Computer-aided identification of the binding sites of protein-ligand complexes

Markus Lumipuu
Master's thesis
University of Eastern Finland
Faculty of Health Sciences
School of Pharmacy
June 2013

UNIVERSITY OF EASTERN FINLAND, Faculty of Health Sciences
School of Pharmacy: Pharmaceutical Chemistry
LUMIPUU, MARKUS SAKARI: Computer-aided identification of the binding sites of protein-ligand complexes
Master's Thesis, 126 pages, 1 appendix (3 pages)
Supervisor: Prof. Ph.D. (Chem.) Antti Poso
June 2013

Keywords: Binding site, binding site identification, molecular modelling, drug development

Abstract

The foundation of drug development is the best possible knowledge of the drugs functional target. In order to be able to study the functional target of a drug, it is crucial to know the binding site of the drug molecule. NMR or X-ray crystallography can be used to define the binding site, but it is not always possible. In such cases, one strives to predict the binding site computationally by special computer-aided identification methods. The most common binding site prediction approaches, a few newer more complex methods and a total of 77 methods are analyzed closer in this Master's thesis. The goal is to bring out the differences between methods, their strengths and weaknesses. In addition, 10 test sets, which were developed for the estimation of binding site prediction ability, are presented and their usability is estimated. Three different databases are used to compare several methods and this enables the reliable comparison of binding site prediction ability of 5, 9 and 20 method sets (in total 26 different methods) with each other.

In the experimental part of this work it is attempted find out what the differences are between good and not so good binding pockets. This study is focused on finding any single or multiple connectivity factors, which would explain the difference in affinities of binding molecules. The basic assumption was that the ligand binding affinity value does not increase significantly above a certain molecular weight. In his study the focus is on small molecule ligands (M_w = 100-600 Da), which are not bound covalently to target protein. The binding affinity database Binding MOAD was used as the source of research material for this study. Only the complexes, in which the binding affinity is reported as a K_i value, are taken into consideration. The material was processed into a comparable form and 127 of 1492 complexes from the chosen inspection range were taken into closer analysis. A new database of these 127 complexes is created. The 3D-structure of each complex was downloaded from the PDB. Each of these downloaded complexes was analyzed visually with the molecular modeling software MOE and all discovered interactions were collected. This collected data and other computational parameters from MOE and SiteMap were added to the created database. In order to find the reason for the difference in binding affinities special focus was put on lipophilicity, molecular weight, and the number of H-bonds. In addition, a new binding site classification method, which contains 7 classes and bases on the position and shape of the binding site, was created.

ITÄ-SUOMEN YLIOPISTO, Terveystieteiden tiedekunta
Farmasian laitos: Farmaseuttinen kemia
LUMIPUU, MARKUS SAKARI: Tietokoneavusteinen proteiini-ligandi kompleksien sitoutumispaikkojen tunnistaminen
Pro Gradu –tutkielma, 126 sivua, 1 liite (sivuja 3)
Ohjaaja: Professori FaT Antti Poso
Kesäkuu 2013

Avainsanat: Sitoutumispaikka, sitoutumispaikan tunnistaminen, molekyylimallinnus, lääkekehitys

Tiivistelmä

Lääkekehityksen peruskivenä toimii lääkkeenvaikutuskohteen mahdollisimman hyvä tunteminen. Jotta vaikutuskohdetta voidaan tutkia perusteellisesti, on tärkeää tuntea vaikutuskohteen eli sitoutumispaikan sijainti kohdeproteiinissa. Nykyisin vaikutuskohta voidaan monesti tunnistaa NMR tai röntgenkristallografia kuvantamisella, mutta aina se ei onnistu. Tällöin vaikutuskohta pyritään määrittämään laskennallisesti erityisillä sitoutumispaikan identifiointimenetelmillä. Tässä pro gradu –tutkielmassa esitellään yleisimpiä sitoutumispaikan tunnistusperiaatteita ja analysoidaan lähemmin 77 eri sitoutumispaikan tunnistusmetodia. Menetelmien eroavaisuuksia, vahvuuksia ja heikkouksia on pyritty tuomaan esille ja lisäksi muutamia tuoreita sitoutumispaikan tunnistusmenetelmiä on arvioitu. Lisäksi on esitelty 10 menetelmien arviointiin kehitettyä testitietokantaa ja arvioitu niiden käytettävyyttä. Kolmen eri testitietokannan avulla on pystytty vertailemaan luotatettavasti 5, 9 ja 20 metodin (yhteensä 26 eri metodin) sitoutumispaikan määrittämiskykyä keskenään.

Tämän lopputyön kokeellisessa osassa pyrittiin selvittämään mitä eroja on hyvillä ja huonoilla sitoutumispaikoilla. Työssä keskityttiin erityisesti löytämään yhtä tai useampaa tekijää, joka voisi selittää yleisellä tasolla sitoutumisaffiniteettien eroja. Alkuoletuksena oli, ettei ligandien sitoutumisaffiniteetti kasva keskimääräisesti enää tietyn molekyylipainon jälkeen. Tutkimuksessa keskityttiin pienmolekyylisten ligandien (M_w = 100-600 Da) ja proteiinien komplekseihin, joissa sitoutuminen ei tapahdu kovalenttisesti. Tutkimusaineiston pohjana käytettiin BindingMOAD sitoutumisaffiniteettitietokantaa, josta keskityttiin vain komplekseihin joiden sitoutumisaffiniteetti oli ilmoitettu K_i muodossa. Aineisto käsiteltiin vertailukelpoiseksi ja aineistosta otettiin tarkasteluvälille sijoittuneesta 1492 kompleksista 127 lähempään tarkasteluun. Näistä luotiin oma tietokanta. Jokaisesta lähemmin tarkasteltavasta molekyylistä ladattiin kiderakenteet PDB:stä, joista jokaista tarkasteltiin visuaalisesti molekyylimallinnus ohjelma MOE:n avulla ja laskettiin kaikkien todennäköisten interaktioiden lukumäärät, mitkä tallennettiin tietokantaan. Lisäksi kerättiin ja tallennettiin useita erilaisia sekä MOE että SiteMap ohjelmien laskennallisia parametreja. Affiniteettien eroavaisuuksien syyn selvittämisessä keskityttiin erityisesti lipofiilisyyteen, molekyylipainoon ja vetysidosten lukumäärään. Lisäksi luotiin seitsemänluokkainen luokittelujärjestelmä sitoutumispaikkojen sijainnin ja muodon perusteella.

Acknowledgements

I made this Master's Thesis at Faculty of Health Sciences, School of Pharmacy, Pharmaceutical Chemistry, University of Eastern Finland during the years 2011-2013. I would firstly like to thank my supervisor, Prof. Ph.D. Antti Poso, for a most interesting topic of study, his valuable help and patient guidance throughout the research project. I would thank Ph.D. Tuomo Laitinen for practical advices and technical support with the modeling programs MOE and Maestro. Also I would thank M.Sc. Åsmund Kaupang for interesting discussions and relaxing musical moments in the modeling lab. Colleague Tatu Pantsar and my cousin Tomi Mononen I would thank for revision of the language of my Master's Thesis.

Finally, I would like to thank my lovely wife Kerttu and my children Hilla and Haiku for their most incredible patience while I was out of circulation in these years.

