). A semi-empirical pair-potential for Ne-Ne was used as a reference state. They further derived BLEEP-II by including interactions of protein and ligand with water molecules (explicitly added using Aquarius2 (Pitt and Goodfellow 1991)). Gohlke et al (Gohlke, Hendlich et al. 2000) derived DrugScore using distance-dependent pair-potentials from a dataset of 6026 protein-ligand complexes and incorporated solvent accessible surface area based solvation potentials from a database of 1376 protein-ligand complexes. Cline et al (Cline, Karplus et al. 2002) used an information theoretic relationship of mutual information to quantify information in amino-acid contact potentials for protein structure prediction. They studied the contribution of amino-acid character in terms of hydropathy, charge, disulphide bonding and residue burial to the mutual information.

The Boltzmann law is very useful for determining the interaction energy values from a database of the observed frequencies of joint occurrences. The variation in temperature factors for the protein-ligand atoms (Finkelstein, Gutin et al. 1995) give rise to heterogeneity in the interaction database which complicates the application of the inverse Boltzmann law. However, even though knowledge-based methods are susceptible to the artefacts in data collection they have performed surprisingly well, often better than force-field based scoring functions (Sternberg, Bates et al. 1999; Wang, Lu et al. 2004).

Here the development of a novel knowledge-based scoring function: ScoreJE - derived from the ligand-protein interacting atomic pairs is presented. Our approach differs from the previous scoring functions in two important aspects. Firstly, it uses over 3,000 structurally non-redundant protein-ligand complexes. This is more complexes than used in constructing previous knowledge-based scoring functions, the only exception being DrugScore, which uses a 30% sequence identity cut-off for the creation of the protein non-redundant dataset. Secondly in using the mathematical relationship of joint entropy for deriving the atomic contact preferences it bypasses the problems implicit in the application of the inverse Boltzmann law, eliminating the need for a reference state. These preferences are derived for describing the energetics of short-range atomic interactions. A Single-body Solvation Potential (SSP) is developed using the joint entropy of protein-water atom contact probabilities and is combined with ScoreJE to obtain SIScoreJE (SSP included ScoreJE). These functions were tested for their ability to predict the binding energies of test datasets containing 100 protein-ligand complexes.

The overall aim was to develop a novel knowledge-based scoring function for predicting protein-ligand interaction energy. The main objective was to calculate a non-redundant set of atomic contact preferences for the protein-ligand and protein-water interactions and to use these to develop a scoring function using information theory. A secondary aim was to evaluate the potential of using information theory and new atom type classification schemes (alongside popular atom-type classification schemes currently in use) to optimally describe protein-ligand interactions.

4.3 Methods

4.3.1 Construction of an atom pair contact database

Nearly 30,000 protein-ligand complexes present in PDBSUM (Laskowski, Hutchinson et al. 1997; Laskowski 2001; Laskowski, Chistyakov et al. 2005) with structural information were extracted from the PDB (Berman, Westbrook et al. 2000). From these only protein-ligand complexes having experimentally determined x-ray crystal structures with a resolution greater than 2.5Å were retained. In addition complexes having either covalently bound ligand or involving co-solvents (Appendix-I) or metallic ions or not having a classification in SCOP (version 1.63) (Murzin, Brenner et al. 1995) were removed. A ligand was classified as non-covalently bound to the protein if none of its atom was within the covalent interaction distance. The covalent interaction distance for a specific pair of protein and ligand atom was the sum of their atomic radii plus a 10% tolerance limit.

Only the best resolution complex with a unique ligand name and unique SCOP superfamily are further retained. These comprised a non-redundant dataset (NRD) with only one ligand representative for each SCOP superfamily. Hydrogen atoms were added to these protein-ligand complexes using the QuacPac software (OpenEye). The definition of an atomic contact was taken from DrugScore (Gohlke, Hendlich et al. 2000) in which two atoms are considered to be in contact if the intervening distance between the atoms is less than the sum of their van der Waals radii plus 1Å. The cut-off therefore includes only short-range interactions. For every protein-ligand complex, the interaction information involving protein atoms in contact with atoms of a single molecule of a specific ligand were placed in the Pair Contact Database (PCD). In total there were 1.1 million atomic contacts. Information about atomic orientation was also stored, for each atomic contact pair. This includes the distance (A1), angular (∟A12), and dihedral (A12-123) relationships between the terminal ligand atom (A) and the ultimate (1), penultimate (2) and antepenultimate (3) atoms of the protein (See Figure 4.1 for definition).

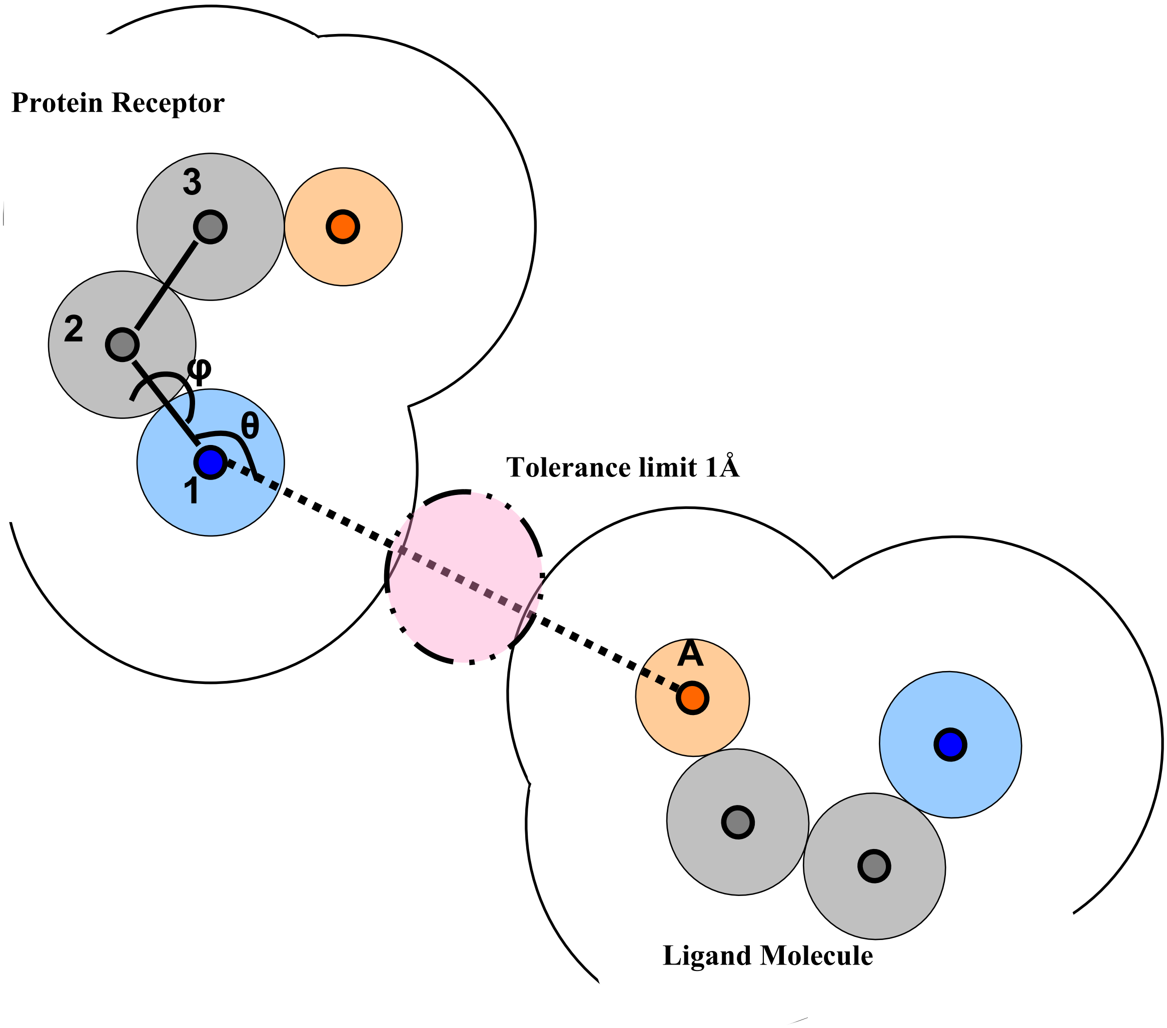


Figure 4.1: Ligand atom A is in contact with Protein atom 1. Protein atom 1 is bonded with non-hydrogen protein atoms 2 and 3. The interactions which are quantified: interatomic A1 distance, planar angle (∟A12) and dihedral (A12-123).

