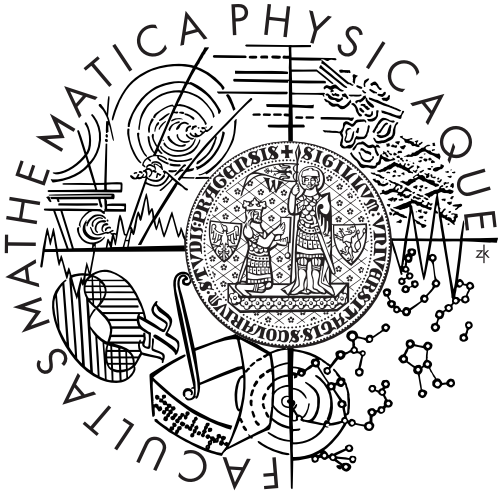
Charles University in Prague

Faculty of Mathematics and Physics

MASTER THESIS



Radoslav Krivák

Algorithms for Protein-Ligand Binding Site Discovery

Department of Theoretical Computer Science
and Mathematical Logic

Supervisor: Mgr. Roman Neruda, CSc.

Co-supervisor: RNDr. David Hoksza, Ph.D.

Study programme: Computer science

Specialization: Theoretical computer science

Prague 2013

First of all I would like to thank my supervisor Roman Neruda for his guidance and helpful suggestions. Equally thankful I am to David Hoksza for co-supervising the thesis and introducing me to the exciting field of structural bioinformatics.

Dedicated to my family.

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

Prague, 1st August 2013 Radoslav Krivák

Title: Algorithms for protein-ligand binding site discovery
Author: Radoslav Krivák
Department: Department of Theoretical Computer Science
and Mathematical Logic
Supervisor: Mgr. Roman Neruda, CSc.
Co-supervisor: RNDr. David Hoksza, Ph.D.

Abstract: Virtually all processes in living organisms are conducted by proteins. Proteins perform their function by binding to other proteins (protein-protein interactions) or small molecules – so called ligands (protein-ligand interactions). Active sites for protein-ligand interactions are pockets in protein structure where ligand can bind. Predicting of ligand binding sites is the first step to study and predict protein functions and structure based drug-design. In this thesis we reviewed current approaches for binding site prediction and proposed our own improvement. We have developed a novel pocket ranking function based on prediction model that predicts ligandability (ability to bind a ligand) of a given point inside of a pocket. Prediction is done considering only a local physicochemical and geometric properties derived from neighbourhood.

Keywords: Structural Bioinformatics, Protein-ligand binding sites, Machine learning

Název práce: Algorithms for protein-ligand binding site discovery
Autor: Radoslav Krivák
Katedra: Department of Theoretical Computer Science
and Mathematical Logic
Vedoucí diplomové práce: Mgr. Roman Neruda, CSc.
Konzultant: RNDr. David Hoksza, Ph.D.

Abstrakt: Prakticky každý proces v živých organismech je zajišťován promocí proteinů. Proteiny vykonávají svoji funkci buď vázáním se na další proteiny (protein-protein interakce), nebo na malé molekuly, tzv. ligandy (protein-ligand interakce). Aktivní místa pro protein-ligand interakce jsou tzv. kapsy v proteinové struktuře, kam se může ligand navázat. Predikce těchto kapes je obyčejně prvním krokem ke studiu funkce daného proteinu a taky základem k strukturálně orientovanému vývoji léků. V této práci jsme udělali přehled existujících metod a představili naše vlastní vylepšení. Vyvynuli jsme novou funkci pro ranking kapes, která je založená na predikci ligandability (schopnosti vázat ligand) budu vně kapsy jen na základě chemických a geometrických vlastností lokálního okolí daného bodu.

Klíčová slova: Strukturální Bioinformatika, Vazební místa pro Ligandy, Strojové učení

Contents

Preface

Back in the year 1943 when physical basis of life was not clear and molecular mechanism of reproduction was yet to be discovered Erwin Schrödinger wrote a short book titled *What is Life?* [Schrödinger, 1944] that inspired a number of pioneers of molecular biology. The underlying idea of the book was that the essence of life is information and the thesis was that this information is stored by conformation of atoms in molecules.

It is not, however, just the genetic information that is of interest for biology. Since 1950's we have collected large amount of three-dimensional structural data that describe individual parts of macromolecular machinery of living organisms. Our understanding of how those parts work together and ability to extract applicable knowledge from this data is but lacking behind the collection efforts. Although we still have only a fraction of structural information about macromolecules in human body, databases are growing exponentially.

This resembles the situation with genetic sequence data at the end of the century. Now, ten years after the completion of the Human Genome Project, it is clear that genomics in medicine did not bring improvements and new treatments many hoped for. Relationships between genes and biological function turned out to be too complex. Thanks to genomics we can, for instance, predict that individuals with certain genetic variation are 25% more likely to get Alzheimer's disease, but not what is the underlying mechanism and what can we do to stop it. This is where structural biology comes into the picture.

We are approaching the age when structural information of all molecules in human body will be known (which is anticipated within two decades [Nair et al., 2009]). At the same time DNA sequencing is getting radically cheaper and soon it will be possible to have a complete genetic information of each person as a presupposition of any medical intervention. By combining this information we can get a full structural information of an individual and offer personalized and fundamentally rationalized (not just statistically effective) treatments.

Intelligent algorithms are therefore needed that can interpret and exploit rapidly growing amount of biological structural data.

Chapter 1

Introduction

The goal of this chapter is to provide a biological, bioinformatical and pharmacological overview of the problem and its context. We will also try to touch on all realities and facts which lead to considerations that were shaping presented work. In case of biochemistry we will go to greater detail only when it will be directly used in our proposed algorithm (such as list of amino acids and their properties) or particularly interesting from a computer science point of view.

1.1 Biochemistry and molecular biology

1.1.1 Proteins

Proteins and their role

Proteins are biological macromolecules responsible for most processes in living organisms. Of all types of biological macromolecules (such as polysacharides and nucleic acids) they are arguably most complex in terms of variety of their forms and functions they perform. To give a better idea about the ubiquity and variety of proteins lets briefly look at some of the most important protein types and their roles.

Fibrillar proteins are building blocks of muscle fibres and generate their coordinated mechanical motion. However, they also have important passive roles and provide bones and ligaments with their characteristic strength. *Transport proteins* store and transport various biologically important substances such as glucose, O_2, metal ions and lipids. *Enzymes* are highly selective biological catalysts, which means they accelerate both rate and specificity of metabolic reactions. An enzyme, through the arrangement of atoms in its enzymatic active site, creates energetically favourable pathways for chemical reaction that would not occur in physiological

conditions at a sufficient rate otherwise. These reactions can have either catabolic (breakdown of complex molecules into simpler ones) or anabolic (resulting in more complex molecules) character.

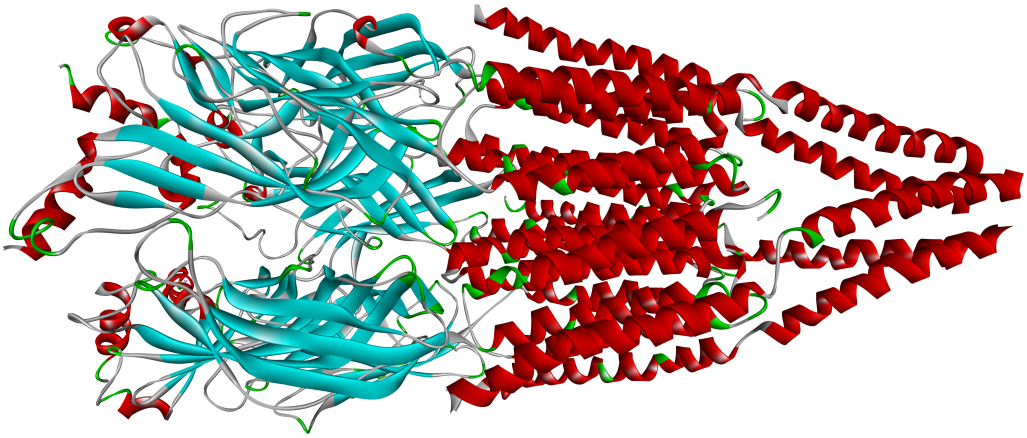


Figure 1.1: Structure of nicotinic acetylcholine receptor in a closed state: a prototypic example of a ligand-gated ion channel (PDB code: 2BG9)

Receptors are proteins that help to regulate those metabolic reactions by mediating chemical signals. Those signals are carried by small signaling molecules such as hormones, neurotransmitters, drugs and other chemical messengers. They do so by recognizing certain molecules and letting them attach. A small molecule that attaches to a receptor (or any protein) is called a *ligand*. This temporary ligand binding results in conformational change of the receptor protein, that — by interacting with other biomolecules — sends the signal further down the signaling pathway or produces the desired biochemical response itself. The prominent type of receptors are so-called GPCRs (G-protein coupled receptors). They are transmembrane proteins that sense signaling molecules outside of the cell and activate signal pathways inside. Receptors within GPCR class share a similar design and mechanism of action but respond to different signaling molecules because of relatively small differences in their ligand binding sites. Another interesting type of receptors are ligand-gated ion channels (Fig. 1.1). They are transmembrane proteins that allow certain ions (such as Na^+, K^+, Ca^{2+}, or Cl^-) to get in or out of the cell. As their name suggests, they open only in the presence of particular ligand(s). Ligand-gated ion channels play central role in nerve cells where they convert chemical signal of released neurotransmitter into an electrical signal.

Amino acids and polypeptides

Structurally a single protein consists of one or more long chains of amino acids linked together by covalent peptide bonds. Those chains are called polypeptides

and are always linear (that is, no branching occurs). Amino acids that are already linked in a polypeptide chain are usually referred to as *residues* (since, at least conceptually, two stand-alone amino acids polymerise through the elimination of one water molecule). Polypeptides in proteins range in length from ~40 to ~34,000 amino acid residues (although few are longer than 1500 residues) [Voet and Voet, 2010]. It should be noted here that not all amino acids found in living organisms are constituents of polypeptides and not all polypeptides are parts of proteins. Some (usually shorter) polypeptides exist independently and play various biologically important roles.

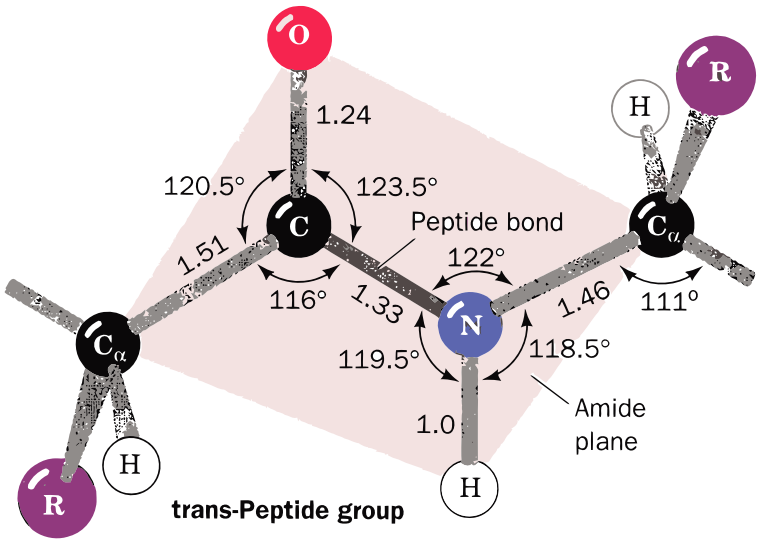


Figure 1.2: Peptide bond [Voet and Voet, 2010]

Polypeptides that proteins are made of are biosynthesized from their monomeric units in ribosomes according to the information stored in DNA. There are 20 standard amino acids. These are the amino acids encoded by triplets of universal genetic code, which is nearly identical for all known life-forms. All proteins are initially synthesized from those 20 standard amino acids.[1] All of these amino acids consist of the same peptide group built around the central α-Carbon atom but differ in their side chains. These side chains are of different sizes, shapes and physicochemical properties. Great variety of proteins largely stem from the varied properties of side chains of 20 standard amino acids.

Standard amino acids are, however, not the only amino acids found in actual proteins. Some residues on certain proteins are modified in the process called post-translational modification. In most cases this modification is essential for the function of the protein.

[1] To be correct there are 2 more amino acids (Selenocysteine and Pyrrolysine) that are incorporated into polypeptide chains by unique synthetic mechanisms. Sometimes they are being included among standard amino acids. Their occurrence is but rare (only 25 human proteins for example contain Selenocysteine [Kryukov et al., 2003]).

Chemical bonds and interactions

In this section we will describe some of the most important chemical interactions that play roles in protein structures themselves (folding and stabilizing, see p. 14) and in their bonding with other molecules. In some cases typical bonding energies are provided for comparative reasons.

- **Covalent bonds.** Covalent bonds are bonds between atoms that share one or more (up to three) electron pairs. They are comparatively "firm" that is, not so easily reversible. Typical length of covalent bonds in organic molecules range from 1 to 1.5 Å. The length and strength of a particular covalent bond does not depend only on the type of the atoms that enter the bond (and the number of electron pairs they share) but is also affected by the conformation of other neighboring atoms in the molecule. Double and triple bonds are stronger than single bond but restrict rotation around bond axis. Special type of covalent bonds can be observed in so-called aromatic rings of benzene and other organic molecules where electrons are delocalized between more than 2 atoms.