Table of Contents

List of Abbreviations

2D	Two-dimensional
3D	Three-dimensional
ABPA	Almost buried hydrogen bonds
aMD	Accelerated molecular dynamics
APO	Proteins 3D structure without bound ligand structure
BD	Blind Docking
Cα	Alpha carbon
CFG	Conserved Functional Group
$-CH_3$	Methyl molecule
C.O.G.	Center of gravity
CSV	Comma-separeted values
DCLM	Double Cubic Lattice Method
DD	Druggability data set
DPA	Dynamics Perturbation Analysis
E_{el}	Energy calculated by the Electrostatic function
E_{lj}	Energy calculated by the Lennard-Jones Function
E_{xyz}	Combination of three calculated interaction energies: Directional hydrogen bond function, E_{el}, and E_{lj})
ET	Evolutionary Trace
FASTA	Text-based format for presenting amino acid sequences
FC	Functional Confidence
FD	Focused Docking
FDPB	Finite Difference solution of the Poisson-Boltzman equation
FFT	Fast Fourier Transform
G_{elec}	Electrostatic free energy
HOLO	Proteins 3D-structure with bound ligand structure
IC_{50}	Concentration at which the enzyme inhibition level is 50 %
JSD	Jensen-Shannon divergence
K_a	Acid dissociation constant
K_d	Dissociant constant

K_i	Binding affinity of an inhibitor
ML	Maximum Likehood
MP	Maximum Parsimony
MPK1	MetaPocket
MPK2	MetaPocket version 2.0
MSA	Multiple Sequence Alingment
MSMS	Program for molecular surface calculation
M_W	Molecular weight
NJ	Neighbour-Joining
NRDD	Nonredundant version of druggability data set (DD)
OMIM	Online Mendelian Inheritance in Man
PDB	Protein Data Bank
PDF	Protein-Depth Flag
PLD	Protein Ligand Database
PRFM	Parameterized Reaction Field Multipoles
SAS	Solvent accessible surface
SASA	Solvent accessible surface area
SES	Solvent excluded surface
SMILES	Simplified Molecular Input Entry Specification
SOM	Self-organizing maps
SPF	Single Point Flag
SS	SequenceSpace
UPGMA	Unweighted Pair Group Method with Arihtmetic Mean
VD	Volume Depth
vdW	van der Waals
vdW-FFT	van der Waals -fast Fourier transform
V_e	Estimation of pocket volume

1 Introducing

Finding selective drugs against specific diseases is the ultimate goal in today's pharmaceutical research (Volkamer *et al.* 2012). Before these can be developed, their targets must be identified. The development of such targeted drugs requires the precision and detailed knowledge of the pharmacological and functional properties of the targets. In 2000, Drews reported that all the marketed drugs are targeting only 482 different targets. The estimation of the target numbers (by the same author) was clearly higher 3000–10000 (Drews 1996, 2000). On the basis of the estimation of DrugBank database the individual potential drug targets known is currently 4081 (Knox *et al.* 2011).

Today's drug development is a multi-step process in which each step requires accurate documentation. Therefore, the current drug development is time-consuming and expensive. Egner and Hillig (2008) have estimated that the average time for drug development is more than 12 years and costs around 1150 billion U.S. dollar. Unfortunately, data from 2003 showed that 60 % of drug discovery projects led to failure even before lead optimization, because underlying targets were found to be undruggable (Brown and Superti-Furga 2003).

Based on the above information, a crucial factor in a successful drug development process is the ability to identify the targets that are druggable. Several methods have been developed to evaluate the druggability of a target e.g. MAP_{pod} (Cheng *et al.* 2007), SiteScore (Halgren 2007), Druggability Score (Halgren 2009), DLID (Sheridan *et al.* 2010), RF-Score (Ballester and Mitchell 2010), DrugPred (Krasowski *et al.* 2011), and DoGSiteScore (Volkamer *et al.* 2012). Nevertheless, these methods are not capable to evaluating the druggability of a target without the knowledge of the binding site location, with exception of DoGSiteScore, which includes a binding site prediction algorithm. A more accurate optimization of a drug candidate is possible when the physicochemical environment of the binding site is known. The development of a drug without knowledge of the binding site could be compared to the building of a skyscraper construction onto unknown soil. It would be possible to achieve of success, but the probability of failure is high and the amount of wasted money and time is significant. Consequently, the identification of the binding site of a target protein is one of the most important parts of the drug development process.

Over 70 different small molecules binding site identification methods are presented in this thesis. In addition, the binding site identification ability of the methods are also compared with each other. The main focus is on the prediction of the binding sites of small molecules within proteins, but some methods are capable of identifying the protein-protein and protein-DNA binding sites. 10 datasets, which were developed to facilitate the evaluation of binding site identification ability, are also presented.

2 Binding site identification

The knowledge of three-dimensional structures of protein targets has the potential to greatly accelerate drug discovery and save resources. In some cases, it is possible to find the drug binding sites of a target protein on the basis of molecules that are known to bind to the protein (e.g. the well-known inhibitor of the target protein). Generally, X-ray crystallography or NMR is used to define the three-dimensional structure of the protein and/or ligand (Muchmore and Hadjuk 2003, Blundel and Patel 2004). The easiest and most certain way to identify the binding site of a drug is to succeed in defining the three-dimensional structure of a complex in which the ligand is bound to the protein. Neverthless, binding site prediction methods are still needed, as experimental identification of a binding site is not always possible.

To date, various binding site prediction methods have been developed. In 1985, Goodford introduced the first computational binding site characterization method, named GRID, that was based on theoretical interaction energies between probes and a protein. GRID was used e.g. for designing the inhibitors of influenza virus sialidase that resulted in a drug named Relenza® from GlaxoSmithKline (von Itzstein *et al.* 1993, Varghese 1999). Over the years, many different approaches were developed and used to identify binding sites. Some methods need information about the possible ligand of the target, but most of the methods only need the 3D-structure of the target protein. In addition, some methods are capable of predicting the binding site based on the amino acid sequence of a protein.

Table 1 The binding site prediction methods

Classification of binding site prediction methods (name, year published)					
Geometry-based	CavitySearch	1990	**Energy-based**	GRID	1985
	Method by Delaney	1992		Surflex-Protomol	1997
	POCKET	1992		vdW-FFT	1998
	Method by Del Carpio *et al.*	1993		Method by Elcock	2001
	VOIDOO	1994		CS-Map	2003
	SURFNET	1995		DrugSite	2004
	APROPOS	1996		Q-SiteFinder	2005
	LIGSITE	1997		PocketFinder	2005
	CAST	1998		DPA and Fast DPA	2006/2008
	DOCK(sphgen)	1982/1998		BindingResponse	2008
	Method by Voorintholt *et al.*	1998		AutoLigand	2008
	Surface patches	2000		Method by Morita *et al.*	2008
	PASS	2000		SiteHound	2009
	LigandFit	2003		Dsite	2010
	CASTp	2003/2006	**Evolutionary-based**	Sequence Space	1995
	SCREEN	2006		Evolutionary Trace method	1996/2002
	TravelDepth	2006		Method by de Rinaldis *et al.*	1998
	LIGSITE^{CS}	2006		Method by Landgraf *et al.*	1999
	PHECOM	2007		Method by Dean and Golding	2000
	Method by Xie and Bourne	2007		Method by Aloy *et al.*	2001
	PocketPicker	2007		ConSurf & ConSurf 3.0	2001/2005
	PocketDepth	2008		Rate4Site	2002
	VisGrid	2008		Method by del Sol Mesa *et al.*	2003
	VICE	2010		Method by Mayrose *et al.*	2004
	CLIPPERS	2010		Method by Innis *et al.*	2004
	GHECOM	2010		PatchFinder	2005/2008
	POCASA (Roll)	2010		HotPatch	2007
	DoGSite	2010/2012		siteFINDER|3D	2007
	LSite	2010		ConSurf 2010	2010
	MSPocket	2011	**Combined approaches**	SURFNET-ConSurf	2006
Blind Docking	Blind Docking	2002		LIGSITE^{CSC}	2006
	Optimized MD simulations	2004		FINDSITE	2008
	Method by Aita *et al.*	2010		FINDSITE^{LHM}	2009
	MolSite	2011		Focused Docking	2009
	Long scale MD	2011		SiteMap	2009
				MetaPocket	2009
				SiteIdentify	2009
				Fpocket	2009
				ConCavity	2009
				MetaPocket 2.0	2011
				DEPTH	2011
				Method by Gu *et al.*	2012

The prediction methods have been classified in many different ways, and in this thesis the methods are divided into five categories (Table 1). Four of the categories (geometry-, energy-, evolutionary-based, and combined approaches) are based on the classification of Volkamer *et al*. (2010). The geometry-based methods have been classified into three sub-types (grid system scanning, a probe sphere filling and alpha shape) based on Kawabata (2010) and Yu *et al.* (2010). The fifth category is added for Blind docking and Molecular Docking Methods.