The program Vega (Pedretti, Villa et al. 2002; Pedretti, Villa et al. 2004) was used to convert the atom types in the atom contact pair database from Tripos to those listed in Table 4.1. The use of multiple atom-type sets allows the comparative study of different atomic forcefields, with the intention of seeing which is best suited for the information theoretic approach Table4.2. In addition to the available force field atom-type descriptions, SATs1 and SATs2 (APPENDIX-1a, 1b) were developed.

Classification Type	Size of Classification	Classification Type	Size of Classification
AMBER*	56	MENG*	52
Broto, P et al*	245	MM2*	64
BLEEP*	32	MM3*	120
CFF91*	87	MM+*	40
CHARMM*	79	MMFF*	58
Crippen et al*	82	SATs1**	16
CVFF*	66	SATs2**	24
GRID*	62	TRIPOS*	32
H-bond (Vega template)*	9	Universal*	47

Table 4.1: The different Atom type classification schemes and the resultant alphabet size. * For these atom-types Atom Type Descriptive Language (ATDL) templates were used as provided in the Vega Software. ** These atom type classification systems were designed in Atom Type Descriptive language.

Atom Type Sets	Combination1	Combination2	Combination3	Combination4
AMBER	1867	16463	15460	142434
BLEEP	344	5103	4274	106358
BROTO	3323	20560	19841	131204
CFF91	2785	21683	20450	156505
CHARMM	1876	15332	13978	142591
CRIPPEN	3665	28626	28238	179601
CVFF	2196	18391	17190	148097
GRID	1862	15060	14575	132136
HBOND	51	1028	691	46294
MENG	918	8617	7523	131492
MM2	1213	11093	9932	126356
MM3	1599	13086	11884	125786
MMFF	1410	13395	12440	136650
MM PLUS	1040	10012	8614	129607
SATs1	138	2697	2053	101157
SATs2	233	4568	3573	129565
TRIPOS	380	5300	4341	113238
UNIV	404	5656	4758	108785

Table 4.2: Increase in interacting atom-atom combinations: Combination1 – When only atomic character are considered, Combination2 – When interatomic A-1 distance and atomic character is considered, Combination3 – When dihedral (A12-123) and atomic character is considered, Combination4 – When interatomic A-1 distance, planar angle (∟A12), dihedral (A12-123) and atomic character is considered.

4.3.2 Calculation of atomic contact preferences

Mutual information and joint entropy are indicative of the extent to which the distributions of two variables are related. While mutual information is a measure of mutual dependence of two variables (Reza 1994), joint entropy is amount of uncertainty associated with two variables (Shannon 1948). Mathematically:

Mutual information is defined by:

$$I(X,Y) = \Sigma_x\Sigma_y(P(x,y)\log[P(x,y)/\{P(x)P(y)\}]) \tag{4.1}$$

Where, p(x,y) is the joint probability distribution function of X and Y, and P(x) and P(y) are the marginal probability distribution functions of X and Y respectively (Shannon 1948) (where, X and Y are the ligand and protein atom types respectively in an interacting atom-atom pair).

Joint entropy is defined by:

$$H(X,Y) = -\ \Sigma_x\Sigma_y(P(x,y)\log[P(x,y)]) \tag{4.2}$$

For a given pair of interacting atoms the value of “$P(x,y)\log[P(x,y)/\{P(x)P(y)\}]$” and “$-P(x,y)\log[P(x,y)]$” were respectively considered as the contribution of an individual pair of atoms towards the obtainable amount of mutual information and joint entropy. A complete set of these pair-wise contributions (termed as MI-coefficients and JE-coefficients) for all the atom-atom contact pairs forms the ensemble of atomic contact preferences between a protein and a ligand in the complexed state. Mathematically: MI-coefficients and JE-coefficients for an atom-atom pair were respectively defined as:

$$\mathrm{MIcoeff}(X,Y) = P(x,y)\log[P(x,y)/\{P(x)P(y)\}] \tag{4.3}$$

$$\mathrm{JEcoeff}(X,Y) = -P(x,y)\log[P(x,y)] \tag{4.4}$$

The coefficients are applied to the set of atomic contacts between a specific protein-ligand complex to obtain the amount of mutual information or joint entropy for that complex. For a protein-ligand complex the sum of coefficients associated with all atom contact pairs was considered as the ScoreMI and ScoreJE respectively:

$$\mathrm{ScoreMI}(P{:}L) = \Sigma_x\Sigma_y P(x,y)\log[P(x,y)/\{P(x)P(y)\}] \tag{4.5}$$

$$\mathrm{ScoreJE(P{:}L)} = -\Sigma_x\Sigma_y \mathrm{P(x,y)log[P(x,y)]} \tag{4.6}$$

Where x and y are the protein and ligand atom types respectively in an interacting atom-atom pair.

4.3.3 Generation of a protein-water contact database and atomic solvation-desolvation measures

The preference of protein atoms to make contact with water atoms was calculated using the same approach as outline above. Coordinates of water molecules were obtained by modelling water molecules on the protein surface using the Aquarius2 software (Pitt and Goodfellow 1991). For a dataset of 999 proteins hydration shells were generated around each protein. QuacPac was used for the addition of protons and Vega was used to convert the atom-types to the correct format from which protein atom-water contact pairs were extracted. These contacts were used to derive SSPs using the following relationship:

$$\mathrm{SSPcoeff(X)} = -\mathrm{P(x,H2O)log[P(x,\ H2O)]} \tag{4.7}$$

SSP coefficients for protein atom-types were added to ScoreJE coefficients to obtain SIScoreJE.

$$\mathrm{SIScoeff(P{:}L)} = \Sigma_x\Sigma_y\mathrm{(JEcoeff(X,Y) + SSPcoeff(X))} \tag{4.8}$$

Where, SSP(X) is the Single-body solvation potential for X protein atom type. SIScoreJE comprised of (JEcoeff(X,Y) + SSPcoeff(X)) values.

4.3.4 Protein-ligand test set

The performance of ScoreJE and SIScoreJE to predict the binding energy was evaluated on a dataset of 100 protein-ligand complexes. None of these was a member of our training dataset. Some of the complexes were obtained from a dataset of 205 protein-ligand complexes , others were taken from scorpio (Ladbury et al 2003) and bindDB (Bader, Donaldson et al. 2001). As the scores developed here consider only protein-ligand complexes, nucleic acid-ligand complexes were also excluded from the test set. The ligands in this dataset were considerably diverse in terms of number of rotatable bonds (0-24), molecular mass (71-824 amu), number of heavy atoms (7-62) and number of aromatic rings (0-4).

Since the training dataset has a single SCOP superfamily representative for a specific ligand in complex with a protein, some of the SCOP superfamilies have more members than others. Intuitively the SCOP superfamilies having a larger number of member proteins could introduce an element of over training; however, since the ligand population has no repetition the effect is minimal. Moreover, in order to eliminate all possible effects of due to SCOP superfamily representation in the training set a “tailor made” training dataset was created for each ligand in the test set. For each test set member the contacts were derived from the NRD (see above) by removing all those proteins belonging to the same SCOP superfamily as the test dataset member. These were then used to derive scoring functions which were unique for each test set member, removing any possible bias due to protein evolutionary relatedness (as defined by SCOP) during cross-validation.

A dataset of 50 protein-ligand complexes was used to determine the efficiency of the ScoreJE in identification of near-native configurations produced during docking. Only those protein-ligand complexes that did not have presence of cofactor or metallic ions in the ligand binding site were considered. The protonation states of ligand and the protein molecules were determined by OpenEye software (QuacPac). For each of the test set member 100 docking solutions were obtained using default parameters of GOLD program.

4.4 Results

4.4.1 Choice of scoring function

Scoring matrices for ScoreMI and ScoreJE from the Pair Contact Databases (PCDs) of 18 different atom-type datasets were calculated according to equation-3 and equation-4 respectively (see methods). For the comparative evaluation of the two scoring functions (ScoreMI and ScoreJE) 100 protein-ligand complexes were used, for which the experimental binding energies were known (complete list is given in appendix-II). For each member of the test set, full cross validation was performed eliminating any possible bias due to protein evolutionary relatedness in the training set (see methods). The ScoreMI and ScoreJE for protein-ligand complexes were calculated by summing up the MIcoeff and JEcoeff assigned for each of atomic contact pair in the interaction database (see methods equation 5 and 6). The correlation coefficients (R2 values) between scores thus calculated (ScoreMI and ScoreJE) and the experimental binding energies were obtained via linear regression all 18 different atom-type sets (see Figure 4.2).