- **Hydrogen bonds** or H bonds are predominantly electrostatic interactions between a weakly acidic hydrogen donor group and a weakly basic acceptor that is an atom with lone pair of electrons or a π bond (schematic representation: D—H···A) [Horowitz and Trievel, 2012]. They are characterized by an H···A distance that is at least 0.5 Å shorter than the calculated van der Waals distance (distance of closest approach between two non-bonded atoms) between the atoms. The energy of a hydrogen bond is small compared to covalent bond energies (for instance, $\sim$460 kJ $\cdot$ mol^{-1}for an O—H covalent bond). Hydrogen bonds have energies that are normally in the range of 12 to 40 kJ $\cdot$ mol^{-1}. The typical length of a hydrogen bond (the distance D···A) is in the range 2.7 to 3.1 Å. Some of the hydrogen bonds in proteins are members of networks in which each donor is H bonded to two acceptors (a bifurcated hydrogen bond) and each acceptor is H bonded to two donors.

- **Ionic interactions.** An ion is a particle or molecular group charged either positively (cation) or negatively (anion). The charge is caused by surplus or deficit of electron(s). The association of two ionic protein groups of opposite charge is known as an ion pair or salt bridge. The distance at which two such groups form a salt bridge is usually taken to be 4 Å.[2] Ionic bonds have relatively high stability (typical energy $\sim$80 kJ $\cdot$ mol^{-1}).

- **Polar interactions.** When bonding electrons are asymmetrically distributed over the atomic nuclei involved, one atom will bear a negative, and its partner a positive *partial* charge. The molecule thus presents a positive and a negative

[2] [Karshikoff and Jelesarov, 2008], all other numbers in this section taken from [Voet and Voet, 2010]

pole, i. e., is a dipole. A partial charge can interact electrostatically with an ion or another dipole. Dipoles may be permanent or induced – a permanent dipole can induce a dipole moment on a neighboring group so as to form an attractive interaction.

- **van der Waals interactions** are non-covalent associations between electrically neutral molecules. Although nonpolar molecules are nearly electrically neutral, at any instant they have a small dipole moment resulting from the rapid fluctuating motions of their electrons. This transient dipole may induce dipole in the neighbouring molecule such that they are attracted to one another (a quantum mechanical effect that cannot be explained in terms of only classical physics). These so-called *London forces* are extremely weak and only significant for contacting groups because their association energy is proportional to r^{-6} [Voet and Voet, 2010]. Nevertheless, the great numbers of interatomic contacts in the closely packed interiors of proteins or between sterically complementary ligands and binding sites can make them very significant.

- **Hydrophobic forces.** Nonpolar substances are hydrophobic, i. e., they have tendency to minimize their contacts with water. To put it very roughly: in a sense by trending towards each other, polar H_2O molecules squeeze apolar particles from their midst [Lüllmann, 2005]. The actual physical mechanism of this effect is more complex and the tendency of nonpolar molecules to cluster together in aqueous environment is entropic in character. In a similar manner we can say that polar molecules or groups are hydrophilic and have a tendency to be exposed to water. Globular proteins have cores comprised mostly of hydrophobic residues and polar residues located on the outside, in contact with the aqueous solvent.

- **Other interactions.** Very recently determined ultra-high-resolution protein structures (1.1 Å resolution allowing to determine hydrogen atoms positions) confirmed prevalence of many poorly understood or at least underappreciated interactions (unusual H bonds, $n \cdots \pi*$ interactions) [Chen et al., 2012, Horowitz and Trievel, 2012].

In case of small molecules we can attribute characterizing labels to the whole molecule. For example H_2O is a polar molecule and benzene (C_6H_6 ring) an aromatic molecule. Bigger molecules can consist of parts with different characteristics. Single molecule can have parts that are polar and others that are non-polar (aliphatic or aromatic). In the same fashion one molecule can be acid and base or anion and kation (zwitterion, German: hybrid) at the same time.

Protein structure

Two amino acid residues connected by a peptide bond are spatially flexible around two rotational axes of C_α—N and C_α—C bonds. This allows polypeptide chains to fold and create complex three-dimensional structures. Structure of proteins can be described in terms of four levels:

- **Primary structure** defines the order of amino acid residues in polypeptide chain(s) of the protein. The sequence of amino acids in a particular protein is given by the sequence of nucleotides in a protein coding part of a gene.
- **Secondary structure** is the local spatial arrangement of polypeptide's backbone without regard to the conformations of their side chains. Several local structural patterns can be recognized, of which α-helices and β-sheets are the most common. Subsequences of polypeptide chain can be annotated according to which secondary structural element they form.
- **Tertiary structure** refers to three-dimensional arrangement of all atoms of the protein including those in residue side chains and in any prosthetic groups (groups of atoms other than amino acids that were added to the protein to help them perform their function).
- **Quaternary structure.** Most of proteins are composed of more than one polypeptide chain. Quaternary structure refers to the spatial arrangement of subunits created by individual chains. These subunits are usually connected by noncovalent and in some cases disulfide covalent bonds.

Apart form the four structural levels we would like to introduce two other important terms related to protein structure: domain and fold. *Domains* are distinctive modules of single-chain proteins or subunits. They usually form a recognizable globular clusters. Many, but not all, proteins consist of several structural domains. Even though they are parts of a single polypeptide chain, they often fold, stabilize and evolve independently. Domains are basic building blocks of structural evolution. One domain may appear (often slightly modified) in different proteins of the same organism.

If we abstract from the exact 3D coordinates of the atoms in a single domain and consider only secondary structural elements of the backbone like α-helices and β-sheets and the way they are arranged with respect to each other we are talking about *fold*. The number of possible different folds may seem to be unlimited. However, comparisons of the now large number of known protein structures have revealed that few protein folds are unique. Theoretical considerations suggest there are less than 8000 naturally occurring folds of which around 1200 have already been observed [Voet and Voet, 2010].

Protein folding

Polypeptide chains of proteins in their natural physiological environment quickly fold into their native three-dimensional conformations. The way how exactly they fold and which of many possible conformations they finally assume is the subject of notorious protein folding problem. However, in recent years it became clear that many biologically important proteins remain natively unfolded [Tsvetkov et al., 2009]. Such intrinsically disordered proteins (IDPs) lack specific tertiary (and even secondary) structures. IDPs constitute a separate class of proteins that are not considered in this thesis. Interestingly, ~33% of eukaryotic proteins are predicted to contain long disordered regions [Ward et al., 2004].

Protein stability and flexibility

The stability of folded protein structures is mostly the result of a fine balance among the various non-covalent interactions and countervailing forces (ionic and dipolar interactions, hydrogen bonds, van der Waals interactions and hydrophobic forces). Covalent disulfide bonds can form between two Cysteine residues of one polypeptide chain. However, they are usually found only in extracellular proteins.

Most of protein structures in humans are only marginally stable under physiological conditions and are easily disrupted by sharp changes in temperature, pH or physical disruptions [Creighton, 1993]. Hyperthermophiles (organisms that grow at temperatures near 100°C) have many homologous proteins that carry almost the same functions. This fact suggests that this marginal stability is an essential property that has arisen through evolutionary design[3]. One explanation for this is that the marginal stability allows more structural flexibility which many proteins require to carry out their functions [Petsko and Ringe, 2004].

1.1.2 Ligands and binding sites

Ligands

Ligand in biochemistry and pharmacology[4] is a substance, typically a small molecule or a short peptide, that binds to a protein to form a complex. In relation to a particular protein, ligand can be native or artificial. Native ligand is "supposed to" bind to the protein as a part of some biological mechanism or regulatory process (case of neurotransmitters for example). Another way to look at it is to say that protein has evolved to recognize certain ligands. Artifical ligands (that can be

[3] More careful way to put it would be to say that clearly-possible higher structural stability does not seem to provide any evolutionary advantage.

[4] in the context of coordination chemistry the term ligand is used to mean any ion or molecular group that bind to a central metal atom

synthetic drugs or natural toxins) may, by binding to it, disable or improve protein function.

In most cases ligand binding is caused and stabilized by non-covalent forces. Binding is therefore usually reversible and its length somewhat proportional to the binding energy. In some cases binding can be permanent (for instance there are cases of enzyme inhibitors that permanently disable the enzyme by forming covalent bonds).

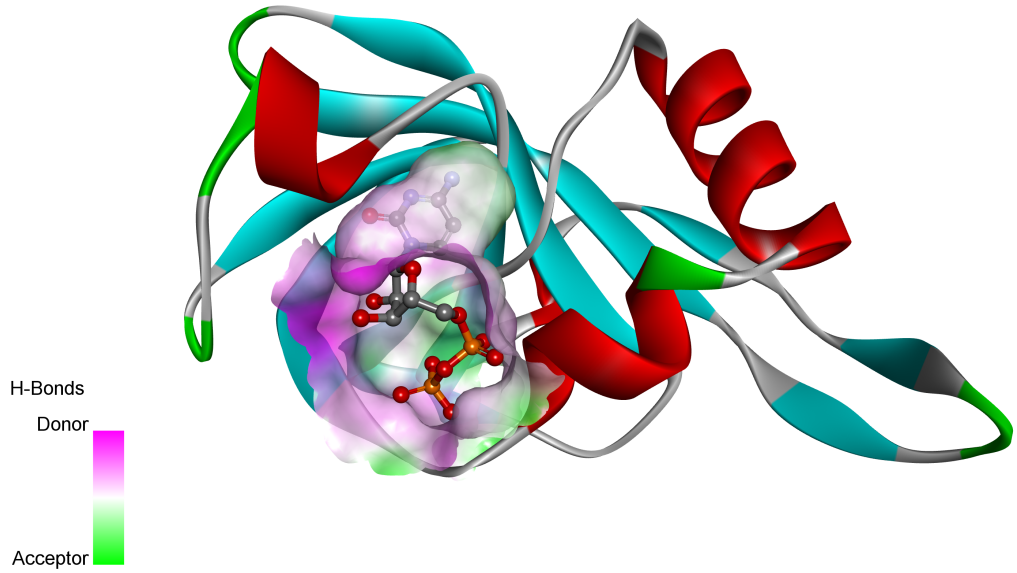


Figure 1.3: Example of ligand in a binding site

Binding sites

Ligand binding sites are usually found in concave pockets and clefts on the protein surface. On enzymes, location of active/binding site correlates with deepest or biggest cleft on protein surface [Laskowski et al., 1996]. Traditional view is that ligand-binding site tends to be the largest or most geometrically complex [Pettit and Bowie, 1999] cavity on the surface. Binding sites for small molecules are often generalized to be hydrophobic and enclosed cavities. However, not all ligand binding sites are deep pockets. It is important to acknowledge that such generalizations are derived from currently known structures of ligand binding complexes and have their exceptions. There are examples of ligands found to bind to rather exposed shallow clefts [Nisius et al., 2012]. Statistical analysis of 15,232 concavities in 756 protein structures showed that true binding sites cannot be determined on the basis of size alone. The chemical properties resulting from composition of surrounding amino acid residues significantly contribute to the determination of ligand binding sites [Soga et al., 2007].

In multi-chain proteins ligand binding sites can be located at the interface between subunits in a way that atoms of both chains contribute to the creation of a binding pocket. This is particularly the case in ligand gated ion channels.

Terminology note: in the following text we will use the terms pocket and binding site interchangeably with the following distinction: the binding site will always mean experimentally confirmed or "true" binding site whereas pocket can be just putative (or predicted) binding site depending on context.

Molecular recognition

The driving force during protein-ligand complex formation is not only shape complementarity but also physicochemical complementarity between two binding partners [Wirth et al., 2013]. However, an old hypothesis that ligand binding in proteins is like "an insertion of key into a lock" [Fischer, 1894] had to be corrected to incorporate flexibility of the protein as well as of the ligand. Binding site can wrap tighter around the ligand and we observe phenomenon called ligand induced fit. Ligands are usually relatively rigid small molecules but often exhibit rotational flexibility along some bonds.

Proteins also interact with other proteins through protein-protein binding sites, or more generally, interfaces. Protein-protein interfaces tend to be large, flat and relatively featureless, with several hotspots that contribute much of the free energy of interaction [Jubb et al., 2012]. Protein-protein binding sites and protein-ligand binding sites are not exclusive concepts and we can talk about ligand binding pockets located at protein-protein interaction interfaces. Similarly there are protein-DNA, protein-RNA and protein-membrane interfaces.

Allosteric regulation

One protein can have multiple ligand binding sits. Allosteric pockets are binding sites that bind small molecules but are different from the primary (orthosteric) binding site of a receptor or the active site of an enzyme. Binding molecules to allosteric sites causes structural changes that modulate protein behavior in respect to its main function conveyed by the active site or primary ligand binding site. Allostery is one of the most common ways of regulation of protein activity [Panjkovich and Daura, 2012, Christopoulos, 2002].

Some proteins are found to perform more than one function. This phenomenon also termed as "protein moonlighting" is being observed in growing list of proteins [Jeffery, 2009]. Existence of protein moonlighting and allosteric regulation encourages and justifies application of binding site prediction methods also to proteins of known functions.