There have been a great number of different binding site identification methods published and 76 of these methods are covered in this thesis (Table 1). The following methods are excluded from the thesis: the method Yao *et al.* (2003), Crescendo (Chelliah *et al.* 2004), the method by Keil *et al.* (2004), the method by Yang *et al.* (2005b), MEDock (Chang *et al.* 2005), PDBSiteScan (Ivanisenko *et al.* 2005), ET viewer (Morgan *et al.* 2006), the method by Rossi *et al.* (2006), FOD (Bryliński *et al.* 2007a, 2007b), THEMATICS (Wei *et al.* 2007), the PLB index (Soga *et al.* 2007a, 2007b), the method by Kim *et al.* (2008), SplitPocket (Tseng *et al.* 2009), method by Qui and Wang (2009), fPOP (Tseng *et al.* 2010), McVol (Till and Ullmann 2010), the method by Dai *et al.* (2011), IBIS (Thangudu *et al.* 2012), the method by Hawkins *et al.* (2012), Provar (Ashford *et al.* 2012), and PLB-SAVE (Lo *et al.* 2013). The exclusion was based on time limitations, and all of the excluded methods have seen little or not at all used after initial publication.

2.1 Geometry-based binding site identification methods

Geometry-based binding site prediction methods try to locate surface cavities and clefts by the analysis of the geometry of the molecular surface (Volkamer *et al.* 2010). These methods have been popular for years and they are commonly used in binding site identification. The main advantage of the geometry-based methods is their computational speed. Generally, geometry-based methods are based on the assumption that the binding site is usually the largest pocket (Kuntz *et al.* 1982, DesJarlais *et al.* 1988, Laskowski 1995, Laskowski *et al.* 1996, Peters *et al.* 1996, Hendlich et al. 1997) when only the size of the pocket matters.

Therefore, the algorithms of the methods commonly used are very simple and the methods do not require a lot of computational capacity. Additionally, the prediction of the binding site normally requires only the 3D crystal structure of the protein.

Geometry-based methods identify the binding pockets either by a grid or they are sphere- or tessellation-based. In tessellation-techniques 3D-structures are created by coating the surface of 3D object with the repetition of 2D planes without overlaps or gaps. Common to all geometry-based methods is simplicity, which is their strength but also their weakness as the binding site, in reality, is not always the largest pocket. For example, Laskowski and coworkers (1996) used SURFNET (Laskowski 1995) to analyze 67 protein structures and in 83 % of the cases the ligand binding site was found to be located in the largest pocket. Because of this problem, the methods have been improved further and new types of detection algorithms have been developed. Geometry-based methods have been the most used methods for detecting the binding sites since the Cavity search method was introduced by Ho and Marshall in 1990. To date, more than 20 methods have been developed (Table 1), and they are discussed more thoroughly below.

2.1.1 Grid system scanning

Grid-based methods use a 3D grid to define the molecular surface. It should be noted that these methods are not based on the method named GRID by developing Goodford, which uses the calculation of interaction energy instead of the geometric method to the characterization of binding sites. At first algorithms create a grid surrounding the protein. Then the algorithms simply find the interface of protein by determining which grid points are no longer parts of the protein. In addition, some Grid-methods use the molecular surface algorithms, so their results are not only dependent on the grid resolution. On the other hand, the methods are dependent on the radius of the sample probe, which defines the surface of the crystal structure by rotating around it. The probe radius is usually based on the water molecule that is 1,4 Å. Some of the methods create the protein surface by replacing protein atoms with 3D spheres of different van der Waals radii (e.g. Kleywegt and Jones 1994, Laskowski 1995, Venkatachalam *et al.* 2003, Yu *et al.* 2010). In the following sections grid-based methods are introduced in more details.

2.1.1.1 Cavity search

Cavity search was the first a geometric-based method that was created to find a binding site. At first, the algorithm of Cavity search isolates the cavity of interest. Then a 3D cast of the internal volume is produced using the techniques of 'solid modeling' (Requicha and Voelcker 1982). This method is not exactly based on a three-dimensional grid, as the algorithm divides the pocket into thin slices and forms two-dimensional cross-sections of these grid cells, which are filled by the 'flood filling' algorithm (Foley and Vam Dam 1982). These slices are assembled together in a three-dimensional pocket that will provide the volume and the shape of the pocket (Requicha and Voelcker 1982).

2.1.1.2 Method by Delaney

Delaney (1992) was the first who introduced the 3D grid-based pocket identification method. This method uses cellular logic operations to distinguish convex and concave regions of a protein structure, which has been mapped onto a 3D logical grid. Simultaneously it uses ‘solid fills’ technique for filling the protein cavities and it automatically defines a boundary between cavity and exterior free space.

2.1.1.3 Method by Del Carpio et al.

Del Carpio and co-workers (1993) developed an algorithm, which uses a surface ‘growing’ process to identify cavities and pockets. At first, the algorithm uses the method of Lee and Richards (1971) to identify the surface of protein (Del Carpio *et al.* 1993). Then it defines the center of gravity (C.O.G.) of the protein. Finally, the algorithm searches the closest surface atom from C.O.G., and defines and flags the atoms that are part of the concave pocket within the line of sight of the first atom (Fig. 1A). Then the algorithm searches the next closest unflagged atom to C.O.G. and repeats the concavity definition process. The algorithm continues until no more concave surface can be found (Fig. 1B). However, any scoring or prediction algorithms are not included in this method.

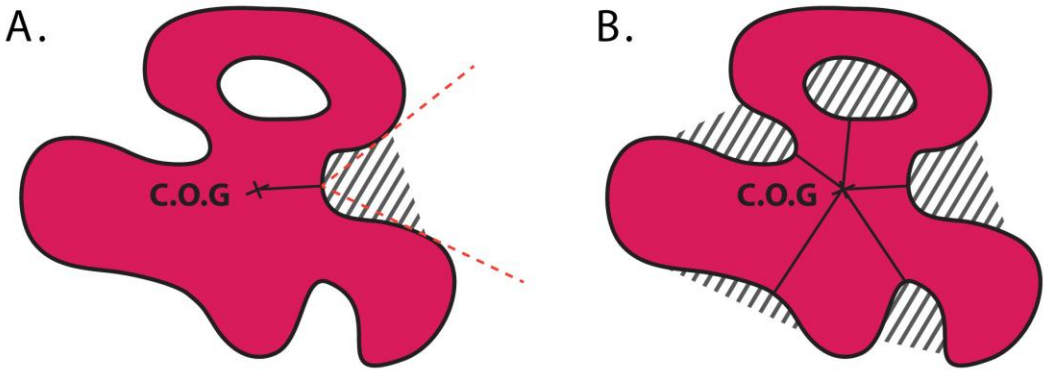

Figure 1 A. The method searches the closest surface atom from C.O.G. Then it defines and flags the atoms that are part of the concave pocket within the line of sight of the first atom (red dashed lines) and repeats this stage until unflagged atoms that are the line of sight cannot be found any more. **B.** The definition process continues until no more concave surface can be found. (Del Carpio *et al.* 1993, Laurie and Jackson 2006)

2.1.1.4 VOIDOO

In 1994, Kleywegt and Jones presented VOIDOO program that uses a mechanism called 'atomic fattening' for detecting the binding cavities and voids. The algorithm maps the protein onto a 3D grid with a spacing of 0,5-1 Å, before it starts to multiply the vdW radius of all atoms by a certain factor, which is typically 1,05 to 1,2. The multiplying is iterated until a particular cavity is found or a fixed number of iterations have been done (Fig. 2B & 2C). VOIDOO is able to find the cavities, which are in contact with the "outside world". In addition VOIDOO is able to separate the cavities, which are connected to one another through small channels. VOIDOO do not include scoring or prediction algorithms as part of the program, so it will not choose or propose any pockets automatically. VOIDOO has been updated 2008 (Kleywegt 2008).