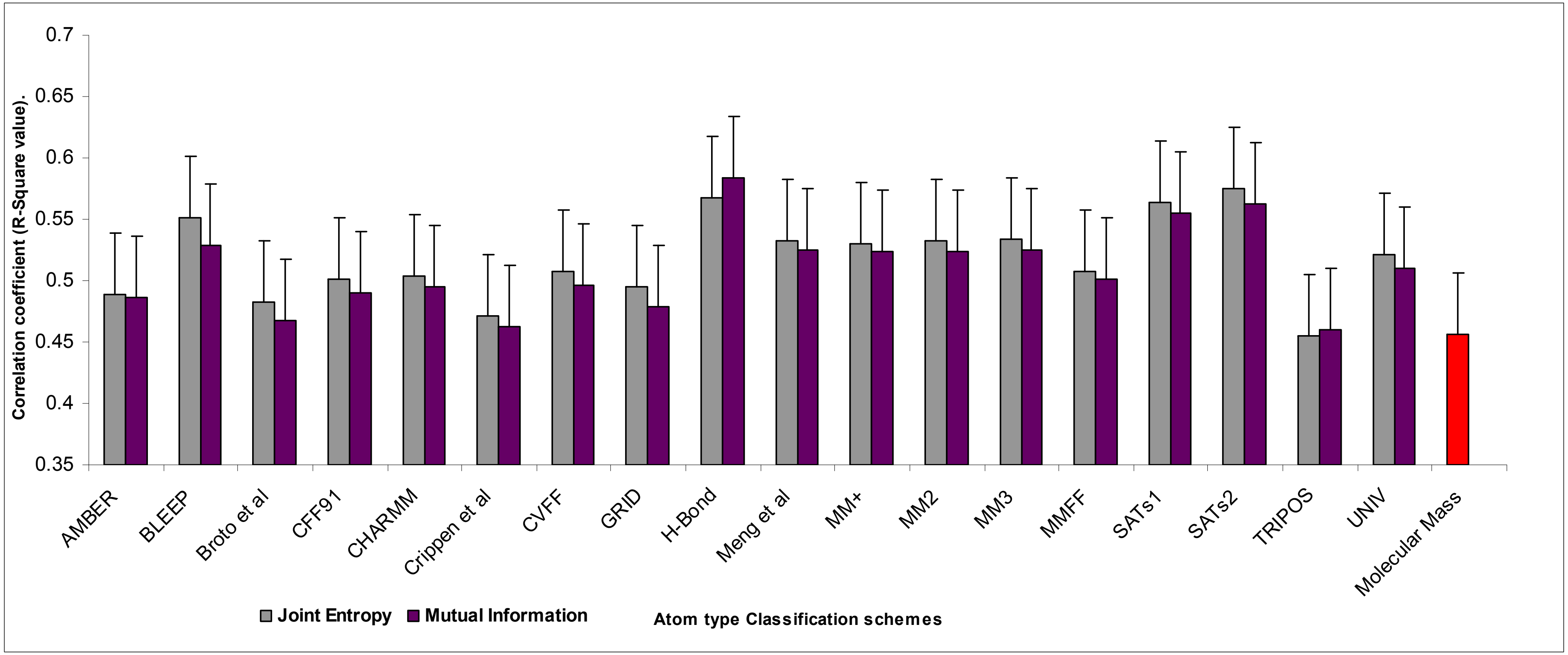


Figure 4.2: Comparative study of correlation coefficients (R2) between experimental binding energies and the scores calculated using ScoreMI (Mutual Information) and ScoreJE (Joint Entropy) during cross-validation for 18 different atom-type schemes.

ScoreJE performs either better or equal to ScoreMI in all but two cases and gives the better correlation between calculated score and binding energies. In the subsequent work ScoreJE was adopted as the basal scoring function of choice.

4.4.2 Choice of scoring parameters

The best possible combination of atom-type set and orientation index was determined. The orientation indices include the distance the between interacting atoms, the angle of ∟A12 and the dihedrals of A12-123 (see methods). These indices are continuous for protein-ligand interacting atom pairs therefore the orientations were binned into discrete values with distances rounded to one decimal place, planar angles and dihedrals were binned in intervals of 10°. ScoreJEcoefficients were then calculated for each of orientation index for each of the 18 atom type datasets and the test set scored. The correlation coefficient (R2) values between the calculated score and experimental binding energies are given in Figure 4.1.

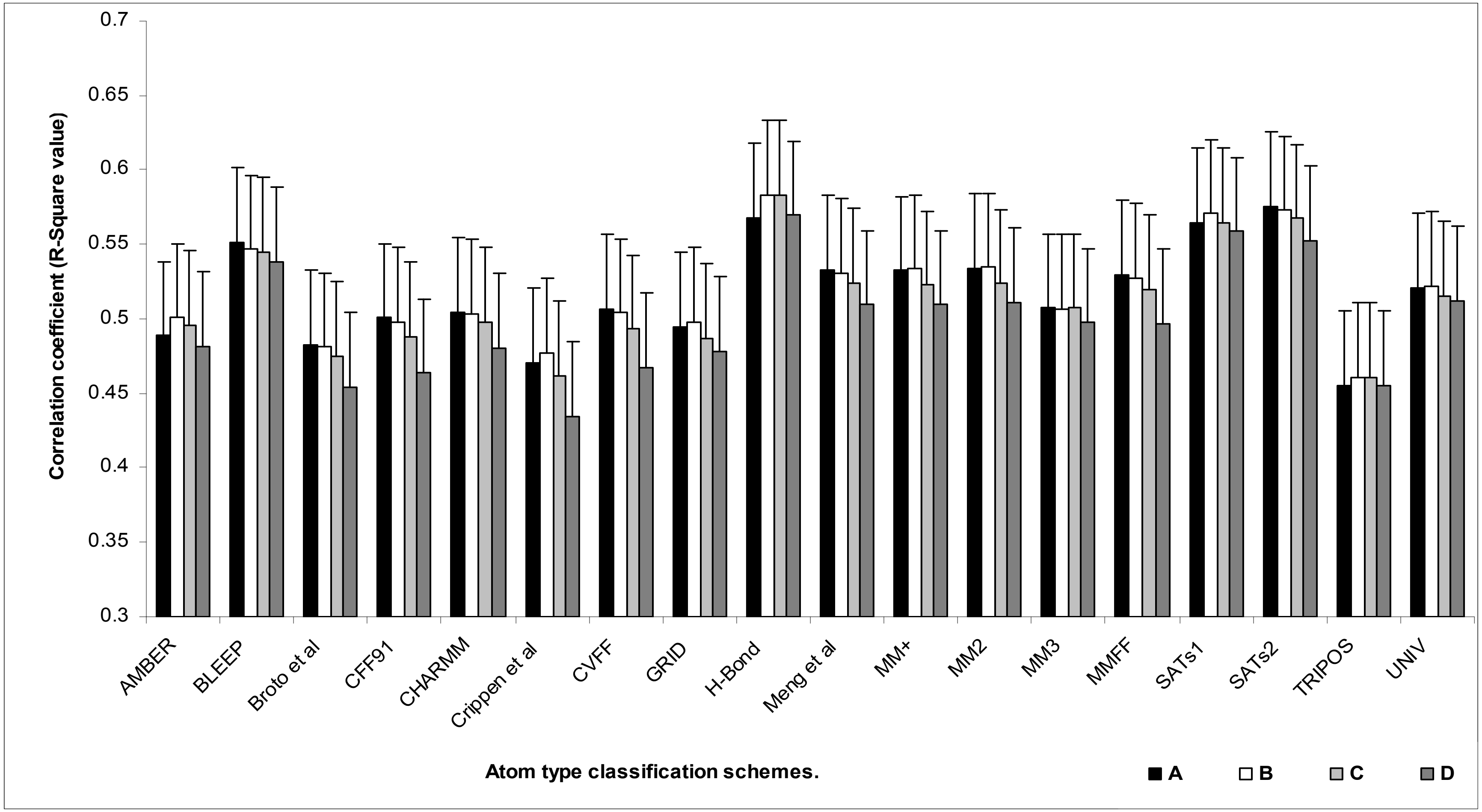


Figure 4.3: Comparative study of correlation between experimental binding energies and the scores calculated using ScoreJE (Joint Entropy) when A – only atomtype contacts are considered; B – atomtype contacts and the interatomic distance is considered (A-1 from Figure 4.1); C – atomtype and dihedral angle (A12-123) is considered; D – atom type, interatomic distance (A-1), planar angle (A12), dihedral angle (A12-123) are considered.

As can be seen, the ScoreJE calculated for atom-type pair alone for SATs2 performed best. Also, in most of the cases atom-type pair alone performed better than the combinations of atom-type, intervening distances, angles and dihedrals. Even though the number of descriptors for each atom-type classification scheme increases for the atom-type orientation index combinations of B, C, and D, the performance in terms of the correlation coefficient between calculated scores and experimental binding energies remains almost the same relative to atom-atom pairs alone.