Relationship of structure and function

Relationship between structure and function in molecular biology is an ever discussed subject. The prevailing paradigm states that structure determines function and is often interpreted to mean that there is a causal relation between structure and function. However, proteins perform functions only through interacting with other molecules. It has been estimated that proteins are able, on average, to interact with as many as five partners through a variety of binding sites. It has been suggested that "structure determines function" should be replaced by "binding determines function" [Van Regenmortel, 2002].

Nevertheless, the important conceptual consideration is this: if we have a structure we should be able to determine function, maybe only by considering relationships to other structures, but without regard on the genetic code and evolutionary relationships.

1.2 Structural Bioinformatics

1.2.1 Determination of protein structures

Work in this thesis revolves around three-dimensional structures of proteins, therefore it is important to understand methods used to obtain them and their limitations. There are two main experimental methods currently being used for determination of macromolecular structures with atomic level precision: X-ray crystallography and nuclear magnetic resonance. In the absence of experimental structures, homology modeling method is used for structural prediction in practice.

X-ray crystallography

X-ray crystallography is responsible for most of the currently known macromolecular structures. First experimentally determined high resolution protein structure was the structure of sperm whale myoglobin published in 1958 by Kendrew and Perutz. Usual procedure is that the symmetrical crystal grown out of identical proteins is repeatedly, from different angles, exposed to X-ray beam with wavelength of $\sim$1 Å. X-rays interact almost exclusively with the electrons and thus produce images of electron densities, which are then aggregated into 3D electron density maps. Positions of individual atoms are then determined and refined with the help of the knowledge of protein sequence and structures of amino acid residues. The refinement process usually involves alternating rounds of automated optimization and manual corrections [Wlodawer et al., 2008].

How precisely it is possible to determine atom coordinates depends on the resolution of the electron density map (which can be understood as a degree

of focus). However, relationship between resolution and uncertainty of atom positions in the final refined structure is not straightforward. On average, the uncertainty of the position of an atom is roughly one fifth to one tenth of the resolution [Martz, 2013]. Hydrogen atoms, having only one electron, are detectable only in the X-ray structures with resolution less than 1.2 Å. The major obstacle to achieving higher resolutions is not technology related but comes from inability to to produce sufficiently ordered crystals. Another limitation of X-ray crystallography is that position of atoms can be altered by crystal packing (particularly near crystal contacts) [Eyal et al., 2005].

Nuclear magnetic resonance

Nuclear magnetic resonance (NMR) spectroscopy started to be used for determination of protein structures in 1980's. Its use is limited to proteins of smaller molecular weight and in most favourable cases resulting structures are comparable to crystal structures with resolutions of 2–2.5 Å. Although structures obtained by NMR are not as detailed and accurate as those obtained crystallographically, the method has other advantages. It works with molecules in solvent and can be used to determine the structures of proteins and other macromolecules that fail to crystallize. Moreover, NMR can probe motions over small time scales and can be used to study protein folding dynamics and conformational changes. It is increasingly being used to confirm protein-ligand bindings.

Homology modeling

Homology or comparative modeling is based on the observation that proteins with similar sequences have similar structures, or at least similar folds. Because models determined by *de novo* structural prediction methods are regarded speculative at best, the term 'homology modeling' is used synonymously with 'structural prediction'. Even with homology modeling at least around 30% sequence identity is required for meaningful models. Structure of the protein with similar sequence (or matching secondary structure) is used as a template according to which model is build and then iteratively refined, optimized and validated. Error of homology models compared to the actual experimental structures can be up to ~1–2 Å C_α atom RMSD (root-mean-square deviation distance between corresponding C_α atoms) [Cavasotto and Phatak, 2009]. Low accuracy of protein structure prediction is still a limiting factor but sometimes it is the only available choice. For example, up until recently structures of GPCRs had to be predicted through homology modeling techniques. Between 2000 and 2007 there was only one representative structure of this class available [Adams et al., 2012].

1.2.2 PDB Database

Protein Data Bank (PDB) is the central curated publicly available database of macromolecular structures. As of July 2013 PDB contains more than 92,000 three-dimensional structures of mostly proteins and protein-ligand complexes out of which 8944 were added in the year 2012 alone. Number of annual new entries has been growing steadily in the recent years (with the exception of 2008). Current prediction is that PDB size will increase 1.5-fold between 2012 (∼85,000) and the end of 2017 (∼134,000) [Berman et al., 2013]. For contrast in the year 2000 size of PDB was only ∼13,600 structures. Source organism of around 30% deposited structures is Homo sapiens. Other structures are coming mostly from model organisms such as Escherichia coli, mouse and yeast or disease causing organisms, i. e., viruses and bacteria.

Structural Genomics initiative

In general Structural Genomics can be described as an effort to determine as many protein strutures from given genome as possible. Traditionally, determination of experimental 3D structures was driven by projects that researched proteins of interest with already known function. To fill in the protein fold space and have at least one representative structure from protein families with no experimental structural information, several Structural Genomics projects were initiated. Structural Genomics it thus source of experimentally determined protein structures of unknown function. Structural Genomics projects have also driven the developments in X-ray crystallography and NMR, so that new protein structures are solved quicker and cheaper [Cavasotto and Phatak, 2009]. It was estimated that comprehensive coverage of UNIprot could be reached in less than 15 years [Nair et al., 2009].

1.2.3 Proteome and chemical space

Human genome contains around 23,500 protein encoding genes. However, it is estimated that there are ∼150,000 different gene transcripts to mRNA due to a increase in complexity introduced in the process of gene expression [Schmidt, 2010]. The chemical space, the complete set of all possible small molecules, was estimated to be as large as 10^{30}-10^{200} structures depending on the parameters used [Bauer et al., 2010].

1.2.4 Machine learning in bioinformatics

Machine learning approaches are ideally suited for domains characterized by the presence of large amounts of data, noisy patterns, and the absence of general

theories (what bioinformatics arguably is). The fundamental idea behind them is to learn the relationships automatically from the data through a process of inference, model fitting and learning from examples. An often-met criticism of machine learning techniques is that they are black box methods: one usually cannot pin down (in simple terms) how a complex Neural Network, Support Vector Machine or Random Forest, reaches a particular answer [Baldi and Brunak, 2001].

It is important to realize however two things. Firstly, many techniques in contemporary molecular biology and pharmacology are used on a purely empirical basis. For example, the mode of action and molecular basis for the pharmacological effect and side effects of many currently used drugs remains unknown.[5] Secondly, machine learning methods applied to ever growing set of biological data can give us positive answer to the question: Is there some new relationship in the data to be exploited (and therefore new theory to be learnt)?

1.3 Pharmacology

The most important motivation for working on this problem, and all other structural bioinformatics problems for that matter, is their application in medicine. To understand where it fits into the picture we will take a brief overview of drug discovery process and recent developments in pharmaceutical industry. The fact that drug discovery is a prominent application of resulting method leads to interesting considerations.

1.3.1 Drugs and mechanism of action

Great majority of currently used drugs are small molecules that target proteins. Drugs produce their pharmacological effects by variety of different molecular mechanisms that always involve binding to a binding site.

Several possible molecular mechanisms of drug action:

- receptor agonism (“activation”) or antagonism (“blocking”)
- enzyme inhibition by binding to the active site
- allosteric modulation of receptors and enzymes
- protein-protein interaction inhibition (or stabilization)

However, the ability of a small molecule to bind to the target protein and thus potentially produce some disease-modifying/therapeutic effect is not sufficient for the compound to be used as a drug. Molecules used as drugs need to have others

[5] 7% of approved drugs are purported to have no known primary target, and up to 18% lack a well-defined mechanism of action [Elisabet et al., 2012]. This statistics, however, do not consider side effects and exact structural mechanism.

necessary or desirable properties:

- water solubility (as it needs to be carried in aqueous blood)
- fat solubility (to be able to cross cell membrane)
- high bioavailability (fraction of the dose that reaches circulation)
- low molecular weight
- low toxicity
- high selectivity

In practice many of these properties are contradictory objectives. Low selectivity means that drug is bind to several proteins (other than desired target) and is the major cause of drug side effects. It is important to note that selectivity as well as toxicity is relative and depend on compound concentration.

1.3.2 Concept of drug-likeness

A traditional method to evaluate drug-likeness is Lipinski's 'rule-of-five' (RO5) [Lipinski et al., 1997, Lipinski, 2000], which consists of four important properties, each related to the number 5:

- molecular mass ≤ 500 Da
- H-bond acceptors [6] ≤ 10
- H-bond donors [7] ≤ 5
- octanol-water partition coefficient[8] $\log P \leq 5$

The rule is based on data in the literature for a large number of compounds, including all known drugs, that correlate physical properties with oral bioavailability [Lombardino and Lowe, 2004]. It has been associated with 90% of orally active drugs that have achieved phase II clinical status [Lipinski, 2004]. Lipinski's rule was formulated to guide chemistry to desired direction and predict success (or failure) of candidate molecules in clinical trials. As with many other rules of thumb, there are many exceptions. RO5 is tied specifically to oral bioavailability and the meaning of "drug-like" is thus dependent on mode of administration. Subsequently many other drug-likeness definitions/indices correcting or extendning RO5 based on statistical analysis of drug databases have been developed [Ghose et al., 1999, Oprea, 2000, Veber et al., 2002]. Lipinski himself acknowledged that defining drug-like by what exists in databases leads to the criticism that most of chemical space will be undefined and that discovery opportunities in unexplored chemistry space will be limited [Lipinski, 2004].

[6] expressed as the sum of OHs and NHs in the chemical structure

[7] expressed as the sum of N and O atoms

[8] measure of how lipophilicity – provides an estimate of the ability of the compound to pass through a cell membrane

1.4 Problem definition and scope

1.4.1 Binding site detection

We can state the problem in the following way: given the protein of known (static) structure, predict which areas around the protein surface are likely to bind an unspecific ligand. Additional information like protein sequence or structure evolutionary conservation may be considered. There is an implicit constraint on the ligand to be a small, relatively rigid molecule and the identified binding sites to have a meaningful size compared to the size of the protein. Too small ligands like individual ions are not biologically interesting. Too big ligands like longer peptides can theoretically be of arbitrary length. For both of those cases specialized approaches are used that take advantage of distinct properties of ions or peptides.

Often several pockets are detected at a protein surface and it is necessary to have some method to select relevant ones. Most of pocket detection programs incorporate some scoring and ranking that is supposed to represent likelihood of the pocket to accept small molecule ligand. Pocket detection programs are often compared by their ability to find true (experimentally confirmed) binding site within top-1 and top-3 detected pockets.

1.4.2 Pocket characterization & druggability

In the recent decade there has been some confusion with the term druggability. Druggability is usually understood as ability of the protein or binding site to bind small drug-like molecule with high affinity. Term druggability used in this sense is thus linked to the notion of drug-likeness [Schmidtke, 2011] but as we mentioned before traditional scope of drug-likeness has been recently challenged. Many druggability prediction methods try to predict not just the *ability* to bind small drug-like molecules bot also *how much* "drug-like" (in the sense of ADMET) those molecules can potentially be at the same time. This gives results that are biased towards chemical subspace of typical currently used drugs. Some authors argued that the term druggability implies too much and suggested the term ligandability (structural druggability) [Hajduk et al., 2005a].

In our opinion it is not clear where (nor it is useful) to make a clear line between pocket ranking problem (used in pocket detection programs) and structural druggability prediction. Indeed, many of the druggability/ligandability prediction methods output score that can be used to reorder pockets found by any pocket detection method and thus lead to better results when considering only first n found pockets. Some pocket detection programs use very simplistic scoring to rank pockets (such as volume) but other have more complex scoring functions that could be viewed as ligandability prediction methods in their own right. The difference

is mainly in datasets they are trained and tested on and their results are thus interpreted differently. The only substantial difference is that recently published druggability methods claim to (or are trained to) distinguish categorically between druggable and non-druggable sites.

1.4.3 Inherent limitations and difficulties

First and apparent limitation comes from the fact that proteins are to the some extent flexible. We can even ask the question: does it make sense to work on algorithms that predict binding sites on static structures when we know that proteins are flexible? Protein dynamics is so intricate that is extremely difficult to incorporate protein flexibility directly into the pocket detection algorithm. As we mentioned before, a small conformational change on one side of the protein may result in a big change on the other side. Although simplified flexibility models based on geometric constraints or residue rotamer libraries have been developed, they reflect only certain aspects of protein flexibility. So far the only way how to meaningfully predict protein flexibility and conformational changes is molecular dynamics simulation. Pocket detection can then be performed on static snapshots of molecular dynamics simulation with subsequent aggregation of the results (like it is done by MDPocket method [Schmidtke et al., 2011]). Static structures is what we currently have available. We expect that structural databases will be gradually enriched by snapshots of different conformations of the same protein.

Chapter 2

Current approaches

Binding site detection methods have been in development for almost 30 years now (first known method being published in 1985). During this time more than 30 different algorithms or improvements have been published [Schmidtke, 2011]. Especially in the recent five years we have observed increased interest in this field (indicated by the number of recently published reviews) and also influx of original or improved methods. It seems that very recently focus has shifted from mere pocket detection to characterization, that is, ranking and druggability prediction of binding sites. In this chapter we will overview the most influential and some of the most recent methods. Table 2.1 contains representative list of pocket detection methods. It is not our goal, however, to list and review all existing methods but rather introduce variety of pocket detection and druggability prediction strategies and somewhat capture developments in the field.