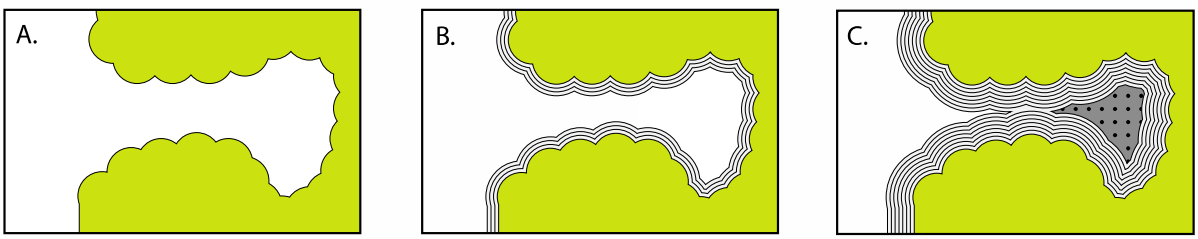

Figure 2 A. The target protein is mapped onto 3D grid. **B.** VOIDOO multiplies the vdW radius of all protein atoms until a fixed number of iterations have been done or **C.** a particular cavity is found. (Kleywegt and Jones 1994)

2.1.1.5 Pocket and LIGSITE

As Delaney method, Pocket (Levitt and Banaszak 1992) method was published at the Journal of Molecular Graphics, but one issue later. Pocket algorithm fills the protein with a grid with 3 Å spacing (Fig. 3A). Then it marks the grid points as part of the 'protein' or the 'solvent'. The algorithm goes along x, y, and z-axes of the grid and marks solvent points that lie between protein points (PSP = Protein-Solvent-Protein) for the potential pocket. Finally, the algorithm finds the largest cluster of 'pocket' points. The algorithm of Pocket is really fast, but it contains one main problem (Hendlich *et al.* 1997): the algorithm does not recognize correctly, or not at all, those pockets with an orientation of 45° to any of the x, y, or z-axes. Therefore, the orientation of the grid has a huge role in the identification of the pockets. Hendlich and co-workers (1997) approached this problem by extending the Pocket algorithm. In addition to the x, y, or z-axes, their LIGSITE program scans along the four cubic diagonals, which reduce this dependence on orientation (Fig. 3B). They tested also scanning along the diagonals in the xy, xz, and yz planes but showed no further improvement. LIGSITE give each solvent accessible grid points value between 0 (not in a cavity or pocket) and 7 (deeply buried), based on how many scanning direction it gets in PSP event. In LIGSITE, the grid spacing can be freely adjusted, but below 2 Å the grid spacing results will not be significantly better, and in addition the calculation time increases exponentially (Hendlich *et al.* 1997).

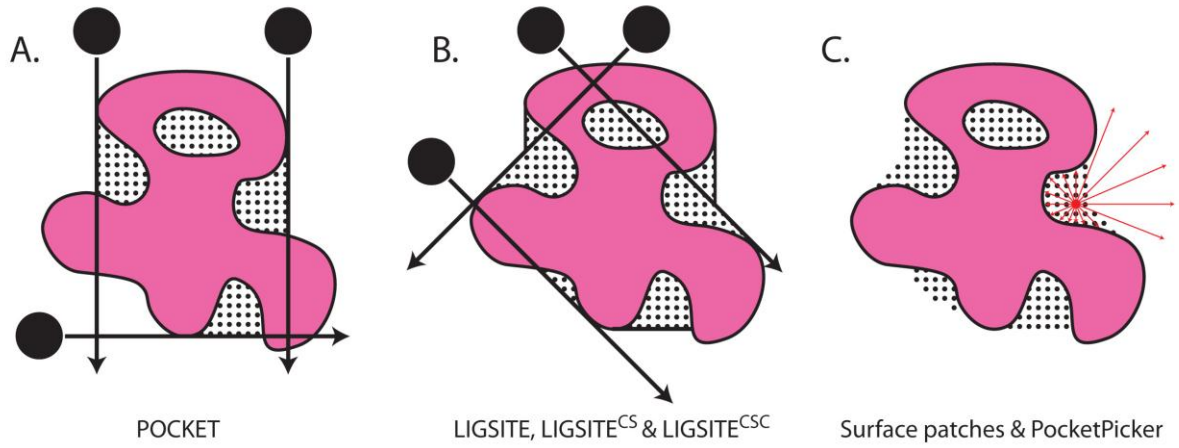

Figure 3 A. Pocket algorithm scans along x, y and z axels. **B.** In addition, LIGSITE, LIGSITE^{CS} and LIGSITE^{CSC} scan along four cubic diagonals. **C.** Surface patches algorithm scans the surroundings in 14 directions and PocketPicker scans in 30 directions from the grid points. (Laurie and Jackson 2006)

2.1.1.6 Method by Voorintholt et al.

Voorintholt and co-workers (1989) introduced very simple binding site prediction method. First, user gives a value of grid spacing between 0,3-1,0 Å and the radius of the sample probe, which describes a potential ligands size. Then, the algorithm creates a 3D grid around the protein and every grid point is scored by the distance descriptor. If a grid point falls within the vdW radius of any protein atoms, it scores a value of 100 and when a grid point is further than the vdW radius, but closer than the vdW radius plus a user selected sample probe radius, it will be get a value between 0-100. Finally, the program shows contours around the cavities, which are large enough to hold the sample probe.

2.1.1.7 DOCK (spghen)

Hendrix and Kuntz (1998) developed a new site descriptor for their molecular modeling program that called DOCK (Kuntz *et al.* 1982). This descriptor helps to find possible binding sites. First, the surface of protein is described with Connolly's molecular surface program (Connolly 1983). Then a solid angle of each surface point is calculated using Connolly's solid angle algorithm (Connolly 1986). Finally, cavities and concavity areas are determined by the solid angles (Fig. 4).