4.4.3 Inclusion of Solvation effects and SIScoreJE

ScoreJE and SIScoreJE were obtained for atomic contact pairs for the SATs2 atom-type classification. The calculated scores for the 100 protein-ligand complexes of the test dataset are plotted against the experimentally determined binding energies in Figures 4.4 and 4.5 for ScoreJE and SIScoreJE respectively.

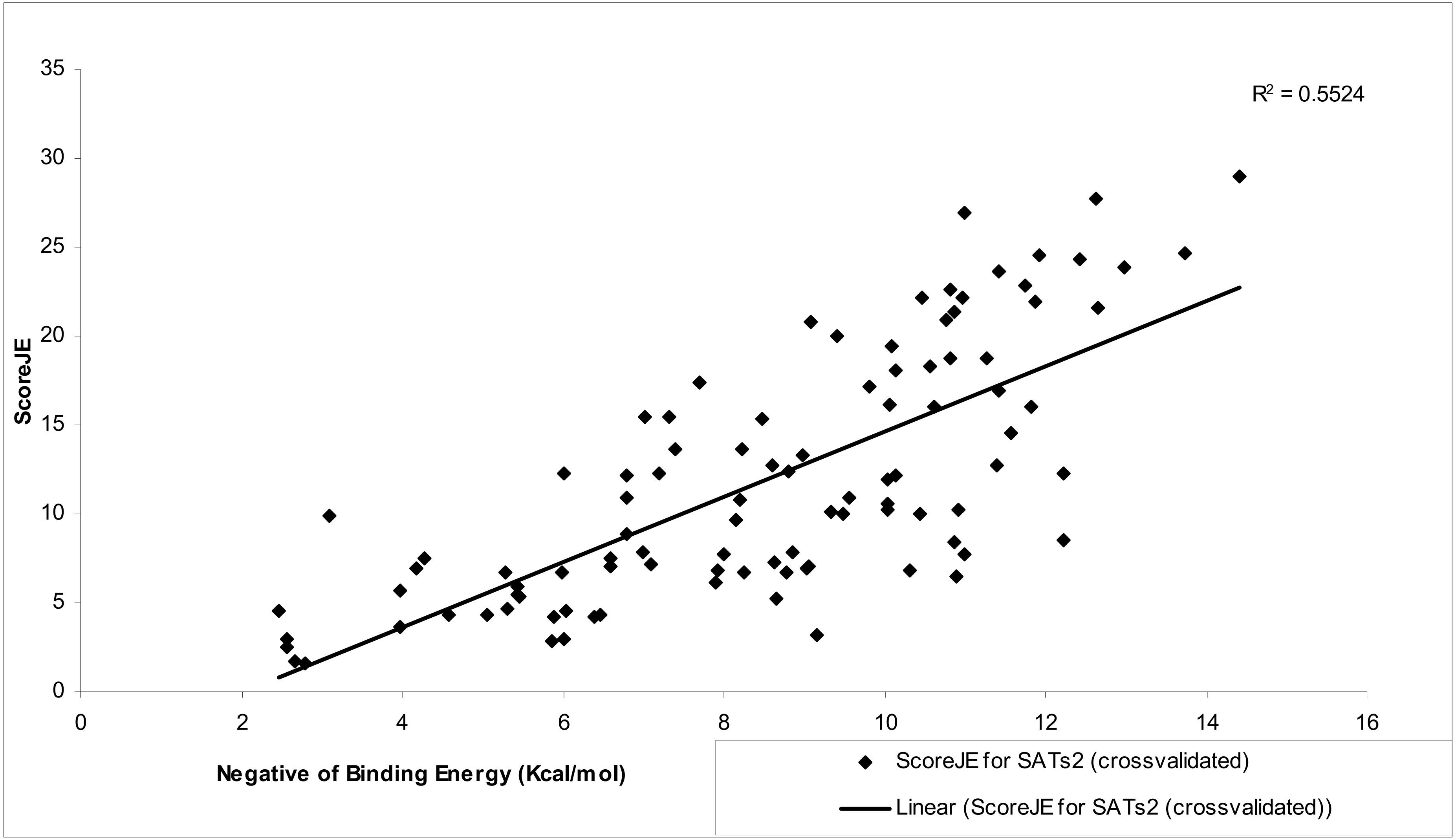


Figure 4.4: ScoreJE vs. Binding energy plot for 100 protein-ligand complexes.

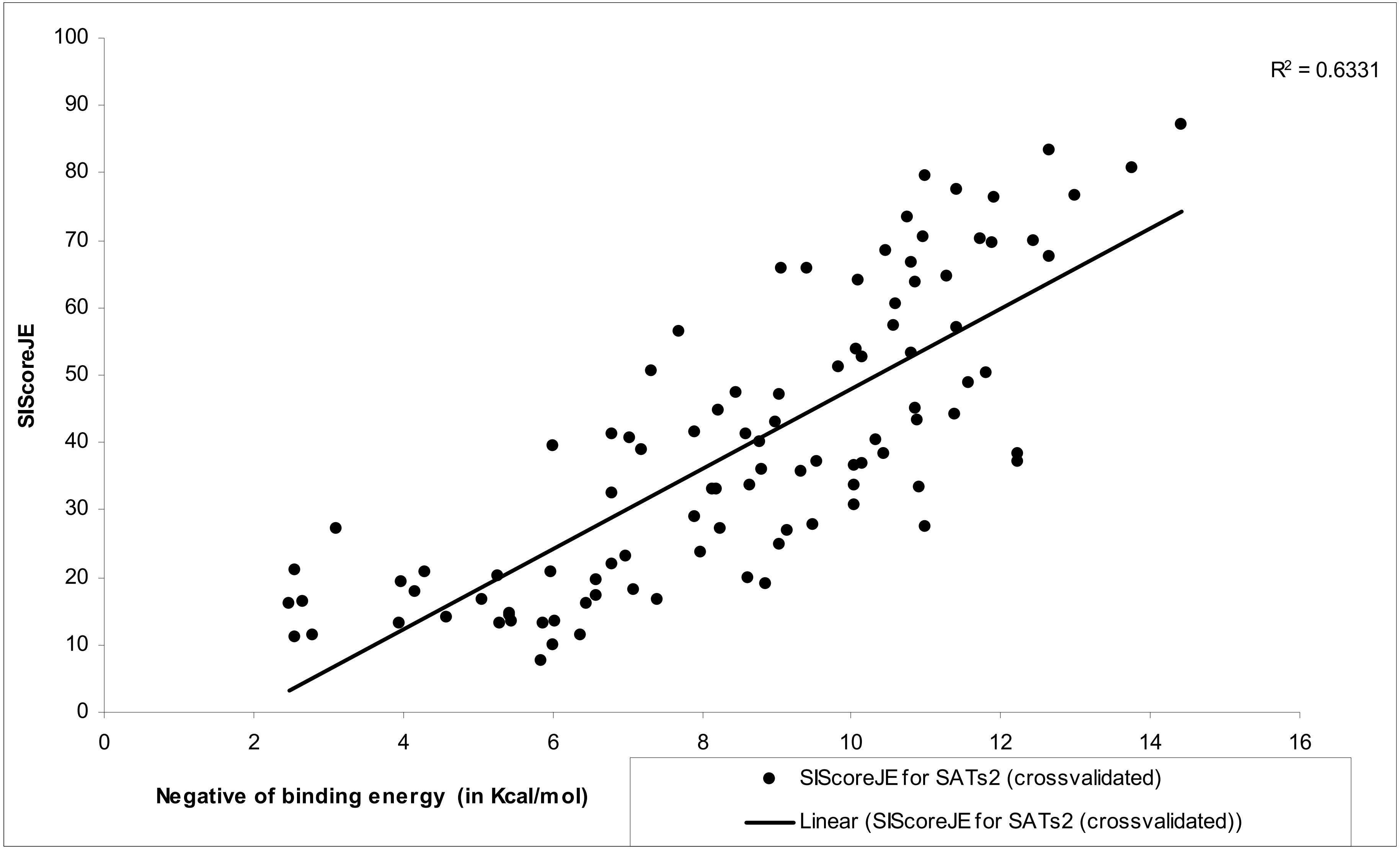


Figure 4.5: SIScoreJE vs. Binding energy for 100 protein-ligand complexes.

ScoreJE includes only protein-ligand direct interactions whereas the SIScoreJE also includes the indirect interactions that take place with solvent molecules modelled using Aquarius2 (see methods). The overall SIScoreJE scores correlated slightly better with the experimental binding energy than those with ScoreJE. In order to understand the influence and utility of SIScoreJE for various functional classes of proteins the above dataset is further subdivided into acid proteases (31), serine proteases (12), carbohydrate binding proteins (16) and miscellaneous groups (41). A summary of results is given in Figure 4.6.