2.1 Pocket detection methods

2.1.1 Geometry based methods

Numerous geometrical methods has been published that focus mainly on the algorithmic side of the problem of finding concave pockets and clefts on the surface of 3D structure. From this simpler methods here we mention only LIGSITE and PocketPicker. None of the other methods can be considered purely geometrical as they consider also other factors such as physicochemical properties. Geometrical methods in general do not consider placement of hydrogen atoms. This makes sense, since most of the known protein structures are not resolved in a sufficient resolution and hydrogen atoms are not necessarily needed to identify concave areas.

Name	Date	Type	Server	Executable	Source
GRID	1985	energy-based			
POCKET	1992	geometric			
LIGSITE	1997	geometric			
PASS[1]	2000	geometric		✓	
MOE Site Finder	2001	geometric		$	
Q-SiteFinder	2005	energy-based	✓	✓	
ICM-PocketFinder[2]	2005	energy-based	✓	$	
LIGSITE^{csc}	2006	evolutionary		✓	✓
CASTp[3]	2008	geometric			✓
PocketPicker	2007	geometric		✓	✓
SiteMap[4]	2007	energy-based		$	
FINDSITE	2008	evolutionary	✓		✓
PocketDepth[5]	2008	geometric	✓		
MetaPocket	2009	consensus	✓		
Fpocket	2009	geometric	✓	✓	✓
ConCavity	2009	evolutionary		✓	✓
SiteHound	2009	energy-based		✓	
McVol[6]	2010	geometric			✓
POCASA[7]	2010	geometric	✓		
MetaPocket 2.0	2011	consensus	✓		
MSPocket[8]	2011	geometric		✓	✓
COFACTOR[9]	2012	evolutionary	✓		

Table 2.1: List of representative pocket detection methods

LIGSITE [Hendlich et al., 1997]

LIGSITE is a simple geometric grid based method. At first regular Cartesian grid is superimposed over the protein and grid points are separated to solvent accessible (outside of the protein) and solvent inaccessible according to the minimum distance from any of the protein's atoms. For every solvent accessible point scanning is done along the 7 axes (three Cartesian axes and four cubic diagonals) for protein-solvent-protein (PSP) "events", that is, weather on the both side of the line protein grid point is to be found. Grid points are that way scored from 0 to 7, number that represents their buriedness. Pockets and cavities are then defined as regions of grid points with a minimum number of PSP events.

[1] [Brady and Stouten, 2000]
[2] [An et al., 2005]
[3] [Binkowski et al., 2003]
[4] [Halgren, 2007]
[5] [Kalidas and Chandra, 2008]
[6] [Till and Ullmann, 2010]
[7] [Yu et al., 2010]
[8] [Zhu and Pisabarro, 2011]
[9] [Roy and Zhang, 2012]

Method itself is a slight extension of older POCKET algorithm (that used only 3 x, y, z axes in buriedness scanning step) [Levitt and Banaszak, 1992], but was in turn used as a basis for numerous improvements and as a reference method for comparison with new methods. Parameters like grid step, minimum number of PSP events and minimum number of grid points in a pocket are user adjustable and method does not address ranking of the found pockets. The disadvantage, as with other grid based approaches, is the dependence on the grid orientation.

PocketPicker [Weisel et al., 2007]

Method is very similar to above mentioned LIGSITE in that it uses a 3D grid and calculates a buriedness of grid points around protein surface. PocketPicker uses a finer and optimized scanning approach. Scans are being performed along 30 directions that are approximately equally distributed around a grid probe. 30 search rays of length 10 Å and width 0.9 Å are checked for presence of a protein atom. Probing whether search ray tube contains any atoms is optimized by dividing all protein atoms into neighborhoods and searching only in neighborhoods that tube intersects. Probes with buriedness-indices ranging from 16 to 26 are then clustered into pockets.

MOE Site Finder [Labute and Santavy, 2001]

Site Finder was released as a part of chemical software package MOE 2001.01 from the The Chemical Computing Group.[10] It is based on the concept of *alpha spheres* and *Voronoi tessellation* of space. An alpha sphere is a sphere that contacts four atoms on its boundary and contains no internal atoms. Voronoi vertex in 3D is the point at which four Voronoi regions intersect and is equally distant from four atoms at the center of those regions. The Voronoi vertices thus coincide with the centers of alpha spheres. The problem of finding Voronoi tessellation and Vornoi vertices is equivalent to the dual problem of finding *Delaunay triangulation* which can be converted to the much simpler problem of finding a convex hull in 4D space.

Pocket detection is then based on the observation that alpha spheres that are too small are located in tightly packed interiors of proteins and big alpha spheres are located at an exposed surface regions. Medium sized alpha spheres are likely to be found in relatively enclosed concave pockets. Alpha spheres are filtered according to size and hydrophobicity (hydrophobic sphares with no hydrophilic neighbours are eliminated). Remaining alpha spheres are subsequently clustered with single-linkage algorithm and too small clusters are removed. Ranking of resulting pockets is done according to hydrophobicity calculated as the number of hydrophobic atoms within a contact distance of any of the alpha spheres of the pocket. The advantage over grid based methods is independence of results on the grid placement and generally better time and space complexity.

[10] http://www.chemcomp.com/

Fpocket [Le Guilloux et al., 2009]

Fpocket released as an open source project is another alpha sphere theory based method. The basic principle is much like that of MOE Site Finder but more complicated filtering, clustering and refinement procedure is employed. Fpocket also uses more advanced scoring function based on several pocket descriptors. For calculating Vornoi vertices Fpocket relies on publicly available package Qhull[11] [Barber et al., 1996]. Clustering og alpha spheres is done in three steps: (i) clustering based of Vornoi vertice neighbourhood, (ii) single linkage clustering of centers of mass of vertice clusters andfinally, (iii) multiple linkage clustering of resulting clusters. At the end too small and hydrophylic pockets are dropped.

Scoring scheme relies on five of extracted pocket descriptors such as normalized number of alpha spheres, hydrophobic density and proportion of apolar alpha spheres. Final scoring function was optimized by partial least squares fitting. Fpocket program has several user-adjustable parameters that influence refinement and clustering process. Program can be thus directed to output many smaller pockets or fewer larger pockets. Default values were determined by an semi combinatorial/empirical optimization on a training set with the two main goals: good pocket-ligand center distance and also good ligand coverage/overlap.

2.1.2 Energy based methods

Energy based methods incorporate some level of physics into the pocket identification process by attempting to calculate binding potentials or binding energies [Hajduk et al., 2005b]. Energetic methods usually build upon some existing force field (software that approximates physicochemical forces on molecular level) and are generally computationally more demanding than other approaches.

GRID [Goodford, 1985]

GRID is the first known published and at the same time first energy based computational procedure that can be used for determining binding sites. It is based on calculation of non-bonded interaction energies between the probe molecule and the target protein at position an a three-dimensional grid. GRID does not identify binding pockets per se, but interaction sites in the target of interest [Henrich et al., 2010].

Q-SiteFinder [Laurie and Jackson, 2005]

Method places a methyl group ($-CH_3$) probe on a grid of 0.9 Å resolution to calculate van der Waals interaction energies between the protein and probes.

[11] http://www.qhull.org/

Energy calculations are made with GRID energy field. Energetically favourable probe coordinates are then clustered according to their spatial proximity and total interaction energies for clusters are calculated. Clusters are then ranked according to their interaction energies such that highest interaction energy corresponds with the first predicted binding site.

SiteHound [Ghersi and Sanchez, 2009]

SiteHound uses Molecular Interaction Fields to identify protein structure regions that show a high propensity for interaction with ligands. SiteHound works in a similar way to Q-SiteFinder but allows to use multiple probes for the detection of different types of binding sites. Another improvement lies in the use of alternative hierarchical clustering algorithm, which improve results for ligands of different shapes.

2.1.3 Evolutionary or threading based methods

Sequence-based approaches are based on the presumption that functionally important residues are preferentially conserved during the evolution, because natural selection acts on function. [Roy and Zhang, 2012]

LIGSITE^{csc} [Huang and Schroeder, 2006]

LIGSITE^{csc} improved on original LIGSITE algorithm in two ways. The first (purely geometric) extension is LIGSITE^{cs}, in which protein's Connolloy (solvent-excluded) surface is used to more precisely calculate buriedness of the grid probe. Monitoring of protein-solvent-protein events is replaced by monitoring surface-solvent-surface events around the probe. Secondly, in LIGSITE^{csc} top three pocket detected by LIGSITE^{cs} are re-ranked according to degree of conservation of amino acid residues around pockets. The average conservation of the residues within 8 Å of the center of a pocket is used as a score for re-ranking.

Conservation scores of amino acid residues are obtained from ConSurf-HSSP database (now ConSurf-DB), which provides evolutionary conservation estimates for proteins of known structure in the PDB [Glaser et al., 2005]. Degree of conservation is assigned by Rate4Site algorithm for scoring amino acid residue conservation based on their calculated evolutionary rate. This algorithm takes into account the phylogenetic relationships between the homologous proteins and the stochastic nature of the evolutionary process. Calculation is based on alignment of similar protein sequences and empirical Bayesian inference.

FINDSITE [Brylinski and Skolnick, 2008, Skolnick and Brylinski, 2009]

FINDSITE method combines evolutionary and structural approach. It is based on observation that even distantly homologous proteins usually have similar folds and bind ligands at similar location. At first ligand-bound structural templates are selected from the database of already known protein-ligand complexes by a threading (fold recognition) algorithm. Used threading algorithm PROSPECTOR_3 is not based just on sequence similarity but combines various scoring functions designed to match structurally related target/template pairs [Skolnick et al., 2004]. Found homologous structures are aligned with the target protein by a global structural alignment algorithm. Ligands on superimposed template structures are then clustered into consensus binding sites. Authors showed that FINDSITE is less sensitive to the distortions in the input structure and was shown to retain its accuracy even on (erroneous) structures predicted by homology modeling (in comparison to LIGSITE^{csc}).

ConCavity [Capra et al., 2009]

Authors presented sequence conservation based improvement of geometrical LIGSITE method. Jensen-Shannon divergence method has been used to calculate conservation scores. This method praviously shown state of the art performance in predicting functionally important residues [Capra and Singh, 2007]. Unlike in LIGSITE^{csc} sequence conservation information is not used just to re-rank pockets but is integrated directly into the pocket detection procedure. Grid points around the protein are assigned a score that represents likelihood to overlap a ligand atom. This score is the combination of buriedness-index (as in LIGSITE) and sequence conservation of neighboring residues. Authors demonstrated that this results in a more accurate pocket shapes, i. e., pockets that have better overlap with larger fraction of the ligand volume. Authors also shown that if only sequence conservation information is considered for pocket prediction, it results to poorer results than considering geometrical information alone. This is due to the fact that residues can be conserved for other reasons than ligand binding e. g. stabilizing structural fold.

2.1.4 Consensus methods

MetaPocket [Huang, 2009]

MetaPocket is consensus based method that in recent update MetaPocket 2.0 [Zhang et al., 2011] aggregates results of 8 different pocket detection algorithms (among them forementioned LIGSITE^{cs}, Q-SiteFinder, Fpocket and ConCavity). Authors demonstrated that MetaPocket performed better than any of the individual methods. All of the individual methods output pockets ranked by the different scoring function. Z-score is therefore calculated separately for each site in different methods to make them comparable. MetaPocket 2.0 takes only first three pockets

from each method into account. The total of 24 pockets are then agregated into meta-pockets with the help of a hierarchical clustering algorithm that identifies overlapping sites. Then the z-score of each cluster is calculated and serves as a scoring function to re-rank final meta-pocket sites.

2.2 Druggability prediction

SiteMap [Halgren, 2009]

Halgren developed druggability prediction model as part of commercial program SiteMap [Halgren, 2007] from Schrödinger LLC.[12] SiteMap itself is a grid based geometrical ligand binding site detection and characterization method.[13] Pocket ranking in SiteMap is done by scoring function *SiteScore* (2.1) based on only three descriptors: number of site grid points that is reflecting pocket size (n), degree of enclosure (e) and hydrophilicity score (p). Druggability scoring function *Dscore* (2.2) is using the same properties but with different coefficients.

$$\text{SiteScore} = 0.0733\sqrt{n} + 0.6688e - 0.2p \tag{2.1}$$

$$\text{Dscore} = 0.094\sqrt{n} + 0.6e - 0.324p \tag{2.2}$$

Dscore is in principle the same model as SiteScore except it was optimized on different dataset containing druggable/difficult/undruggable sites. SiteScore is thus deemed to represent likelihood of the pocket to bind any ligand and Dscore likelihood to bind drug-like ligand specifically. This model reflects traditional view that ligand binding sites are usually the biggest, most buried and hydrophobic pockets.