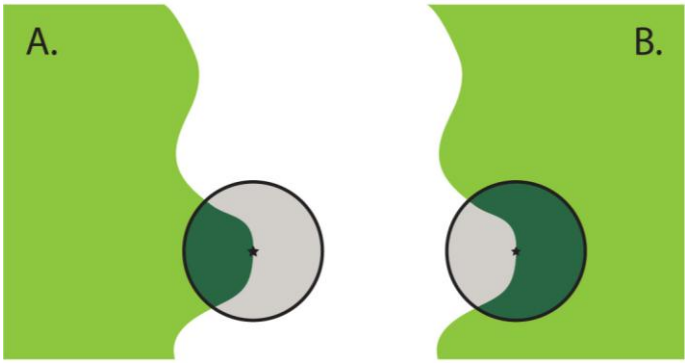

Figure 4 The green area represents the interior of the protein. The interior of the protein, which lies inside test sphere (dark green), defines the surface solid angle of surface point (black star). **A.** In this rounded surface approximately ¼ of the test sphere overlapping with protein, therefore solid angle is ¼ × 4π or π. **B.** Solid angel is approximately ¾ × 4π or 3π in this concavity area. (Hendrix and Kuntz 1998)

2.1.1.8 Surface patches

In 2000, Roche's researchers introduced their fully automated method called Surface patches (Stahl *et al.* 2000). In addition to the identification of protein, Surface patches

generates the topological correlation vectors of the *solvent accessible surface* (SAS) and visualizes these vectors onto a planar display by means of *self-organizing maps* (SOM). SAS calculation algorithm has been described elsewhere (Stahl *et al.* 1999) and it based on Connolly algorithm (Conolly 1983). The pocket prediction method of Surface patches is very similar to LIGSITE. The method uses a grid with 1 Å spacing and grid points within 0,8 Å from the vdW surface of the protein atom are marked as 'protein'. The remaining grid points are marked as the 'solvent'. Unlike the LIGSITE method, Surface patches does not scan the grid points along the axes and diagonals of the whole grid and doesn't use PSP events to define the pockets. It scans the surroundings in 14 directions from each 'solvent' point (Fig. 3C), which are no farther than 2 Å from the 'protein' points, and determination of the pocket points is based on the specific terms (Stahl *et al.* 2000).

2.1.1.9 LigandFit

Venkatachalam and co-workers (2003) developed a docking program named LigandFit that also includes a binding site identification algorithm. This algorithm use 0,5 Å spacing in a grid and classifies every grid point either 'occupied' or 'free'. Those grid points, which lie closer than the radius of the protein atom (2,5 Å at heavy atom and 2 Å at hydrogen), will be marked as 'occupied' and every other grid point is 'free'. Next, the algorithm removes the 'free' grid points, which are not in the site region. For this operation algorithm employ a cubically shaped 'eraser' that moves along the axes of the grid system normal to the six faces of the rectangular parallel piped (Fig. 5). Eraser stops whenever it comes into contact with a protein atom and then it moves sweeping to the next grid line. Finally, all the 'free' grid points that belong to the same site are collected into a single group by using a flood-fill procedure (Foley and Vam Dam 1982, Rogers 1985, Venkatachalam *et al.* 2003). The success of this algorithm is strongly dependent on the 'eraser' size. Much smaller 'eraser' than the mouth of the pocket can be reached inside the pocket, when the grid points that describe the pocket are classified incorrectly to 'free'. Accordingly 'eraser' size should be in proportion to the dimension of the mouth of the pocket.

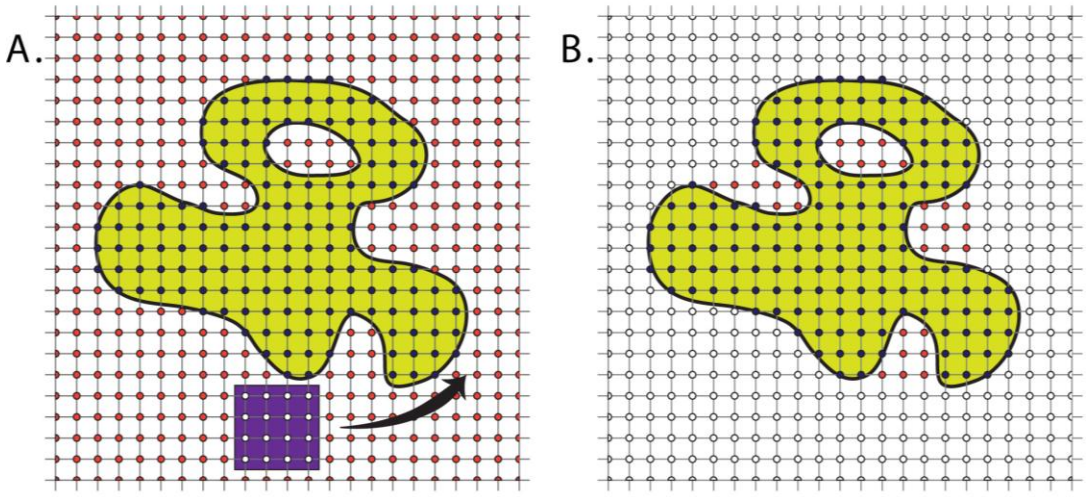

Figure 5 A. Red grid points are 'free' and dark blue points are 'occupied'. The cubical shaped 'eraser' (a purple square) moves along the axes. **B.** All of the remaining 'free' grid points that belong to the same site are collected into single group. (Venkatachalam *et al.* 2003, Laurie and Jackson 2006)

2.1.1.10 LIGSITE^{CS}

Huang and Schroeder (2006) improved LIGSITE program with two extensions. First, instead of capturing PSP events by using protein's atoms, they used the protein's Connolly surface (Connolly 1983) to capture the more accurate surface-solvent-surface events (Huang and Schroeder 2006). This extension they called LIGSITE^{CS} (CS = Connolly surface). Second extensions they called LIGSITE^{CSC} (Connolly surface and conservation) and it re-ranks identified pockets by the degree of conservation of the involved surface amino-acid residues. As this method is not purely a geometric-based method, it is discussed in more detail in section 2.5.2.

2.1.1.11 PocketPicker

The basic idea of PocketPicker is very similar to a number of the above-presented programs (e.g. LIGSITE). Identification of the potential binding site will be made by the calculations of both the buriedness and the shape (Weisel *et al.* 2007). PocketPicker uses a 3D grid with 1 Å spacing around the protein and pocket detection algorithm focuses on grid points that are located closely above the protein surface (maximal distance is 4,5 Å to the nearest protein

atom). A probe is placed in each grid point, which is located inside this region and the probe scans the molecular surrounding in 30 different directions (Fig. 3C). As the following step PocketPicker identifies the potential binding site by the calculations of both the buriedness and the shape. PocketPicker is benchmarked to SURFNET, PASS, CAST, LIGSITE, LIGSITE^{CS}, and LIGSITE^{CSC} with 48 unbound/bound protein structures data set developed by Huang and Schroeder (2006). Benchmarking proved that PocketPicker have better prediction success accuracy than other method except LIGSITE^{CSC}, which is slightly better.

2.1.1.12 PocketDepth

PocketDepth method uses a depth based clustering to a binding site prediction (Kalidas and Chandra 2008). PocketDepth's algorithm can be separated to 6 major steps: 3D grid construction, grid cell labeling, drawing grid bars, depth factor computing, clustering, and ranking, which is shown in the flow-chart. Two different parameters may be used for the ranking of the binding site, 'depth' and 'surface', separately or in an automatic combination. 'Depth' parameter provides better predictability than the 'surface' or a combination thereof; even so the results are not very good.

2.1.1.13 VisGrid

Li and co-workers (2008) had a unique and interesting approach to identify the binding site. Their VisGrid algorithm uses the visibility criterion for characterizing the local geometric features of protein surfaces (Li *et al.* 2008). In addition to identifying the pockets, VisGrid is able to identify hollows (cave or cavity), large protrusions, and flat regions of the protein structure. VisGrid is also able to name atoms and residues, which are part of the identified region. To begin the binding site identification process VisGrid places protein structure onto 0,9 Å spacing 3D grid and each cell has a value 'empty', 'filled' or 'surface'. The cells are marked as 'Filled' if they are within a sphere, which is centered on a protein atom with the radius of the vdW radius of atom plus the radius of the water molecule (1,4 Å). All other cells are marked as 'empty'. Those 'filled' cells that are adjacent to at least one 'empty' cell are

marked as ‘surface’. After this, the visibility of every ‘surface’ cell was calculated and cells, which belong the same visibility range, are collected into groups. The algorithm makes the characterization of residues by the visibility threshold (Fig. 6). Li and coworkers have also tested the robustness of the algorithm by predicting binding sites in distorted protein structures by Molecular Dynamics simulation. The success of the method is illustrated by the fact that the number of false positive predictions is not increased on distorted structures.