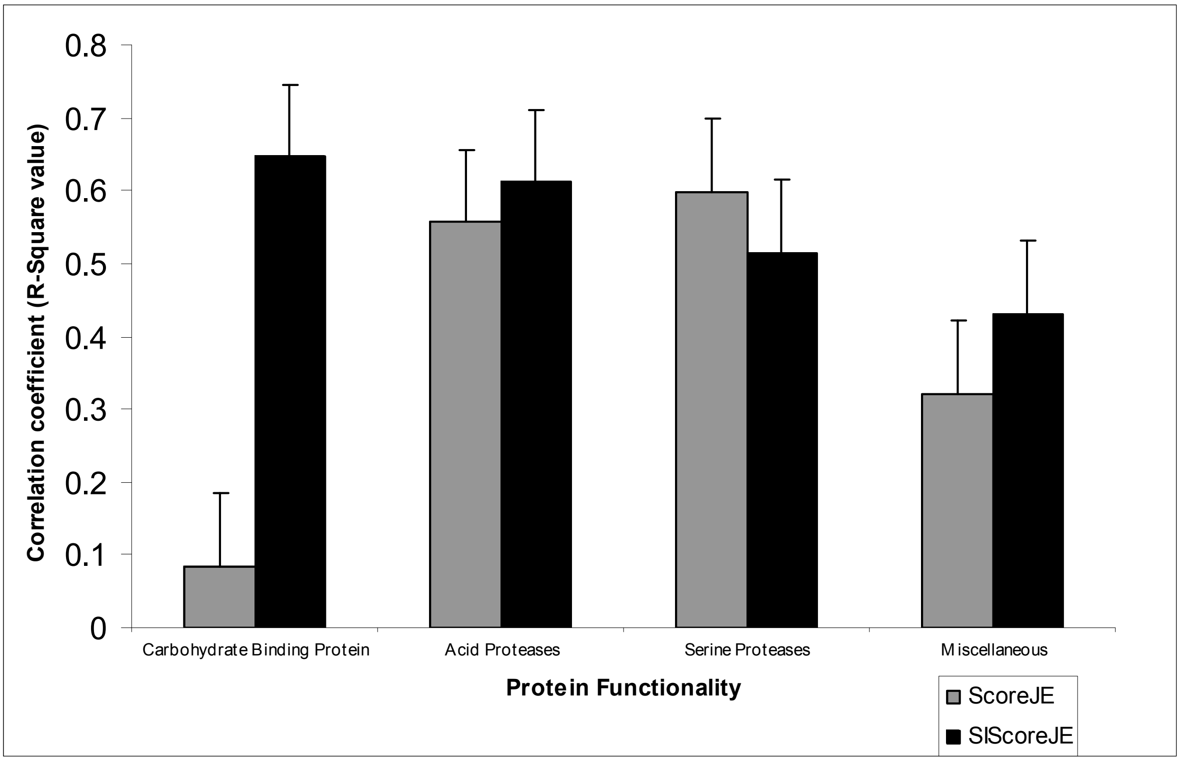


Figure 4.6: The correlation coefficient between experimental binding energies and ScoreJE (grey) and experimental binding energies and SIScoreJE (Black) for different functional classes of protein-ligand complexes.

Predicting the binding energies of carbohydrates to their cognate binding proteins has previously been reported to be very problematic (Taylor and Burnett 2000), however, it is here that SIScoreJE performs significantly better than ScoreJE. For the serine proteases ScoreJE performs slightly better than SIScoreJE. A majority of protease binding sites have a predominance of hydrophobic character. Understandably, SIScoreJE is therefore less likely to have a significant influence in these complexes, as solvent is excluded from the ligand binding site to a greater extent in these cases and therefore less likely to play a dominant role. Whereas the carbohydrate binding sites are generally well solvated and hence, the correlation of SIScoreJE scores for this class improves significantly over the basal scores of ScoreJE.

4.4.4 Comparison of different scoring functions

The ability of ScoreJE and SIScoreJE to correlate with binding energy was compared against GoldScore, ChemScore and X-Score(Wang, Lu et al. 2004). The test set of 100 protein-ligand complexes were re-scored using GOLD::GoldScore, GOLD::ChemScore, X-Score. While GOLD::GoldScore provides a measure of fitness for the ligand-protein complex it was not found to be very effective in predicting the G of interaction in□binding energies. The GOLD::ChemScore estimates the addition to the fitness score. X-Score gives three different scores. A consensus score (an average of the three as suggested in (Wang, Lu et al. 2004)) was taken as a measure for the binding energy. The degree of correlation between the scores and experimental binding energy for the scoring functions is given in Figure 4.7.

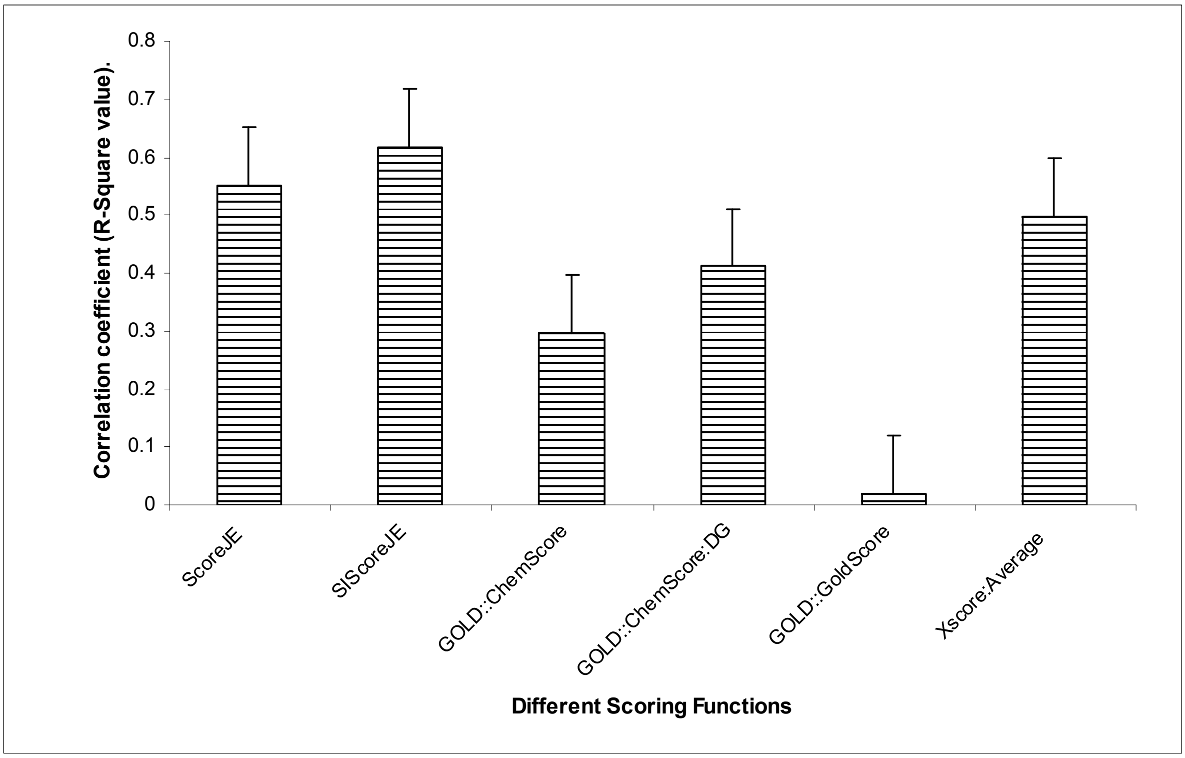


Figure 4.7: comparative analysis of the scoring functions.

X-Score performed better than GoldScore and ChemScore which is consistent with the results obtained by Wang et al (Wang, Lu et al. 2004). However ScoreJE and SIScoreJE have the best correlation between predicted score and experimental binding energy.

4.4.5 Identification of Near-Native Configurations

The effect on the performance of our scoring functions for ranking the poses generated during docking runs, when information on orientation indices is included in training of the scoring functions was further analysed. The mutual interaction preferences in scoring function were recalculated by considering the interacting atom-type identities along with information about the inter-atomic distances (version 2); planar angle (version 3); dihedral angle (version 4) and a combination of distances, planar angle, dihedral angle (version 5). For docking the GOLD(Verdonk, Cole et al. 2003) program was used to generate 100 docking solutions for each of the 53 protein-ligand complexes in docking test set (see methods). These poses were ranked according to the RMSD from the native ligand conformation present in the crystal structure. Conformations which had RMSD less than 2Å were considered as positives and the rest as negatives. The poses were also evaluated using our scoring functions and receiver operating characteristic (ROC) curves were plotted (Figure 4.8). Ranking of the docked poses according to various scoring functions indicate that GoldScore(Verdonk, Cole et al. 2003) performs best followed closely by ScoreJEversion4.