DrugPred [Krasowski et al., 2011]

Druggability is here understood as "ability of the putative binding site to bind orally available molecules with high affinity". Authors compiled a new non-redundant dataset (termed NRDLD) of 71 druggable and 44 less druggable proteins (mostly enzymes) and used it do develop and train structure-based druggability predictor. Druggable class contains only proteins that were experimentally confirmed to bind drug-like molecule that adhere to the strict criteria, one of which was Lipinski's rule of five. Less druggable class contains pockets that were experimentally confirmed to bind (some) ligand but no drug-like ligand binding has been reported.

Several 1-dimensional pocket descriptors have been defined: 16 capturing

[12] chemical software company, http://www.schrodinger.com/

[13] pocket detection method itself has not been reviewed mainly for the lack of reported details

polarity, size and compactness and 40 capturing amino acid composition. All descriptors were tested for normal distribution and only 22 descriptors that appeared to be normally distributed were retained. Dataset was split to training and validation set containing 76 and 37 proteins respectively and linear model was then trained using partial least-squares projection to latent structures discriminant analysis (PLS-DA). Number of descriptors have been further reduced in a manual iterative process by removing descriptors with weak predictive power.

Final model is a linear combination (2.3) of five descriptors:

- relative polar surface area (psa_r)
- total hydeophobic surface area (hsa_t)
- contact surface area (csa)
- relative occurrence of hydrophobic amino acids (haa)
- sum of the hydrophobicity indices of amino acids ($hiaa$)

$$\text{DrugPred} = -0.2psa_r + 1.16hsa_t + 0.11csa + 0.22haa + 0.22hiaa + 1.3 \quad (2.3)$$

Although three of five components reflect pocket hydrophobicity, removing either of them would result in worse model. Interesting here as well is the interplay between relative and absolute value descriptors (roughly reflecting pocket size squared), that, in our view, encourages the use of non-linear model.

VolSite [Desaphy et al., 2012]

Dataset from DrugPred study has been used to train Support Vector Machine (SVM) based druggability predictor. The term druggability is here understood in its weaker form (as authors reiterated) to mean "structural druggability (ligandability)", although model was trained on the same dataset as DrugPred method. Input of the model are descriptors derived from a grid based characterization of a binding site. At first cube lattice of 14^3 cells, each cell having 1.5 Å edge, is imposed over a pocket. Buriedness index is calculated for each cell in a similar way than in PocketPicker method (as proportion of 120 regularly spaced rays from the cell that are intersecting protein). Cells are separated into three classes: cells that are part of the protein (IN), cavity cells, cells outside of the cavity and protein (OUT). Every cell that is inside of a pocket is assigned one of 8 labels that represents physicochemical properties: H-bond acceptor, H-bond donor, H-bond acceptor and donor, negative ionizable, positive ionizable, hydrophobic, aromatic or null. Label is assigned according to the closest protein atom (in such a way that it is complementary to the characteristic physicochemical property of the amino acid residue atom is from).

From this arrangement 73 descriptors are calculated:

- (1) total number of cavity cells
- (2-9) proportion of cells having each of the 8 physicochemical labels
- (10-73) histograms of buriedness of cells from from each of the 8 labels (8×8 ranges)

Dataset was split in the same way as in DrugPred study into training set (76 entries) and validation set (37 entries). SVM model with RBF (radial basis function) kernel was trained in a 5-fold cross-validation procedure with systematic optimization of *c* and *gamma* parameters.

2.3 Summary

2.3.1 Recent reviews

Although first algorithm appeared almost three decades ago, systematic development of binding site detection methods can be observed mainly in last ten years. Review articles started to appear only in last few years. First comprehensive review of pocket detection and characterization methods focused on conceptual differences between geometric/energy/evolutionary approaches [Henrich et al., 2010]. Leis et al. reviewed protein-ligand as well as protein-protein binding site detection methods [Leis et al., 2010]. Authors presented visual side-by-side comparison of results of five ligand binding site prediction servers on the small set of proteins and emphasized influence of ligand induced protein conformational changes on prediction success. Another review focused on drug discovery applications of pocket detection and pocket similarity methods [Pérot et al., 2010]. Latest review [Chen et al., 2011] represents the first independent attempt to systematically assess performance of pocket detection methods (although only a limited number of methods have been compared). Results of this benchmark will be discussed in section 2.3.3 (page 33).

Review of several pocket detection and druggability prediction approaches as well as a critical view on the field can be found in PhD Thesis of P. Schmidtke [Schmidtke, 2011], one ot the two principal authors of Fpocket. Schmidtke points out to the "identification paradox": while pocket detection methods are most often evaluated by identification success among top-1 or top-3 ranked pockets, identification and ranking are two completely independent tasks. Pure identification success is rarely reported and it is very likely that many programs allow identification of all binding sites.[14] He argues that modern pocket identification algorithms should be acknowledged to be excellent and more focus should be directed towards pocket classification and scoring.

[14] has been reported for Fpocket (refering to unpublished data) and at least one other method [Singh et al., 2011]

So far the longest and most complete list of different binding site detection methods was very recently compiled in [Lumipuu, 2013].

High level overview of the new druggability prediction field, introducing computational as well as experimental methods can be found in [Barril, 2012]. Another review of binding site analysis and druggability prediction approaches can be found in [Nisius et al., 2012].

2.3.2 Availability

In terms of availability, several of these methods can be used as web servers but only few are available as distributable packages [Le Guilloux et al., 2009, Ghersi and Sanchez, 2011]. Methods exposed as web services are usually free but can be used only on one individual protein structure at a time and it is impossible to incorporate them into virtual screening pipelines for multiple targets. Even fewer of these methods have been released as open source software. Comprehensive reviews of availability of different methods can be found in [Leis et al., 2010, Khazanov, 2012].

2.3.3 Performance comparison

So far we have not discussed performance of reviewed methods. In case of pocket detection problem, field suffers from the lack of widely established benchmarking dataset and clearly defined assessment procedure. Studies that introduced new methods have compared them to few existing solutions with the expected conclusion that the proposed method performs better or as well as existing solutions. Comparison of different methods is further complicated by limited availability and technical differences (e.g some methods define pocket boundaries by listing pocket surface atoms or binding residues while others define binding site only by a center point). Another problem that complicates objective comparison are differences in methodologies that are being used for assessing success rate of methods.

Dataset of 48 bound/unbound structures introduced in LIGSITE^{csc} study [Huang and Schroeder, 2006] is practically the only dataset that was steadily used to compare different methods. Great majority of the methods reported success rate of around 90% considering Top-3 and above 70% considering Top-1 predicted binding sites. The size of the dataset and small differences in results, however, limit how meaningful comparison based on this dataset are.

First independent systematic comparison between a representative set of methods was presented in already mentioned critical review [Chen et al., 2011]. Authors created a new non-redundant benchmarking dataset of 251 proteins containing more than 400 binding sites (which we term CHEN2011 dataset). Every structural

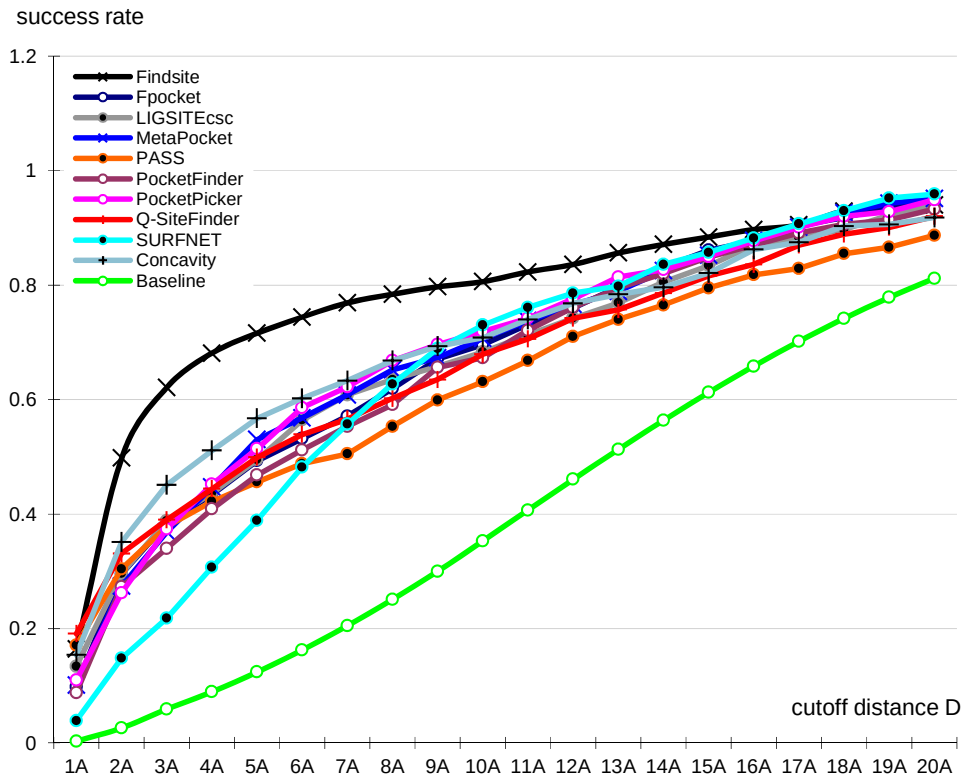


Figure 2.1: Results of ten pocket detection methods on CHEN2011 dataset. Identification success is measured using D_{CA} criterion (the minimal distance between center of the binding site to any atom of the ligand). Binding site is considered correctly predicted if the minimal distance between that site and Top-n predicted sites is bellow the cutoff distance D (x-axis), where n is the number of true binding sites on the particular protein in dataset. Cutoff distance of 4 Å was commonly used in previous studies. Methods are also compared against a baseline predictor that randomly selects a surface patch on the target protein. (Underlying data have not been published. Reused with permission. [Chen et al., 2011])

protein family is represented by one protein.[15] Unlike in previous studies and datasets that followed 1-protein/1-binding site rule, proteins in this dataset are annotated with multiple binding sites (in case more binding sites for given protein are they are known). Acknowledging the fact that proteins have multiple binding sites and incorporating it into new dataset is a major conceptual contribution of this study. While other studies usually based evaluation on Top-1, Top-3 and/or Top-5 predicted pockets, here Top-n predictions are used for evaluation. That is, for protein with n binding sites, top n predictions for each methods are considered. Moreover, instead of using one arbitrary distance threshold, success rate was evaluated across integer thresholds from 1 Å to 20 Å (figure 2.1).

Results show that evolutionary structure-conservation based method FINDSITE clearly outperforms other methods, especially when considering (in our view) meaningful threshold distances 4–10 Å. However, authors also showed that success rate of FINDSITE quickly drops with decreasing maximal structural similarity

[15] based on SCOP structural classification of proteins

between query protein and proteins in template library. Performance of all other methods (below 50% on threshold 4 Å) was substantially lower that previously published results on other datasets. Authors showed that predictions from different methods are complementary and demonstrated it by designing a simple consensus based predictor that outperforms best single method. Another presented conclusion is that considerable fraction (over 30% [16]) of the binding sites is not identified by any of the considered methods.

2.4 Conclusions

Different approaches to pocket identification will find use in different scenarios. If we have only one or a small number of protein targets at hand, we can perform detailed individual inspections of proteins manually. The combination methods is likely to give the most meaningful results. In fact, in this case there is no reason not to use both evolutionary conservation and energetic methods to get all possible informations/predictions about the investigated protein. On the other hand, fast geometric methods with knowledge based scoring functions will find use in massive virtual screenings.

Although evolutionary based methods currently give the best results, they are dependent on similarity/homology between the target protein and proteins in the database they work with. The most successful FINDSITE method has additional disadvantage in that it works only with single-chain proteins and is not able to detect binding sites on the boundaries between subunits. There is also a case for methods that use just structural information. Every year more than 8,000 protein structures are added to the PDB and it is useful to scan all of them for possible new binding sites that can be than used to discover new interactions with ligands.

2.4.1 Pitfalls of current methods

Size of and exact boundaries of predicted pockets are arbitrary. We can usually increase success rate of an algorithm on a given dataset by tuning its parameters to give us pockets with bigger/smaller size. When it comes to pocket ranking and druggability prediction methods this could be a problem because almost all of them are based on (global) pocket descriptors based on the shape and size of the whole pocket.

[16] using threshold of 4 Å considering Top-n predictions

Chapter 3

Proposed Improvement

We have developed a pocket ranking function based on prediction model that predicts ligandability (likelihood to bind a ligand) of a given point inside of a pocket. Prediction is done considering only a local physicochemical and geometric properties derived from point's neighbourhood.

3.1 Initial considerations

The basic assumption was that the the physicochemical properties calculated from local neighbourhood of a given point can predict ligandability.

Fpocket

Fpocket [Le Guilloux et al., 2009] was selected as a basis four our solution because it is elegant and fast geometric algorithm. Of all pocket detection methods that are available as open source, Fpocket is best documented and it is well maintained as a software project.

Fpocket is an geometric method that is based on Vornoi tessellation of space and clustering of alpha-spheres (discussed on p. 26). Pocket as defined by Fpocket is a cluster of alpha spheres that are touching atoms on a protein surface. The center of an alpha-sphere is Vornoi vertex that is equally distant to the four closest atoms. Alpha-spheres that constitute a pocket have a radius in range 3–6 Å (this range is parameterizable but for our research default parameters of Fpocket were used).