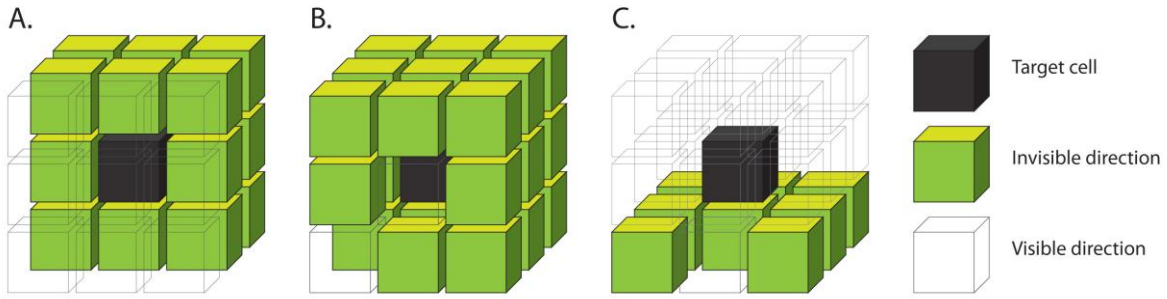

Figure 6 The visibility of each ‘surface’ cell is defined as the number of visible directions from the cell. A direction is considered to be visible when the line of sight from the target cell toward that direction does not hit any ‘surface’ or ‘filled’ cell up to 20 layers. The visible directions from the target cell are 9 (**A**), 2 (**B**), and 18 (**C**) in the first layer. (Li *et al.* 2008)

2.1.1.14 VICE

Tripathi and Kellogg (2010) began to develop VICE (Vectorial Identification of Cavity Extents) method because in their opinion most of the previously developed programs contained several weaknesses. Authors aim was to develop algorithm, which is more flexible to identify a wide variety (shapes and sizes) of the cavities. In addition, the method should be adjustable and it should be able to characterize the unusual cavities from the multi-domain proteins with channels or tunnels. At the end quantitative prediction was deemed important (Tripathi and Kellogg 2010). Algorithm VICE was constructed around the HINT toolkit (Kellogg *et al.* 2005) and in addition, it had several new subroutines for 3D map manipulation and analysis (Tripathi and Kellogg 2010). An approach of binding site detection is very similar to LIGSITE, PocketPicker and Surface Patches. VICE provides several user adjustable options for optimizing the calculation. This gives the possibility to change the focus from the whole protein to a small specific region for a detailed inspection.

VICE surround protein with 3D grid (spacing 1 Å) and those grid points which are within vdW radii of protein atoms, are immediately discarded. Algorithm scans the surrounding area in the remaining grid points and each of the points scored by the degree of buriedness. Finally, the program can delete irrelevant pockets automatically or, if desired, it displays the intermediate raw maps and allows interactive application of the refinement. VICE's ability for binding site prediction is better than most of the previous methods and it can locates and delineates most of the different kinds of pockets. Most importantly, VICE can identify successfully over 80 % of the binding pocket in the structures on unbound proteins (Tripathi and Kellogg 2010).

2.1.1.15 DoGSiteScore, LSite and DSite

Volkamer and co-workers (2012) DoGSiteScore is a fully automated geometry-based prediction method. It is able to detect pockets and sub-pockets within protein structures and also to predict the druggability of predicted pockets. The detection of binding pockets uses DoGSite method (Volkamer *et al.* 2010), which is based on "Difference of Gaussian" (DoG) filter (Marr and Hildreth 1980) on a 3D grid to detecting possible binding pockets (Volkamer *et al.* 2012). The main difference between DoGSite and DoGSiteScore is that the latter is also able to estimate the druggability of detected pockets. DoGSite was compared with VICE (Tripathi and Kellogg 2010), Fpocket (Le Guilloux *et al.* 2009), PocketPicker (Weisel *et al.* 2009), LIGSITE (Hendlich *et al.* 1997), CAST (Liang *et al.* 1998b), PASS (Brady and Stouten 2000), SURFNET (Laskowski 1995), DSite, and LSite (Volkamer *et al.* 2010) on a dataset of 48 unbound/bound structures. LSite and DSite are Volkamer's and co-worker's versions of LIGSITE and DrugSite (An *et al.* 2004) algorithms. In contrast to the original LIGSITE, LSite do not use a fixed buriedness cut-off. DSite is the modified version of DrugSite algorithm, which is actually the energy-based method. In contrast to the original DrugSite, DSite uses a Gaussian filter in place of moving average filter. Among the studied methods VICE was clearly the best method to identify the binding site with unbound structures, when only first predictions are taken into account. With bound structures and one of top three predictions as acceptance criteria DoGSite is as good as VICE and both are better than other compared methods.

2.1.2 Probe sphere filling

Probe sphere filling methods use spherical probes to identify the binding sites. These methods have a number of different approaches, such as gap-sphere (Laskowski 1995), rotating the probe (Yu *et al.* 2010, Zhu and Pisabarro 2011), multiscale probe (Kawabata 2010), the combination of big and small probes (Kawabata and Go 2007), probes placing tangential to the triplets of protein atoms (Brady and Stouten 2000), but all of these have the same goal. All of these methods are based on the assumption that the most potential binding site is usually the largest pocket (Kuntz *et al.* 1982, DesJarlais *et al.* 1988, Laskowski 1995, Laskowski *et al.* 1996, Peters *et al.* 1996, Hendlich *et al.* 1997), so the binding site identifications are based on the size or volume of pockets.

2.1.2.1 SURFNET

Roman Laskowski presented in 1995 first probe sphere filling based method, which called SURFNET. The basic idea of SURFNET is to create a three-dimensional structure of the binding site by filling these areas with so-called 'gap-spheres' (Laskowski 1995). At first, algorithm replaces protein atoms with spheres with vdW radii. At the next stage it places 'gap-spheres' in midway between a pair of atoms in a protein molecule (Fig. 7A). If there is overlapping by any other neighboring atoms, the radius of 'gap-sphere' will be reduced until no overlap occurs. Then the remaining 'gap-spheres' define pockets and cavities, and each 'gap-spheres' will be converted to 3D shapes, which together form the pocket area (Fig. 7B).