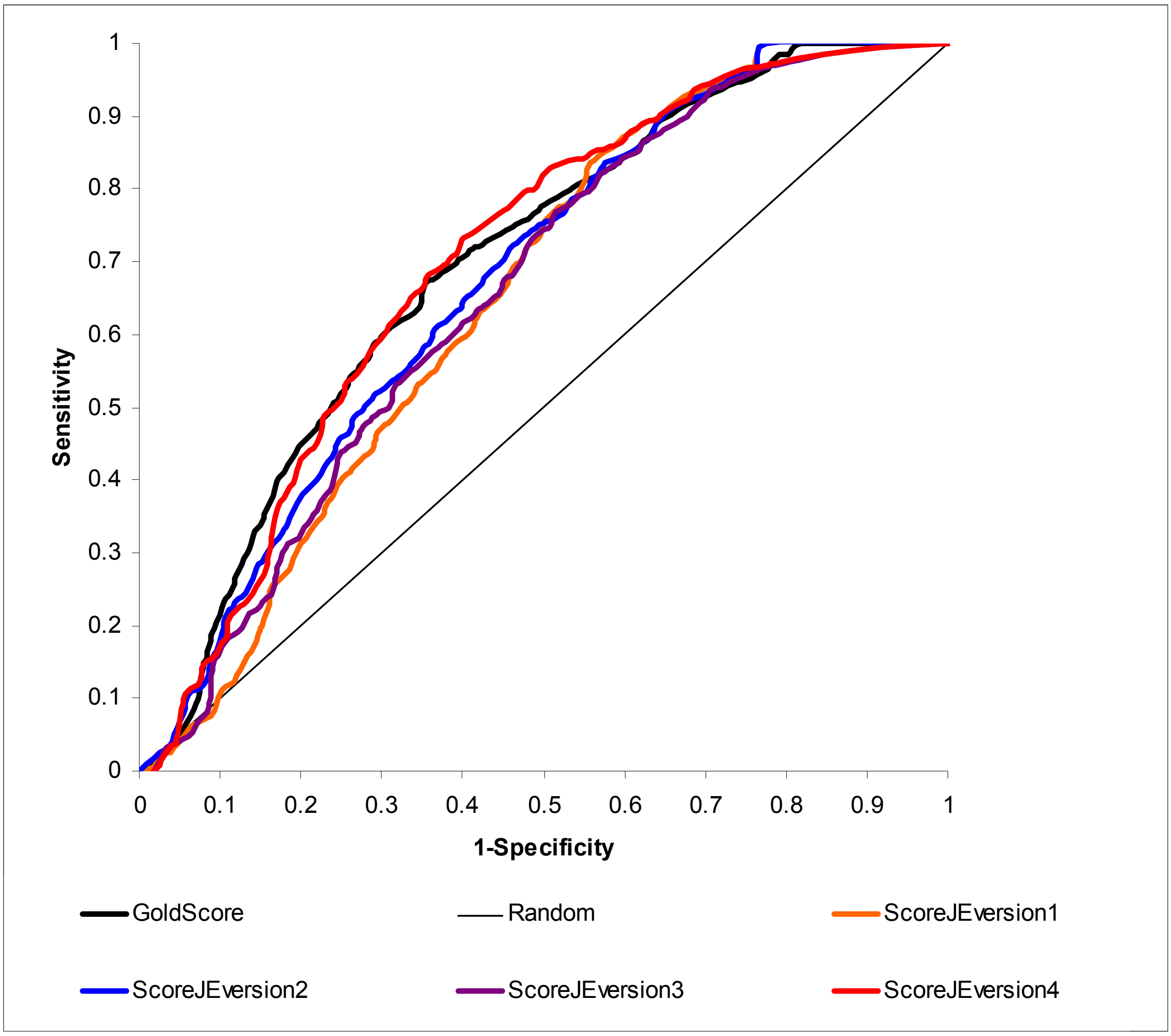


Figure 4.8: Receiver Operating Characteristic curves for comparative study of efficiency of GoldScore and various versions of ScoreJE in identifying near-native docking solutions for a dataset of 50 protein-ligand complexes. ScoreJE version1 when – only atomtype contacts are considered, cersion2 – atomtype contacts and the inter-atomic distance is considered (A-1 from Figure 4.3), version3 - atomtypes and dihedral angle (A12-123 from Figure 4.3) is considered, version4 – atom types, distances, angles and dihedrals are included in calculation of ScoreJE.

4.5 Discussion

Mutual information is a widely used statistic in several fields. Having first appeared in Shannon's paper(Shannon 1948) it has gained widespread acceptance in the applications of information theory. Where as joint entropy (Reza 1994) measures the amount of uncertainty or entropy associated with two random variables, mutual information measures the information. The protein-ligand interaction information obtained from the crystallographically determined structures was converted into protein-ligand atomic contact preferences. These were combined with predicted solvent interactions to create solvation potentials in SIScoreJE.

In our study joint entropy and mutual information coefficients were considered as scores representing the atomic contact preference scores. The ScoreJE for an interaction depends on the joint probability of occurrence of the protein-ligand interacting atom pair and ScoreMI depends on the marginal probabilities of the individual atom types. The success of joint entropy over mutual information is evident (see Figure 4.2) for almost all atom-type definitions. This could be attributed to the nature of these two quantities. Mutual information is the amount of information one can calculate for the occurrence of an event on the basis of knowledge about the occurrence of another related event. In protein-ligand interaction terms: mutual information reduces the choice of a ligand atom identity that can form an interaction with a given protein atom in a particular ligand binding pocket. While this is useful it does not provide the measure of the amount of information the system will gain once the interaction takes place. Joint entropy, on the other hand is a measure of uncertainty associated with two random variables. As information decreases in uncertainty, the joint entropy provides a more accurate estimate about the information the system gains once the ligand atom (with highest joint probability of occurrence) forms an interaction with the protein atom. Only when the joint probability of occurrence of the two entities is absolute does the mutual information becomes equal to the joint entropy. Perhaps localized regions on protein surface with high cumulative joint probability of occurrence have greater ligand binding potential. Such regions have been termed as "hotspots" on the protein surface(Gohlke, Hendlich et al. 2000).

In order to quantify the effects of orientation of interacting atoms on the efficiency of ScoreJE additional parameters (see methods) were included in the calculation of the scoring function. Two interesting trends can be seen in the Figure 4.3. Firstly the basal scoring function (based on atom-type identity alone) performed best. This could be an artifact due to the nature of the test dataset (consisting of crystal structures). Addition of orientation indices to the basal scoring function (calculated on the basis of atom-type identities) provides an enhanced ability to distinguish the stereo-chemically unfavorable interactions. However, as the probability of occurrence of stereo-chemically unfavorable contacts in high-resolution crystal structures is very low, the increased ability of the scoring function to identify unfavorable contacts remains unutilized. To test this hypothesis ligands were docked into their cognate receptor site using the GOLD program and 100 poses were generated. Since the docking of a ligand to its receptor site creates a number of stereo-chemically unfavorable atomic interactions, the dataset of docked poses was used to study whether the inclusion of the different orientation indices has any effect on near-native pose identification. The ScoreJE scoring function version5 (calculated by including the inter-atomic distances, planar angles and dihedrals) performed almost as well as GoldScore. The basal scoring function did not perform as well and demonstrates the differential ability of the expanded scoring functions to distinguish between the stereo-chemically unfavorable and favorable contacts in a docking context. The performance of the other versions of scoring function was comparable and not significantly worse than the basal scoring function. The number of interacting atom-type combinations increased dramatically (

Table 4.2) on inclusion of additional orientational information. As the amount of information in the interaction database was constant the average amount of the information available per interacting atom-type combination reduced. However, the effect of this reduced information, did not affect the ability of the scoring function to predict the binding energies (see Figure 4.3).

The development of the SSP was based on the same principle of joint entropy. As the number of water molecules in the crystal structures were inadequate to generate the solvation potentials the interaction between the protein atoms and the modelled water molecules were used. In order to make the single body solvation potentials free from bias 1000 proteins were used for modelling the water molecules using Aquarius2 (see methods). A large number of interactions between the ligand and protein atoms during complex formation occur as a consequence of the hydrophobic effect(Williams and Bardsley 1999). Inclusion of SSPs in ScoreJE improved the degree of correlation between the predicted scores and the experimental binding energies (Figure 4.4 and 4.5). However this improvement was mostly as a result of the carbohydrate binding proteins (Figure 4.5). This leaves room for the development of better alternative solvation models with ScoreJE, creating a more efficient scoring function.