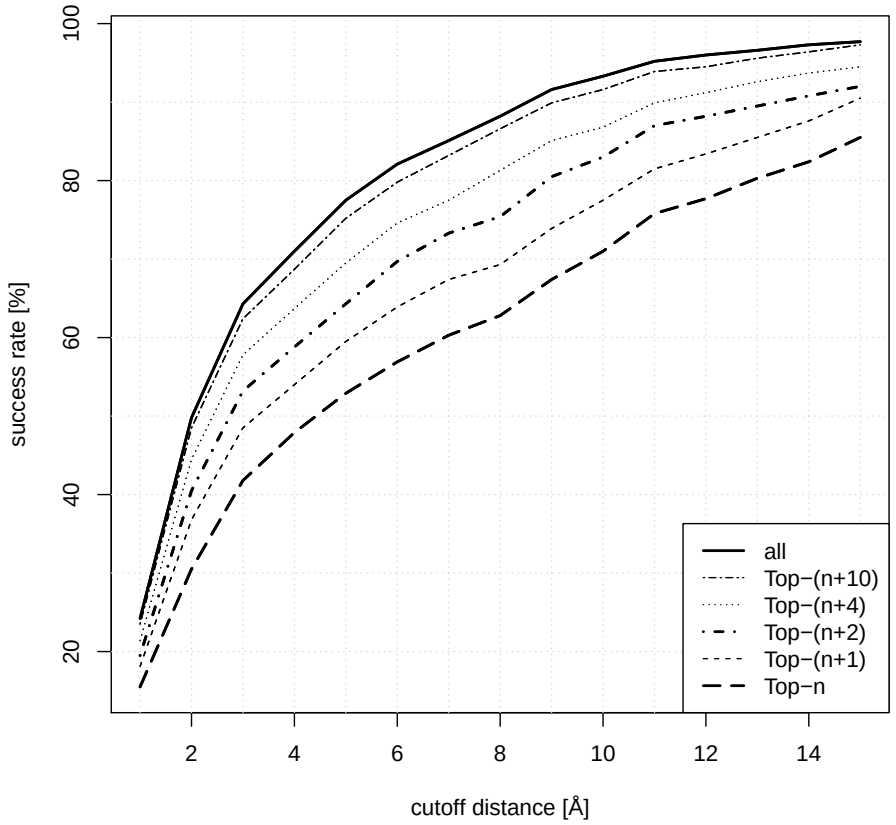


Figure 3.1: Success rate of Fpocket on CHEN2011 dataset considering gradually larger sets of pockets ranked at the top. Figure shows room for improvement by using better pocket ranking function. Binding site is considered correctly predicted if the minimal distance between that site and Top-n predicted sites is bellow the cutoff distance D (x-axis), where n is the number of true binding sites on the particular protein in dataset. Compare with Fig. 2.1 (p. 34).

The importance of pocket ranking

Results of the benchmark study (p. 34) compare success rate of pocket detection methods based on considering only Top-n predicted pockets. To find out pure identification success of Fpocket we re-run Fpocket on CHEN2011 dataset and analyzed the results. We used the same evaluation criteria as in original study (D_{CA} metric) but we looked at success rates considering larger sets of pockets ranked at the top (see Figure 3.1). Ultimately, the pure identification success (success rate taking all predicted pocket) was considered. When considering probably the most meaningful and commonly used cutoff distance of 4 Å this number is much higher than for commonly considered Top-1/Top-3 pockets. Although (using default parameters), Fpocket was not able to detect all binding sites (as was recently boldly suggested by one of the Fpocket authors – see p. 32), results show that there is a clear room for improvement of F-pocket results by creating a better pocket scoring function.

Most of pocket detection methods find much more pockets on a give structure than there is actual binding sites. For example Fpocket outputs average of 12.4 pockets for one protein on CHEN2011 dataset.

3.2 Materials and Methods

3.2.1 Data sets

- **CHEN2011** New dataset of 251 proteins containing 476 ligands that was used to benchmark pocket detection methods in recent comparative review [Chen et al., 2011]. It can be considered a "hard" dataset as most of methods performed unexpectedly poorly.
- **UB48** Datast of 48 unbound/bound structures [Huang and Schroeder, 2006] contains 48 liganated proteins in bound and unbound state. This dataset has been most widely used for comparing pocket detection methods and can be considered as "easy" dataset. We have joined bound and unbound subsets into one dataset of 96 proteins containing altogether 106 binding sites.
- **ASTEX** Astex Diverse set is a collection of 85 proteins containing 133 ligands that was introduced as a benchmarking dataset for molecular docking methods [Hartshorn et al., 2007].

In the case of UB48 and ASTEX datasets some protein structures contained very small ligands. We have decided to ignore all ligands smaller than 8 atoms. This step was not absolutely necessary as it affected only handful of structures. We don't use these datasets to compare our method with other methods by directly comparing results from literature. They are used only to train and compare different classifiers and results of our ranking function (with respect to original Fpocket ranking that is evaluated by the same criteria). On the other hand, dataset and methodology used are the same as in original CHEN2011 study, so the results are directly comparable.

3.3 Proposed method

3.3.1 Vector of physicochemical properties

First we define vector of physicochemical properties PC which can be assigned to any atom on the protein/pocket surface based on the amino acid residue.

Table 3.1: Vector PC of amino acid residue properties. Most of properties have absolute (0/1) character

Property	Value
`hydrophobic`	1 for hydrophobic residues
`hydrophilic`	1 for hydrophilic residues
`hydrophatyIndex`	side-chain hydrophaty index [1] with values in range $\langle -4.5, 4.5 \rangle$
`aliphatic`	1 for aliphatic residues
`aromatic`	1 for aromatic residues
`sulfur`	1 for sulfur containing residues
`hydroxyl`	1 for hydroxyl group containing residues
`basic`	1 for basic residues
`acidic`	1 for acidic residues
`amide`	1 for amide group containing residues
`charge`	1 for positively charged and -1 for negatively charged residues
`hBondDonor`	1 for H-bond donor residues
`hBondAcceptor`	1 for H-bond acceptor residues
`hBondDonorAcceptor`	1 for residues that are H-bond donors & acceptor residues
`polar`	1 for polar residues
`ionizable`	1 for ionizable residues

Aggregation function

To calculate property vector of point inside a pocket property vectors of the pocket surface atoms are aggregated into one using simple aggregation function. Contribution of particular surface atom is weighted by its distance to the inner pocket point.

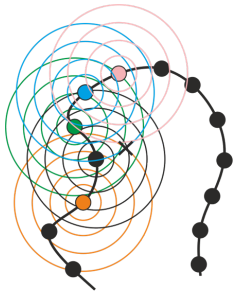


Figure 3.2: Schematic depiction of calculating physicochemical properties of the inner pocket point (X) from the properties of pocket surface atoms

Atomic neighbourhood of vertex V:

$$\mathrm{A}(V) = \{\text{pocket surface atoms within 8 Å radius around V}\} \tag{3.1}$$

Aggregated physicochemical properties of vertex V calculated from its atomic neighbourhood:

$$\mathrm{PC}(V) = \frac{1}{m} \sum_{\mathrm{A}_i \in \mathrm{A}(V)}^{m} \mathrm{PC}(\mathrm{A}_i) \cdot w(dist(V, \mathrm{A}_i)) \tag{3.2}$$

Weight function:

$$w(d) = \begin{cases} 1, & d \leq 4\ \text{Å} \\ (4/d)^2 & d\ > 4\ \text{Å} \end{cases} \tag{3.3}$$

Additional properties

We also define vector WA of additional properties for given alpha-sphere that is calculated from its 8 Å neighbourhood.

Table 3.2: Vector WA of additional pocket aplha sphere properties used as an input for classifier

Property	Value
`atoms`	absolute number of atoms within 8 Å radius
`atomDensity`	sum of atoms weighted by distance
`atomC`	sum of carbon atoms weighted by distance
`atomO`	sum of oxygen atoms weighted by distance
`atomN`	sum of nitrogen atoms weighted by distance
`hDonorAtoms`	sum of H-bond donor atoms weighted by distance
`hAcceptorAtoms`	sum of H-bond acceptor atoms weighted by distance
`vornoiNeighbours`	absolute number of pocket Vornoi vertices in 4 Å radius
`alphaSphereRadius`	radius of the alpha sphere
`Ala ... Val`	relative occurrence of amino acid residues weighted by distance. represents 20 properties each for one standard amino acid

Final vector

Final vector describing pocket alpha sphere neighbourhood has 45 dimensions.

$$(\mathrm{PC}, \mathrm{WA}) \in R^{45}$$

Most of the properties of the vector PC and some of the properties from WA have values that fin in range $\langle 0, 1 \rangle$. Absolute value properties have their upper bound given by practical restrictions (there can be only so much atoms packed in the sphere with radius 8 Å).

3.3.2 Classifiers

To predict ligandability of a given alpha sphere we employed and compared various machine learning methods that we trained on datasets of vectors extracted from Fpocket outputs.

However, it became clear that the performance of our new pocket scoring function is limited more by data and set of features used rather than performance of particular prediction model (however measured). Selecting a best machine learning model based on a fixed dataset was not the main focus of our work. We tried and compared different methods including Random Forests, Support Vector Machines, Neural networks with very similar results which will be discussed in the next chapter.

Random Forests

Random Forests [Breiman, 2001] is an ensemble machine learning method that constructs set decision trees each using only random subset of data vectors considering random subset of features.

Random Forests have been used throughout our research because of several inherent advantages. The generalization error for forests converges to a limit as the number of trees in the forest increases. Moreover Random Forests are generally very robust with respect to noise and train fast on large datasets.

3.3.3 Ranking function

Classifiers that we used can — additionally to predicted class — output histograms of class probabilities. In our case of binary classification it is ordered pair $[P_{class0}, P_{class1}]$. Pocket score is then sum of predicted probabilities of all alpha-spheres that constitute a pocket (represended by their centers – Vorni vertices $\{V_i\}$).

$$\text{PocketScore} = \sum_{i=1}^{n} P_{class1}(V_i) \tag{3.4}$$

Originally we experimented with relative pocket score (divided by n, number of alpha spheres), however we found that absolute score steadily gives us better results. Using simply the sum of predicted probabilities has two advantages:
(i) Size of predicted pocket do not always match the size of the ligand. For putative pockets of unusually large size it is more important that there is a presence of ligand-binding sub-pocket rather than relative ligand-bindability of all pocket volume.
(ii) Non-relative pocket score gives bias to the larger pockets, which is actually desirable as ligands are usually found in largest pockets. If we had a random

classifier our pocket score would practically reorder pockets by size, whereas a perfect classifier would still distinguish true binding site from decoy pockets. Performance of pocket score based solely on pocket size thus forms the bottom line of our scoring function on which we can improve even by poorly-performing classifiers.

Chapter 4

Evaluation and Results

4.1 Classification results

Feature vectors have been extracted from Fpocket outputs on considered datasets of protein-ligand complexes creating corresponding datasets of ligand binding and non-binding alpha-spheres. Random Forest classification models have been trained on given vector datasets to predict ligandability of alpha-spheres.

4.2 Evaluation methods

Pocket detection criteria

D_{CA} (distance between pocket center and any atom of the ligand) has been used as positive hit criterion.

4.3 Pocket detection results

Despite the difficulties with generalization / data sampling our new scoring function significantly improved on the results given by default Fpocket scoring function. Generally speaking improvement is around 5-7% points when considering cutoff in the meaningful range 3-5% and Top-n to Top-$n+1$ pockets. This may not seem dramatic but consulting Fig. 2.1 show that our method would outperform all other methods that are not using any external evolutionary information.

Next we present of our scoring function on several datasets. We have been careful to never evaluate pocket scoring function on the same dataset that we trained the classifier.

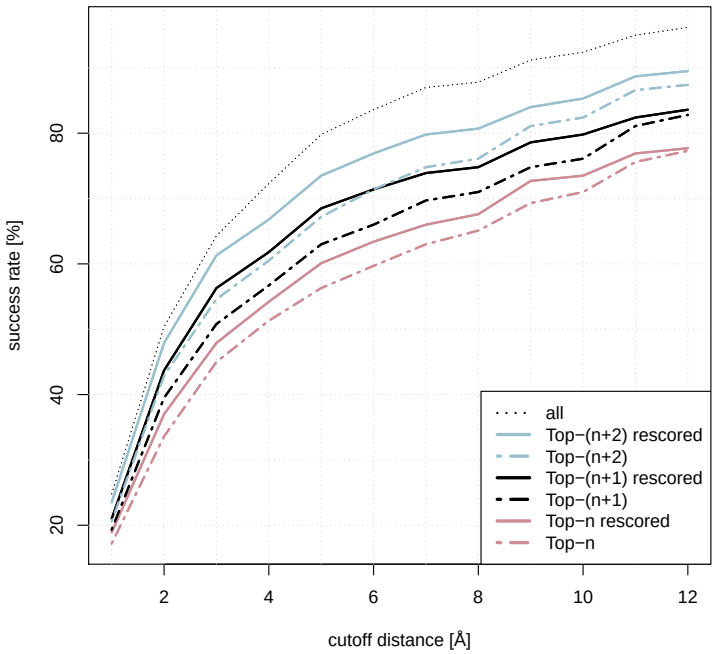


Figure 4.1: Success rates on CHEN2011_HALF1 dataset.
Classifier: Random Forest trained on CHEN2011_HALF2 dataset.