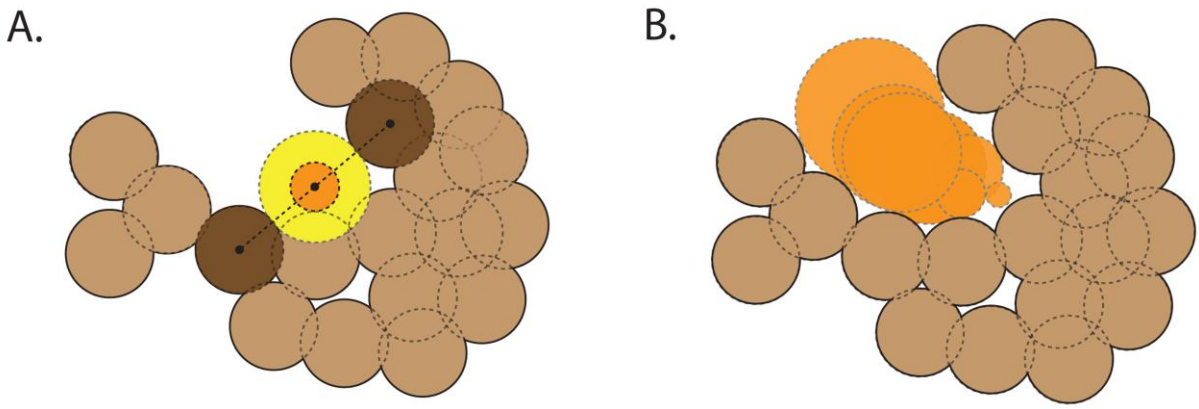

Figure 7 A. SURFNET algorithm places 'gap-sphere' (yellow sphere) in midway between a pair of atoms (darker brown spheres). The radius of 'gap-'sphere' will be reduced until no overlap with neighboring atoms occurs (orange sphere). **B.** The remaining 'gap-spheres' define pockets. (Laskowski 1995, Laurie and Jackson 2006)

2.1.2.2 PASS

The PASS method coats the protein surface by creating a probe spheres tangentially to all unique triplets of atoms (Fig. 8A & 8B), layer upon layer, until the new probe spheres can no longer be created (Brady and Stouten 2000). Then only the probes with low solvent exposure are kept. Then, PASS determines the potential binding sites using ASP (Active Site Point) method (Fig. 8C).

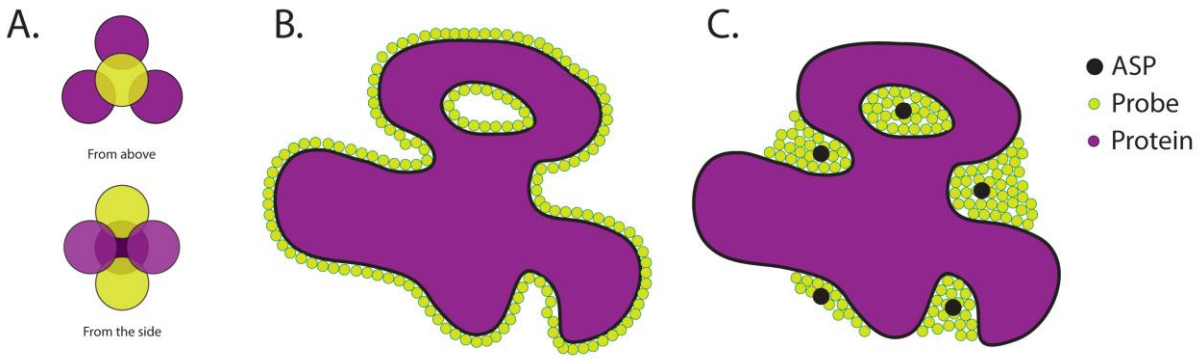

Figure 8 A) Tangentially placed probe sphere. **B**) First, PASS coats protein surface by probe. **C)** After multiple layers creating and filtering PASS identifies ASP, which determines the potential binding sites. (Brady and Stouten 2000, Laurie and Jackson 2006)

2.1.2.3 SCREEN

Nayal and Honig (2006) developed SCREEN (Surface Cavity REcognition and EvaluatioN) called method, which defines surface cavities geometrically in terms of the empty space between the protein's molecular surface and an envelope surface. SCREEN constructs these surfaces by using GRASP program (Nicholls *et al.* 1991), which creates the surface by rolling spherical probes (the protein's molecular surface using a 1,4 Å probe radius and the envelope surface using a large probe radius akin to that of a typical small ligand dimension) (Nayal and Honig 2006). The prediction of the most potential binding pocket is made by using the Random Forest (Breiman 1996, 2001) machine learning technique (Nayal and Honig 2006). Nayal and Honig compared SCREEN and CASTp with the dataset by Perola *et al.* (2004) and the results proved SCREEN more successful in identifying the binding site.

2.1.2.4 PHECOM

In 2007, Kawabata and Go presented a new method, which is called PHECOM (Probe-based HECOMi finder). The basic idea of their method is to determine the binding site using two adjustable parameters, which describe the 'depth' and 'size' of pockets (Kawabata and Go 2007). To describe these parameters, method simply uses two different sized probe spheres, of which the smaller one can enter inside the potential pocket, but the larger probe cannot. Firstly, PHECOM creates randomly at probe spheres on the protein surface with 1,87 Å radius (the size of a single methyl group) (Fig. 9A). Like in PASS method (Brady and Stouten 2000), also in PHECOM each probe sphere must be tangentially with tree protein atoms (Kawabata and Go 2007). Then method creates a new set of probe spheres with a user-selected radius (4-12 Å) (Fig. 9B). Those large probe spheres and the parts of small probes, which are within the large probe spheres, are removed. The remaining parts define possible binding pockets and cavities (Fig. 9C).

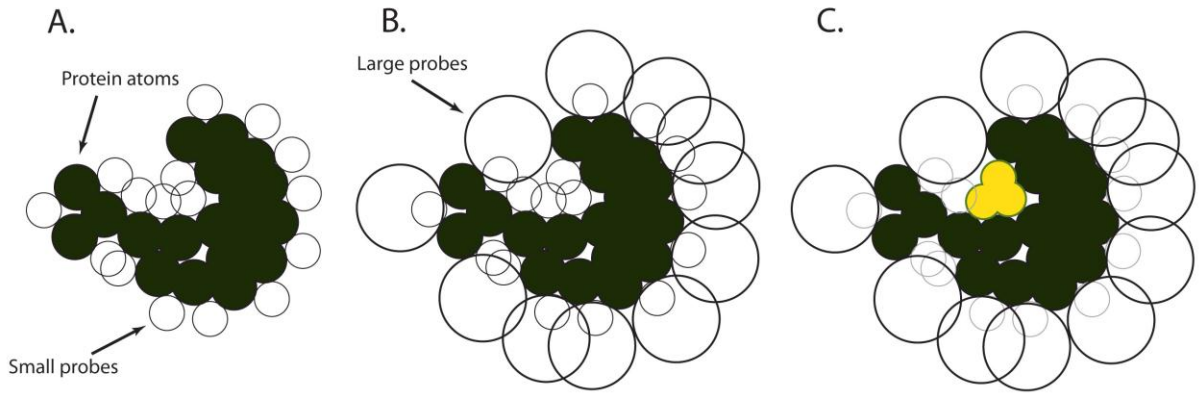

Figure 9 A. PHECOM creates randomly at the tangentially placed small probe spheres on the protein surface. **B.** The method creates the new set of probe spheres with larger radius. **C.** The large probe spheres and the parts of small probes, which are within the large probe spheres, are removed and the remaining parts define possible pockets. (Kawabata and Go 2007)

2.1.2.5 GHECOM

Although PHECOM gave better (and faster) results than the PASS and SURFNET, Kawabata was not completely satisfied with the method (Kawabata 2010). In his opinion the probe spheres, which are placed tangentially with three protein atoms, were main weakness of PHECOM. Therefore PHECOM was not capable to determine all the possible regions. Other weakness was computational calculation time, as it was approximately proportional to the square of the number of protein atoms. To remedy the deficiencies Kawabata developed the method further. The result was GHECOM (Grid-based HECOMi finder) algorithm, which uses the 3D grid presentation of protein and probes spheres. To detect binding pockets and cavities GHECOM uses multi-scale large probe, which are based on the theory of mathematical morphology (Masuya and Doi 1995). In practice, the algorithm uses many large probes with different radii in the same time at the same point and this allows GHECOM to provide new perspectives on the definition of binding sites. Kawabata proposed that these new concepts should be called ‘multi-scale molecular volume’ and ‘multi-scale pocket’.