Evaluation of the accuracy of ScoreJE, SIScoreJE, GOLD:GoldScore, GOLD:ChemScore and X-Score to predict protein-ligand binding affinity was carried out and SIScoreJE and ScoreJE were seen to perform better than the rest. XScore is an empirical scoring function and has been seen to perform better than most scoring functions currently in use (Verdonk, Cole et al.). Similarly, ChemScore is an empirical scoring function that is widely applied (available in Sybyl and GOLD) in docking. That ScoreJE and SIScoreJE are able to perform better than GOLD:GoldScore, GOLD:ChemScore and XScore.

4.6 Conclusions

This chapter describes the development of a novel, knowledge based scoring function designed to estimate the protein-ligand interaction energy. The ScoreJE was tested on a set of 100 protein-ligand complexes. The ability of the scoring function in ranking the protein-ligand docking solutions has been investigated. The ScoreJE scoring function which included the information of orientation along with the identities of the interacting atoms performs at the same level as GOLD:GoldScore.

4.7 References

"http://www-mitchell.ch.cam.ac.uk/dataset205.html."

Bader, G. D., I. Donaldson, et al. (2001). "BIND--The Biomolecular Interaction Network Database." Nucleic Acids Res 29(1): 242-5.

Berman, H. M., J. Westbrook, et al. (2000). "The Protein Data Bank." Nucleic Acids Res 28(1): 235-42.

Cline, M. S., K. Karplus, et al. (2002). "Information-theoretic dissection of pairwise contact potentials." Proteins 49(1): 7-14.

Finkelstein, A. V., A. M. Gutin, et al. (1995). "Perfect temperature for protein structure prediction and folding." Proteins 23(2): 151-62.

Gohlke, H., M. Hendlich, et al. (2000). "Knowledge-based scoring function to predict protein-ligand interactions." J Mol Biol 295(2): 337-56.

Gohlke, H., M. Hendlich, et al. (2000). "Predicting binding modes, binding affinities and 'hot spots' for protein-ligand complexes using a knowledge-based scoring function." Perspectives in Drug Discovery and Design 20(1): 115-144.

Gohlke, H. and G. Klebe (2001). "Statistical potentials and scoring functions applied to protein-ligand binding." Curr Opin Struct Biol 11(2): 231-5.

Hendlich, M., P. Lackner, et al. (1990). "Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force." J Mol Biol 216(1): 167-80.

Klebe, G. (2006). "Virtual ligand screening: strategies, perspectives and limitations." Drug Discov Today 11(13-14): 580-94.

Ladbury, J. E. Scorpio database http://www.biochem.ucl.ac.uk/scorpio/scorpio.html.

Laskowski, R. A. (2001). "PDBsum: summaries and analyses of PDB structures." Nucleic Acids Res 29(1): 221-2.

Laskowski, R. A., V. V. Chistyakov, et al. (2005). "PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids." Nucleic Acids Res 33(Database issue): D266-8.

Laskowski, R. A., E. G. Hutchinson, et al. (1997). "PDBsum: a Web-based database of summaries and analyses of all PDB structures." Trends Biochem Sci 22(12): 488-90.

McDonald, I. K. and J. M. Thornton (1994). "Satisfying hydrogen bonding potential in proteins." J Mol Biol 238(5): 777-93.

Muegge, I. and Y. C. Martin (1999). "A general and fast scoring function for protein-ligand interactions: a simplified potential approach." J Med Chem 42(5): 791-804.

Muegge, I., Y. C. Martin, et al. (1999). "Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein." J Med Chem 42(14): 2498-503.

Murzin, A. G., S. E. Brenner, et al. (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." J Mol Biol 247(4): 536-40.

Pedretti, A., L. Villa, et al. (2002). "VEGA: a versatile program to convert, handle and visualize molecular structure on Windows-based PCs." J Mol Graph Model 21(1): 47-9.

Pedretti, A., L. Villa, et al. (2004). "VEGA--an open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming." J Comput Aided Mol Des 18(3): 167-73.

Pickett, S. D. and M. J. Sternberg (1993). "Empirical scale of side-chain conformational entropy in protein folding." J Mol Biol 231(3): 825-39.

Pitt, W. R. and J. M. Goodfellow (1991). "Modelling of solvent positions around polar groups in proteins." Protein Eng 4(5): 531-7.

Rauh, D., G. Klebe, et al. (2004). "Understanding protein-ligand interactions: the price of protein flexibility." J Mol Biol 335(5): 1325-41.

Reza, F. (1994). "An Introduction to Information Theory." Book: 104-106.

Shannon, C. E. (1948). "A Mathematical Theory of Communication." The Bell System Technical Journal 27(4): 623-656.

Sharp, K. A., A. Nicholls, et al. (1991). "Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects." Science 252(5002): 106-9.

Sippl, M. J. (1990). "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 213(4): 859-83.

Sternberg, M. J., P. A. Bates, et al. (1999). "Progress in protein structure prediction: assessment of CASP3." Curr Opin Struct Biol 9(3): 368-73.

Tanaka, S. and H. A. Scheraga (1976). "Statistical mechanical treatment of protein conformation. I. Conformational properties of amino acids in proteins." Macromolecules 9(1): 142-59.

Taylor, J. S. and R. M. Burnett (2000). "DARWIN: a program for docking flexible molecules." Proteins 41(2): 173-91.

Verdonk, M. L., J. C. Cole, et al. (2003). "Improved protein-ligand docking using GOLD." Proteins 52(4): 609-23.

Verkhivker, G., K. Appelt, et al. (1995). "Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity." Protein Eng 8(7): 677-91.

Wallqvist, A., R. L. Jernigan, et al. (1995). "A preference-based free-energy parameterization of enzyme-inhibitor binding. Applications to HIV-1-protease inhibitor design." Protein Sci 4(9): 1881-903.

Wang, R., Y. Lu, et al. (2004). "An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes." J Chem Inf Comput Sci 44(6): 2114-25.

Williams, D. H. and B. Bardsley (1999). "Estimating binding constants - The hydrophobic effect and cooperativity." Perspectives in Drug Discovery and Design 17(1): 43-59.

Zentgraf, M., H. Steuber, et al. (2007). "How reliable are current docking approaches for structure-based drug design? Lessons from aldose reductase." Angew Chem Int Ed Engl 46(19): 3575-8.

4.8 Appendices

APPENDIX-1a

```
General Description:
 ~~~~~~~~~~~~~
 General atom type - Bond order - Ring indicator - Aromatic indicator

 General atom types:          Bond order:
 ~~~~~~~~~~~~~~~~~~~~              ~~~~~~~~~~~~
 X = Any atom                0   = Atom not bonded
 # = Heavy atom              1-6 = Bond order
 - = None                    9   = Any bond order

 Ring Indicator:            Aromatic Indicator:
 ~~~~~~~~~~~~~~~~               ~~~~~~~~~~~~~~~~~~~~
```

```
0     = Don't check ring            0 = Don't check
3...6 = From 3 to 6 member ring       1 = Aromatic
9     = Generic ring

******************************
****   ScoreJE Atom Type set-1   ****
******************************
 Type Atm    Bonded Atoms
========================================================================
=====


  C0    C-400  (#-900 #-900 #-900 #-900)
  C0    C-400  (H-100 H-100 H-100)
  C0    C-400  (H-100 H-100)
  C0    C-400  (H-100)
  C+1   C-300  (N-300 N-300 N-300)
  C-H   C-300  (H-100 O-100 O-100)
  C-2   C-300  (O-100 O-100 O-100)
  C-1   C-300  (#-900 O-100 O-100)
  C     C-300  (O-100 #-900 #-900)
  C3    C-300  (#-900 #-900 #-900)
  C3    C-300  (H-100 H-100)
  C3    C-300  (H-100 O-100)
  C3    C-300  (H-100)
  C4    C-200  (#-900 #-900)
  C4    C-200  (H-100)
  C5    C-100  (C-200)

 Donors


  hn   H-100  (N-400)
  hn   H-100  (N-300)
  ho   H-100  (O-200)
  hf   H-100  (F-100)
 Acceptors


  n    N-300 (C-400 C-400 C-400)
  n    N-300 (C-400 C-400 H-100)
  n    N-300 (C-400 H-100 H-100)
  n    N-300 (H-100 H-100 H-100)
  np   N-200 (#-951 #-951)
  n    N-200
  o    O-200
  o    O-100
  f    F-100


  x    X-900
```

APPENDIX-1b

```
*******************************
****   ScoreJE Atom Type set-2   ****
*******************************