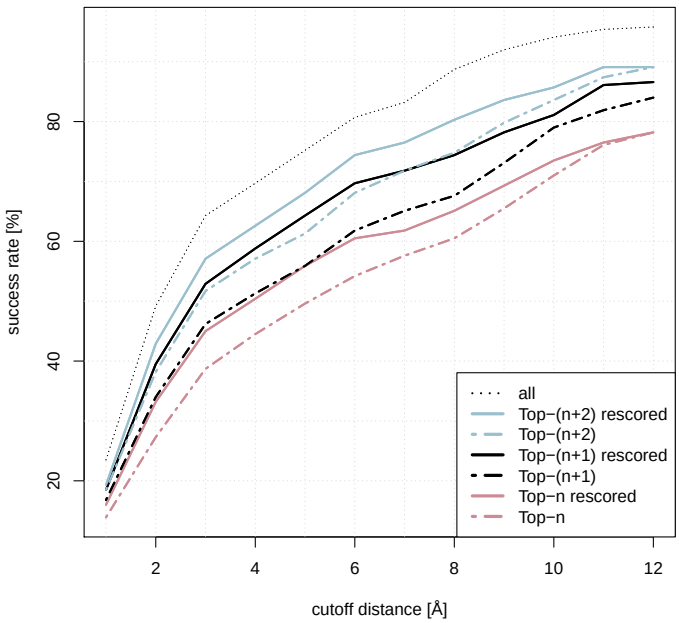


Figure 4.2: Success rates on CHEN2011_HALF2 dataset.
Classifier: Random Forest trained on CHEN2011_HALF1 dataset.

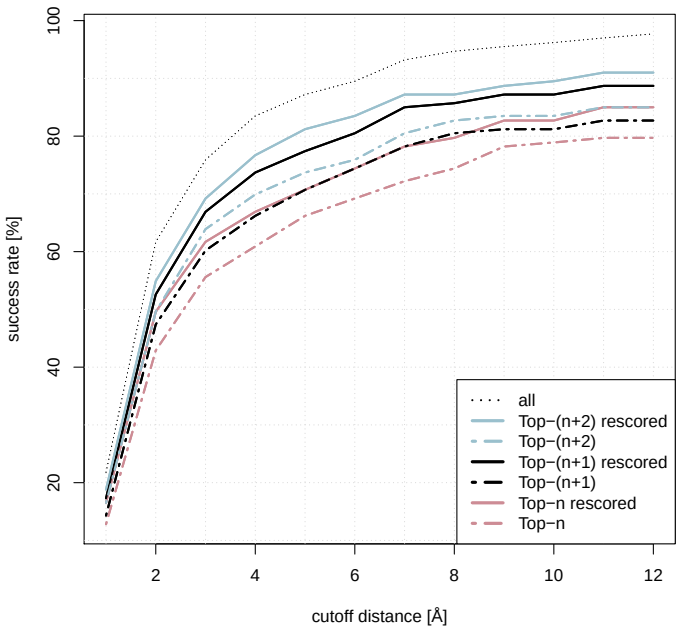


Figure 4.3: Success rates on ASTEX dataset.
Classifier: Random Forest trained on UB48 dataset.

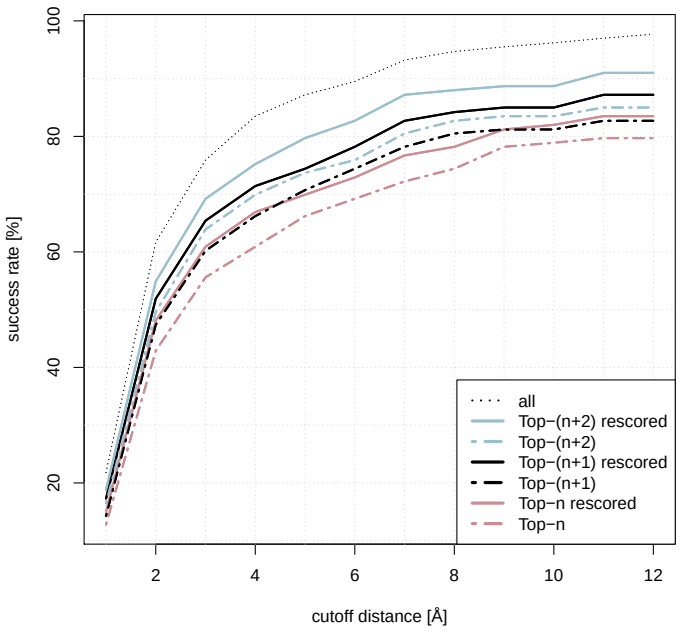


Figure 4.4: Success rates on ASTEX dataset.
Classifier: Random Forest trained on CHEN2011 dataset.

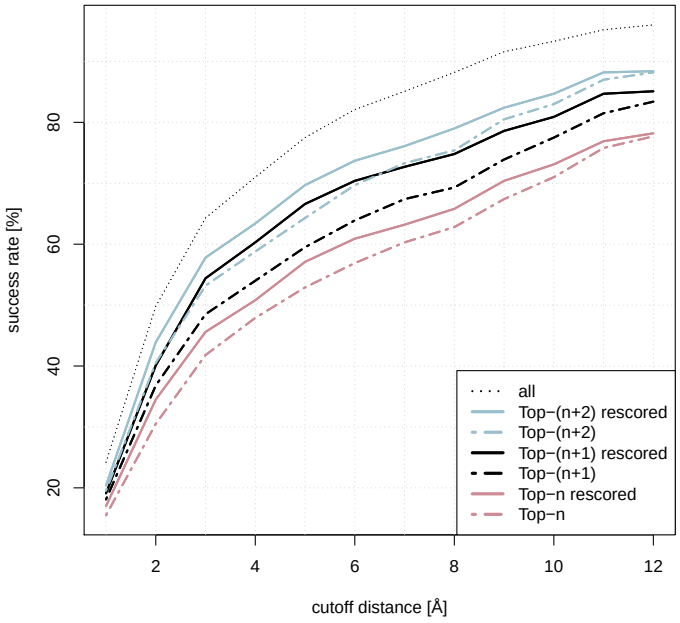


Figure 4.5: Success rates of on CHEN2011 dataset.
Classifier: Random Forest trained on UB48 dataset.

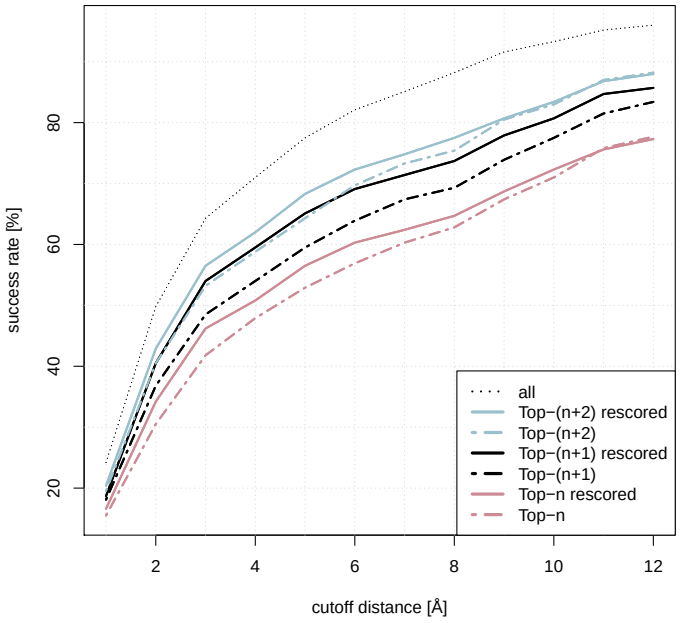


Figure 4.6: Success rates on CHEN2011 dataset.
Classifier: Random Forest trained on ASTEX dataset.

Chapter 5

Summary and Outlook

5.1 Summary and contributions

We have been successful in developing a novel pocket ranking method that outperforms default scoring function of Fpocket. Our scoring function is based on solely on local structural information and learned predictive model. Unlike any other scoring functions employed in current pocket detection algorithms, our approach does not consider global pocket descriptors (which is an advantage since boundaries of predicted pockets are arbitrary), neither consults external databases for evolutionary conservation information.

5.2 Future work

Druggability prediction method

The same model that we used to distinguish between true binding sites and decoy pockets can possibly be trained to distinguish between druggable and undruggable binding sites.

Fragment based pocket similarity

Confirmation of the hypothesis that local neighbourhood can predict ligandability suggests that it would be possible to create pocket similarity method that is fragment based and not dependent on global shape of predicted pockets which boundary is arbitrary. This naturally leads to fast ligand complementarity and rough binding affinity estimation method. Fragment based character will allow creation of indexable binding site database that can be query by similar pocket or complementary ligand.

Making method publicly available

There are three options: (1.) Reimplementation as a new scoring function in Fpocket. (2.) Stand-alone program - this would mean to re-implement Vornoi tessellation and alpha sphere clustering but would have few advantages: (i) ability to use classifier to make more meaningful pocket boundaries and possibly detect pockets currently ignored by Fpocket (ii) easily implementable multi-threading and simple distributed computation for large datasets (3.) Web service (shiny but in our opinion not as useful).

Bibliography

[Adams et al., 2012] Adams, R., Worth, C., Guenther, S., Dunkel, M., Lehmann, R., and Preissner, R. (2012). Binding sites in membrane proteins–diversity, druggability and prospects. *European journal of cell biology*, 91(4):326–339. 18

[An et al., 2005] An, J., Totrov, M., and Abagyan, R. (2005). Pocketome via comprehensive identification and classification of ligand binding envelopes. *Molecular & cellular proteomics : MCP*, 4(6):752–761. 25

[Baldi and Brunak, 2001] Baldi, P. and Brunak, S. a. (2001). *Bioinformatics: The Machine Learning Approach, Second Edition (Adaptive Computation and Machine Learning)*. The MIT Press, 2 edition. 20

[Barber et al., 1996] Barber, C. B., Dobkin, D. P., and Huhdanpaa, H. (1996). The quickhull algorithm for convex hulls. *ACM TRANSACTIONS ON MATHEMATICAL SOFTWARE*, 22(4):469–483. 27

[Barril, 2012] Barril, X. (2012). Druggability predictions: methods, limitations, and applications. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. 33

[Bauer et al., 2010] Bauer, R., Wurst, J., and Tan, D. (2010). Expanding the range of 'druggable' targets with natural product-based libraries: an academic perspective. *Current opinion in chemical biology*, 14(3):308–314. 19

[Berman et al., 2013] Berman, H., Coimbatore Narayanan, B., Di Costanzo, L., Dutta, S., Ghosh, S., Hudson, B., Lawson, C., Peisach, E., Prlić, A., Rose, P., Shao, C., Yang, H., Young, J., and Zardecki, C. (2013). Trendspotting in the protein data bank. *FEBS letters*, 587(8):1036–1045. 19

[Binkowski et al., 2003] Binkowski, T., Naghibzadeh, S., and Liang, J. (2003). CASTp: computed atlas of surface topography of proteins. *Nucleic acids research*, 31(13):3352–3355. 25

[Brady and Stouten, 2000] Brady, G. and Stouten, P. (2000). Fast prediction and visualization of protein binding pockets with PASS. *Journal of computer-aided molecular design*, 14(4):383–401. 25

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45. 41

[Brylinski and Skolnick, 2008] Brylinski, M. and Skolnick, J. (2008). A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proceedings of the National Academy of Sciences of the United States of America*, 105(1):129–134. 29

[Capra et al., 2009] Capra, J., Laskowski, R., Thornton, J., Singh, M., and Funkhouser, T. (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS computational biology*, 5(12). 29

[Capra and Singh, 2007] Capra, J. and Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics (Oxford, England)*, 23(15):1875–1882. 29

[Cavasotto and Phatak, 2009] Cavasotto, C. and Phatak, S. (2009). Homology modeling in drug discovery: current trends and applications. *Drug discovery today*, 14(13-14):676–683. 18, 19

[Chen et al., 2012] Chen, J., Hanson, B., Fisher, S., Langan, P., and Kovalevsky, A. (2012). Direct observation of hydrogen atom dynamics and interactions by ultrahigh resolution neutron protein crystallography. *Proceedings of the National Academy of Sciences of the United States of America*, 109(38):15301–15306. 12

[Chen et al., 2011] Chen, K., Mizianty, M., Gao, J., and Kurgan, L. (2011). A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. *Structure (London, England : 1993)*, 19(5):613–621. 32, 33, 34, 38, 57

[Christopoulos, 2002] Christopoulos, A. (2002). Allosteric binding sites on cell-surface receptors: novel targets for drug discovery. *Nature reviews. Drug discovery*, 1(3):198–210. 16

[Creighton, 1993] Creighton, T. (1993). *Proteins: structures and molecular properties*. W.H. Freeman. 14

[Desaphy et al., 2012] Desaphy, J., Azdimousa, K., Kellenberger, E., and Rognan, D. (2012). Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *Journal of chemical information and modeling*, 52(8):2287–2299. 31