 Type Atm    Bonded Atoms


=============================================================
======


  C0    C-400  (#-900 #-900 #-900 #-900)
  C0    C-400  (H-100 H-100 H-100)
  C0    C-400  (H-100 H-100)
  C0    C-400  (H-100)
  C+1   C-300  (N-300 N-300 N-300)
  C-H   C-300  (H-100 O-100 O-100)
  C-2   C-300  (O-100 O-100 O-100)
  C-1   C-300  (#-900 O-100 O-100)
  C     C-300  (O-100 #-900 #-900)
```

C3	C-300 (#-900 #-900 #-900)
C3	C-300 (H-100 H-100)
C3	C-300 (H-100 O-100)
C3	C-300 (H-100)
C4	C-200 (#-900 #-900)
C4	C-200 (H-100)
C5	C-100 (C-200)

O1	O-200 (H-100 C-300 (O-100 O-100))
O1	O-200 (C-900 C-900)
O1	O-200 (C-900 C-991)
O1	O-200 (C-991 C-991)
O1	O-200 (C-300 (O-100))
O1	O-200 (H-100 C-391)
O1	O-291
O2	O-100 (C-300 (O-100 O-100))
O2	O-100 (C-200 (N-200))
O2	O-100 (S-400 (N-900))
O2	O-100 (N-900)
O2	O-100 (S-900)
O2	O-100 (C-991)
O2	O-100 (C-900)
O2	O-100

N1	N-400 (#-900 #-900 #-900 #-900)
N1	N-400 (H-100 H-100 H-100)
N1	N-400 (H-100 H-100)
N1	N-400 (H-100)
N2	N-391 (H-100)
N2	N-300 (H-100 H-100 C-300 (O-100))
N2	N-300 (H-100 H-100 S-300 (O-100))
N2	N-300 (H-100 H-100 C-300 (S-100))
N2	N-300 (H-100 H-100 C-300 (N-200))
N2	N-300 (H-100 H-100 C-300)

N2	N-300 (H-100 H-100 N-200)
N2	N-300 (H-100 H-100 C-391)
N1-	N-300 (O-100 O-100 O-100)
N2	N-300 (H-100 C-300 (O-100))
N2	N-300 (H-100 S-300 (O-100))
N2	N-300 (H-100 C-300 (S-100))
N2	N-300 (H-100 C-300 (N-200))
N2	N-300 (#-900 #-900 #-900)
N2	N-300 (H-100 C-391)
N2	N-300 (H-100 H-100)
N2	N-300 (H-100)
N3	N-251
N3	N-291
N3	N-200 (C-300 C-300)
N3	N-200 (#-900 #-900)
N3	N-200 (H-100)
N3	N-200 (H-100)
N3	N-100 (C-300)
N3	N-100 (C-200)

Donors

hn	H-100 (N-400)
hn	H-100 (N-300)
ho	H-100 (O-200)
hf	H-100 (F-100)

Acceptors

n	N-300 (C-400 C-400 C-400)
n	N-300 (C-400 C-400 H-100)
n	N-300 (C-400 H-100 H-100)
n	N-300 (H-100 H-100 H-100)
np	N-200 (#-951 #-951)

n	N-200
o	O-200
o	O-100
f	F-100
x	X-900

APPENDIX-II

Binding Energy Evaluation Test Set

1a4k 1a9m 1aaq 1abe 1add 1ae8 1ajx 1am6 1apb 1apu 1b5g 1b6j
1b6k 1b6l 1b6m 1bap 1bb0 1bdr 1bmn 1bra 1bv7 1c5c 1c83 1c84
1c86 1c87 1c88 1c8k 1cbs 1cbx 1cf8 1com 1cps 1ctr 1d3p 1d4l
1d4p 1dbb 1dbj 1dbk 1dhf 1dmp 1dog 1drf 1dwb 1eap 1fax 1fkg
1g2k 1gno 1gpy 1hew 1hih 1hos 1hps 1hte 1htg 1hvh 1hvl 1lgr
1mbi 1mcf 1mnc 1mtw 1nnb 1nsd 1okl 1okm 1phf 1qbt 1rgk 1rgl
1tng 1tnh 1tnj 1tnl 1tph 1ulb 1uvs 2cpp 2ctc 2dbl 2er9 2gbp
2ifb 2upj 2web 3cla 3ptb 4phv 4tln 5abp 5cpp 5gpb 5hvp 6abp
7abp 7dfr 8abp 9icd

Chapter 5: Conclusion

Rational drug design has been the ultimate aim of the structural bioinformatics. The targeting of protein-protein interaction on the basis of knowledge about the receptor structure is difficult. However the successful targeting of GGTase-II activity by MK_INH_X21986 once again underlines the possibilities in this field. Here, it must be noted that later the compound was also found to inhibit the homologs of GGTase-II. Careful docking analysis predicted FTase inhibition due to the competition of peptide substrate with the MK_INH_X21986. This was indeed found to be the case in wet lab experiments. The cross reactivity question was not addressed when the molecule was being designed as it was assumed that the shape of REP interacting hydrophobic groove on GGTase-II surface is unlikely to find anything similar in the homologous structures. This was a mistake. The strategy for targeting protein-protein interaction interface must be such that only unique druggable sites are chosen in the first place. Alternatively targeting a site which may have close resemblances on the surface of homologous proteins can be done through substractive docking. This strategy was indeed used in the later process and a virtual library of molecules was created that targeted only the GGTase-II hydrophobic groove.The validity of the model is yet to be tested.

Correct identification of unique druggable sites is a major challenge. If such sites could be identified with reasonable confidence than the task of rational drug design shall become easier. Identification of the putative sites that can bind ligand molecules is routine. However none of the existing programs are able to assign the character of the ligand molecules that shall target a specific site. Such a tool would be of immense value for computational chemists as they do not have to waste time targeting sites that bind non-drug like compounds (example carbohydrates). This was the aim of the second project. InCa-SiteFinder that is created performs exceptionally well in identification of the druggable sites and distinguishing these from the carbohydrate binding sites. Some of the sites are classified as having dual propensities and these are the sites a computational chemist should be aware of. Targeting such sites may not be as successful as targeting purely drug-like binding sites. On the other hand if such a site houses the enzymatic centre of a protein interaction site then targeting such site also indicates the possible mechanism of the small molecule. If the inhibition of the enzymatic activity or protein-protein interaction is brought about by targeting some other site then the mechanism can be allosteric inhibition.

The value of the differential propensity (DPS) score to distinguish the preference of the predicted site for carbohydrate over drug-like compound ligands is a key factor in the success of InCa-SiteFinder. Sites with high positive values are almost always carbohydrate binder. On the other hand the sites with greater negative values are mostly drug-like compound binding sites. This is valuable information as the carbohydrate binding sites that do not bind drugs-like molecules are easily identified. More remarkable still is the identification of drug-like binding sites with greater success than the identification of carbohydrate binding sites. Though our aim was the prediction of carbohydrate binding sites we have noted the potential use of InCa-SiteFinder in identification of drug-like binding sites, although we have not attempted to optimise the method for this purpose. The sites with dual propensity to bind carbohydrate and drug-like compounds have the DPS values between 10 and -20. It is interesting to speculate that the tool developed here may form the basis for a method that could not only discriminate between different types of functional site, but also facilitate the process of structure-based drug design. In the later case an ability to characterise sites that are amenable to binding drug-like molecules are of great interest for medicinal applications, including blocking protein-protein interactions and for design of competitive inhibitors.

Rational drug design cannot be successful if the affinity of virtual molecules and the structure of the protein receptor can not be estimated in fast and accurate manner. The current day scoring function are not very efficient. The correct way is to solve the wave equations for multi-electron systems and the computational power is simply not enough. In such scenario knowledge based scoring functions should be a better option. However most of the development of such scoring functions has been done in not very scientific way. For instance all of the existing scoring functions are based on the training databases which were not made unbiased. The inherent redundancy of the information in the training set biases the scoring function for particular class of proteins. Towards this end our scoring function ScoreJE and SIScoreJE are considerable improvement. Not only they out performs some of the most widely used scoring functions but are fast enough to be integrated in the docking algorithms.

Future perspectives

The story has just begun……

The completion of a tool for identification and prediction of druggable sites and a scoring function for the estimation of the binding affinity *in silico* underlines the need to have a third program which can analyse the predicted sites for drug-like compound binding and generate a varying Gaussian propensity based pharmacophore and screen the commercially available compounds. A pseudocode was completed but could not be implemented on account of the scarcity of time. Continuation and completion of the project demands its implementation of this third tool because without it the sitefinder and binding affinity scoring function are like wood and gasoline waiting for the matchbox to take shape.

The targeting of the GGTase-II is yet an open question. Testing of second library should answer some of the questions about the cross reactivity.