[Elisabet et al., 2012] Elisabet, G., Setola, V., Hert, J., Crews, B., Irwin, J., Lounkine, E., Marnett, L., Roth, B., and Shoichet, B. (2012). Identifying mechanism-of-action targets for drugs and probes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(28):11178–11183. 20

[Eyal et al., 2005] Eyal, E., Gerzon, S., Potapov, V., Edelman, M., and Sobolev, V. (2005). The limit of accuracy of protein modeling: influence of crystal packing on protein structure. *Journal of molecular biology*, 351(2):431–442. 18

[Fischer, 1894] Fischer, E. (1894). Einfluss der configuration auf die wirkung der enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3):2985–2993. 16

[Ghersi and Sanchez, 2009] Ghersi, D. and Sanchez, R. (2009). EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics (Oxford, England)*, 25(23):3185–3186. 28

[Ghersi and Sanchez, 2011] Ghersi, D. and Sanchez, R. (2011). Beyond structural genomics: computational approaches for the identification of ligand binding sites in protein structures. *Journal of structural and functional genomics*, 12(2):109–117. 33

[Ghose et al., 1999] Ghose, A. K., Viswanadhan, V. N., and Wendoloski, J. J. (1999). A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *Journal of combinatorial chemistry*, 1(1):55–68. 21

[Glaser et al., 2005] Glaser, F., Rosenberg, Y., Kessel, A., Pupko, T., and Nir, B. (2005). The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. *Proteins*, 58(3):610–617. 28

[Goodford, 1985] Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, 28(7):849–857. 27

[Hajduk et al., 2005a] Hajduk, P., Huth, J., and Tse, C. (2005a). Predicting protein druggability. *Drug discovery today*, 10(23-24):1675–1682. 22

[Hajduk et al., 2005b] Hajduk, P. J., Huth, J. R., and Tse, C. (2005b). Predicting protein druggability REVIEWS. 10(23). 27

[Halgren, 2007] Halgren, T. (2007). New method for fast and accurate binding-site identification and analysis. *Chemical biology & drug design*, 69(2):146–148. 25, 30

[Halgren, 2009] Halgren, T. (2009). Identifying and characterizing binding sites and assessing druggability. *Journal of chemical information and modeling*, 49(2):377–389. 30

[Hartshorn et al., 2007] Hartshorn, M., Verdonk, M., Chessari, G., Brewerton, S., Mooij, W., Mortenson, P., and Murray, C. (2007). Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of medicinal chemistry*, 50(4):726–741. 38

[Hendlich et al., 1997] Hendlich, M., Rippmann, F., and Barnickel, G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of molecular graphics & modelling*, 15(6):359–63, 389. 25

[Henrich et al., 2010] Henrich, S., Outi, S., Huang, B., Rippmann, F., Cruciani, G., and Wade, R. (2010). Computational approaches to identifying and characterizing protein binding sites for ligand design. *Journal of molecular recognition : JMR*, 23(2):209–219. 27, 32

[Horowitz and Trievel, 2012] Horowitz, S. and Trievel, R. (2012). Carbon-oxygen hydrogen bonding in biological structure and function. *The Journal of biological chemistry*, 287(50):41576–41582. 11, 12

[Huang, 2009] Huang, B. (2009). MetaPocket: a meta approach to improve protein ligand binding site prediction. *Omics : a journal of integrative biology*, 13(4):325–330. 29

[Huang and Schroeder, 2006] Huang, B. and Schroeder, M. (2006). LIGSITEcsc: predicting ligand binding sites using the connolly surface and degree of conservation. *BMC structural biology*, 6. 28, 33, 38

[Jeffery, 2009] Jeffery, C. (2009). Moonlighting proteins–an update. *Molecular bioSystems*, 5(4):345–350. 16

[Jubb et al., 2012] Jubb, H., Higueruelo, A., Winter, A., and Blundell, T. (2012). Structural biology and drug discovery for protein-protein interactions. *Trends in pharmacological sciences*, 33(5):241–248. 16

[Kalidas and Chandra, 2008] Kalidas, Y. and Chandra, N. (2008). PocketDepth: a new depth based algorithm for identification of ligand binding sites in proteins. *Journal of structural biology*, 161(1):31–42. 25

[Karshikoff and Jelesarov, 2008] Karshikoff, A. and Jelesarov, I. (2008). Salt bridges and conformational flexibility: Effect on protein stability. *BIOTECHNOLOGY AND BIOTECHNOLOGICAL EQUIPMENT*, 22(1). 11

[Khazanov, 2012] Khazanov, N. (2012). *Large-Scale Analysis of Protein-Ligand Binding Sites Using the Binding MOAD Database*. PhD thesis, University of Michigan. 33

[Krasowski et al., 2011] Krasowski, A., Muthas, D., Sarkar, A., Schmitt, S., and Brenk, R. (2011). DrugPred: a structure-based approach to predict protein druggability developed using an extensive nonredundant data set. *Journal of chemical information and modeling*, 51(11):2829–2842. 30

[Kryukov et al., 2003] Kryukov, G. V., Castellano, S., Novoselov, S. V., Lobanov, A. V., Zehtab, O., Guigó, R., and Gladyshev, V. N. (2003). Characterization of mammalian selenoproteomes. *Science*, 300(5624):1439–1443. 10

[Kyte and Doolittle, 1982] Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105 – 132.

[Labute and Santavy, 2001] Labute, P. and Santavy, M. (2001). Locating binding sites in protein structures. `http://www.chemcomp.com/journal/sitefind.htm`. (Online; accessed 2013-07-16). 26

[Laskowski et al., 1996] Laskowski, R., Luscombe, N., Swindells, M., and Thornton, J. (1996). Protein clefts in molecular recognition and function. *Protein science : a publication of the Protein Society*, 5(12):2438–2452. 15

[Laurie and Jackson, 2005] Laurie, A. and Jackson, R. (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics (Oxford, England)*, 21(9):1908–1916. 27

[Le Guilloux et al., 2009] Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10. 27, 36

[Le Guilloux et al., 2009] Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10:168. 33

[Leis et al., 2010] Leis, S., Schneider, S., and Zacharias, M. (2010). In silico prediction of binding sites on proteins. *Current medicinal chemistry*, 17(15):1550–1562. 32, 33

[Levitt and Banaszak, 1992] Levitt, D. G. and Banaszak, L. J. (1992). Pocket: A computer graphies method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of Molecular Graphics*, 10(4):229 – 234. 26

[Lipinski, 2000] Lipinski, C. (2000). Drug-like properties and the causes of poor solubility and poor permeability. *Journal of pharmacological and toxicological methods*, 44(1):235–249. 21

[Lipinski, 2004] Lipinski, C. A. (2004). Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies*, 1(4):337 – 341. 21

[Lipinski et al., 1997] Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1–3):3 – 25. <ce:title>In Vitro Models for Selection of Development Candidates</ce:title>. 21

[Lombardino and Lowe, 2004] Lombardino, J. and Lowe, J. (2004). The role of the medicinal chemist in drug discovery–then and now. *Nature reviews. Drug discovery*, 3(10):853–862. 21

[Lüllmann, 2005] Lüllmann, H. (2005). *Color Atlas of Pharmacology*. Basic sciences. Thieme Georg Verlag. 12

[Lumipuu, 2013] Lumipuu, M. (2013). Computer-aided identification of the binding sites of protein-ligand complexes. Master's thesis, University of Eastern Finland, Faculty of Health Sciences, School of Pharmacy. 33

[Martz, 2013] Martz, E. (2013). Nature of 3d structural data @online. `http://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/nature_of_3d_structural_data.html`. (Online; accessed 2013-07-10). 18

[Nair et al., 2009] Nair, R., Liu, J., Soong, T., Acton, T., Everett, J., Kouranov, A., Fiser, A., Godzik, A., Jaroszewski, L., Orengo, C., Montelione, G., and Rost, B. (2009). Structural genomics is the largest contributor of novel structural leverage. *Journal of structural and functional genomics*, 10(2):181–191. 7, 19

[Nisius et al., 2012] Nisius, B., Sha, F., and Gohlke, H. (2012). Structure-based computational analysis of protein binding sites for function and druggability prediction. *Journal of biotechnology*, 159(3):123–134. 15, 33

[Oprea, 2000] Oprea, T. (2000). Property distribution of drug-related chemical databases. *Journal of computer-aided molecular design*, 14(3):251–264. 21

[Panjkovich and Daura, 2012] Panjkovich, A. and Daura, X. (2012). Exploiting protein flexibility to predict the location of allosteric sites. *BMC bioinformatics*, 13:273. 16

[Pérot et al., 2010] Pérot, S., Sperandio, O., Miteva, M., Camproux, A., and Villoutreix, B. (2010). Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug discovery today*, 15(15-16):656–667. 32

[Petsko and Ringe, 2004] Petsko, G. and Ringe, D. (2004). *Protein Structure and Function*. Primers in Biology. NSP, New Science Press. 14

[Pettit and Bowie, 1999] Pettit, F. K. and Bowie, J. U. (1999). Protein surface roughness and small molecular binding sites. *Journal of Molecular Biology*, 285(4):1377 – 1382. 15

[Roy and Zhang, 2012] Roy, A. and Zhang, Y. (2012). Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure (London, England : 1993)*, 20(6):987–997. 25, 28

[Schmidt, 2010] Schmidt, W. (2010). Current concepts of pharmacogenetics, pharmacogenomics, and the "druggable" genome. In Müller, M., editor, *Clinical Pharmacology: Current Topics and Case Studies*, pages 205–223. Springer Vienna. 19

[Schmidtke, 2011] Schmidtke, P. (2011). *Protein-ligand binding sites Identification, characterization and interrelations*. PhD thesis, University of Barcelona. 22, 24, 32

[Schmidtke et al., 2011] Schmidtke, P., Axel, B., Luque, F., and Barril, X. (2011). MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics (Oxford, England)*, 27(23):3276–3285. 23

[Schrödinger, 1944] Schrödinger, E. (1944). *What is Life?: The Physical Aspect of the Living Cell*. What is Life?: The Physical Aspect of the Living Cell. Cambridge, Cambridge University Press. 7

[Singh et al., 2011] Singh, T., Biswas, D., and Jayaram, B. (2011). AADS–an automated active site identification, docking, and scoring protocol for protein targets based on physicochemical descriptors. *Journal of chemical information and modeling*, 51(10):2515–2527. 32

[Skolnick and Brylinski, 2009] Skolnick, J. and Brylinski, M. (2009). FINDSITE: a combined evolution/structure-based approach to protein function prediction. *Briefings in bioinformatics*, 10(4):378–391. 29

[Skolnick et al., 2004] Skolnick, J., Kihara, D., and Zhang, Y. (2004). Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins*, 56(3):502–518. 29

[Soga et al., 2007] Soga, S., Shirai, H., Kobori, M., and Hirayama, N. (2007). Use of amino acid composition to predict ligand-binding sites. *Journal of chemical information and modeling*, 47(2):400–406. 15

[Till and Ullmann, 2010] Till, M. and Ullmann, G. (2010). McVol - a program for calculating protein volumes and identifying cavities by a monte carlo algorithm. *Journal of molecular modeling*, 16(3):419–429. 25

[Tsvetkov et al., 2009] Tsvetkov, P., Reuven, N., and Shaul, Y. (2009). The nanny model for IDPs. *Nature chemical biology*, 5(11):778–781. 14

[Van Regenmortel, 2002] Van Regenmortel, M. (2002). A paradigm shift is needed in proteomics: 'structure determines function' should be replaced by 'binding determines function'. *Journal of molecular recognition : JMR*, 15(6):349–351. 17

[Veber et al., 2002] Veber, D., Johnson, S., Cheng, H., Smith, B., Ward, K., and Kopple, K. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of medicinal chemistry*, 45(12):2615–2623. 21

[Voet and Voet, 2010] Voet, D. and Voet, J. G. (2010). *Biochemistry, 4th Edition*. Wiley, 4 edition. 10, 11, 12, 13, 57

[Ward et al., 2004] Ward, J., Sodhi, J., L, M., Buxton, B., and Jones, D. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of molecular biology*, 337(3):635–645. 14

[Weisel et al., 2007] Weisel, M., Proschak, E., and Schneider, G. (2007). PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central journal*, 1. 26

[Wirth et al., 2013] Wirth, M., Volkamer, A., Zoete, V., Rippmann, F., Michielin, O., Rarey, M., and Sauer, W. (2013). Protein pocket and ligand shape comparison and its application in virtual screening. *Journal of computer-aided molecular design*. 16

[Wlodawer et al., 2008] Wlodawer, A., Minor, W., Dauter, Z., and Jaskolski, M. (2008). Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *The FEBS journal*, 275(1):1–21. 17

[Yu et al., 2010] Yu, J., Zhou, Y., Tanaka, I., and Yao, M. (2010). Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics (Oxford, England)*, 26(1):46–52. 25

[Zhang et al., 2011] Zhang, Z., Li, Y., Lin, B., Schroeder, M., and Huang, B. (2011). Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics (Oxford, England)*, 27(15):2083–2088. 29

[Zhu and Pisabarro, 2011] Zhu, H. and Pisabarro, M. (2011). MSPocket: an orientation-independent algorithm for the detection of ligand binding pockets. *Bioinformatics (Oxford, England)*, 27(3):351–358. 25

List of Figures

List of Tables