Although GHECOM uses multi-scale probe, it is still significantly faster than the predecessor - PHECOM. GHECOM did very well in comparison with the Q-SiteFinder, PASS, SURFNET, and PHECOM, but the test set used in the study contains only bound state structures (Kawabata 2010). Kawabata had noted himself, that this does not reflect how well the algorithms can

identify pockets of the unbound structures. GHECOM can be used via web browsers in address: http://strcomp.protein.osaka-u.ac.jp/ghecom/

2.1.2.6 POCASA

POCASA (Pocket-Cavity Search Application) program is based on the ROLL algorithm that scans the surface of the protein by rolling (Yu *et al.* 2010). The method is actually a mix of the grid-based and probe-based methods, but the rotating probe is a more important part for prediction than the grid, therefore it is classified to the probe-based method. Initially, the program creates a structure of the protein in the 3D grid, such e.g. as SURFNET, by replacing each protein atom (except hydrogens) with vdW radii spheres. Then, for each point of the network are given a value, either 1 (protein) or 0 (free). After this, program cuts 3D grid to 2D slices (Fig. 10A) and the probe scans the grid points of each slice by rolling along the protein surface without any overlapping protein (Fig. 10B). Slicing and scanning are made in directions of x, y, and z axes, this provides a much better resolution of the protein surface. The probes rolling direction along the protein surface is controlled by the inner border tracing algorithm (Sonka *et al.* 1998). The free grid points that fall within the probe rolls are marked as the 'probe surface' (Fig. 10C) (Yu *et al.* 2010). Then the free grid points between 'probe surface' and protein surface or those, which are surrounded by the protein surface, are identified the pockets and the cavities. POCASA remove distracting noise points with the assistance of two parameters: 'Single Point Flag' (SPF) and 'Protein-Depth Flag' (PDF). Finally, the potential binding pockets are ranked by a special 'Volume Depth' (VD) parameter.

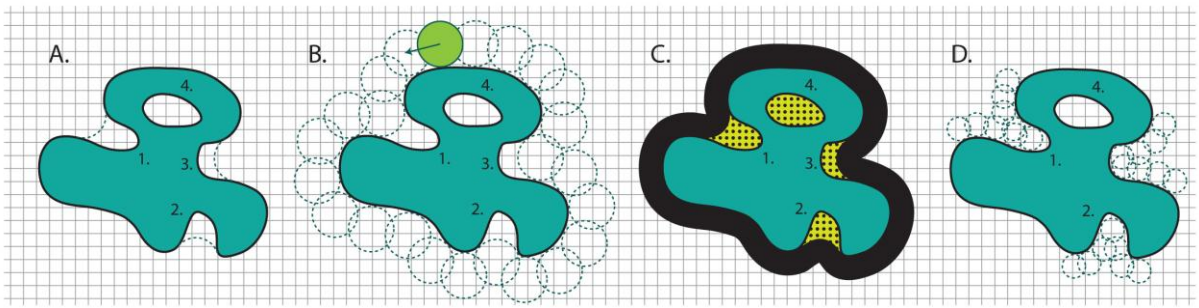

Figure 10 Schematic representation of ROLL. **A.** The 2D slice of protein. Regions 1-3 are defined as pockets and region 4 is defined as a cavity. **B.** The probe (green sphere) roll along the protein surface without overlapping protein. **C.** The black are is the 'probe surface'. Then the free grid points between 'probe surface' and protein surface (regions 1.-3.) or those, which are surrounded by the protein surface (region 4.), are identified the pockets or the cavities (yellow areas) **D.** The too small probe causes pocket 3. to disappear and pockets 1.–2. to become smaller, while it did not affect cavity 4. (Yu *et al.* 2007)

2.1.2.7 MSPocket

MSPocket (Molecular Surface Pocket) identifies a possible pocket on the solvent excluded surface (SES) of protein (Zhu and Pisabarro 2011). The method determines SES by using MSMS program (Sanner and Olson 1996). MSMS consist of four algorithms: 1. Reduced Surface computation, 2. Analytical solvent excluded surface computation, 3. Treatment of singularities, and 4. Triangulation of the SES (Sanner and Olson 1996). MSMS use both a rotating probe than the alpha-shape to the binding site identification therefore MSPocket is not purely probe-based or alpha shape method. However, the method is completely geometry-based and the probe is a more important role, so that is why this is classified as the probe-based method. The identified pockets may be ranked according by various measures: the number of pocket vertices, the number of pocket atoms associated with pocket vertices, the pocket surface area, or the pocket volume (Zhu and Pisabarro 2011). By default, MSPocket uses the estimation of pocket volume (V_e) for pockets ranking.

2.1.3 Alpha shape

In the 1990's, Edelsbrunner and colleagues developed the rapid computer-aided three-dimensional structure describing method, which is based on alpha-shape theory (Edelsbrunner and Mücke 1994, Facello 1995, Edelsbrunner *et al.* 1995, 1996, Edelsbrunner and Shah 1996). The alpha shape method constructs the three-dimensional alpha-shape of an object, such as the protein, by the Delaunay triangulation algorithm and the Voronoi diagrams (Edelsbrunner and Mücke 1994). Voronoi diagram (Fig. 11A) and Delaunay triangulation are mathematically equivalent for each other; therefore, they are particularly useful when combined (Liang *et al.* 1998b). Accuracy of the alpha-shape can be affected by the alpha value, which is roughly similar to the probes radius effect on the accuracy of shape with the probe-based method.

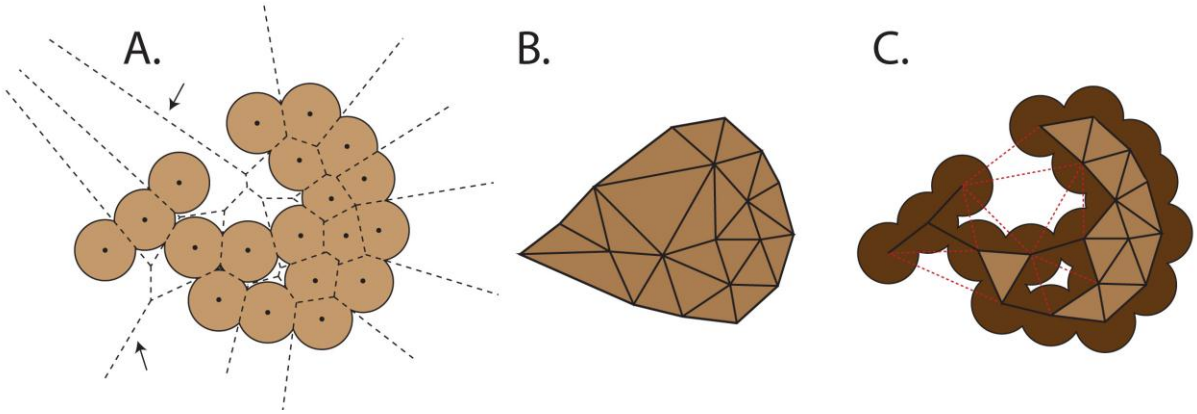

Figure 11 A. Two-dimensional (2D) model of Voronoi diagram. Arrows indicate 2 of 10 Voronoi edges, which are completely outside the molecule (red dashed line in Fig 10C) **B.** 2D model of Convex hull of the atom centers, and **C.** 2D model of Alpha shape of molecule. (Liang *et al.* 1998b)

In simple terms, at first the protein is triangulated by the Delaunay tetrahedrons. Then, the alpha form of target protein is obtained by removing the empty Delaunay tetrahedrons that have a part of the tetrahedron outside the protein. Finally, the pockets and the voids can be recognized by the alpha form and/or the Discrete-flow method.

Peters and co-workers were the first that took advantage of the alpha-shape to identify the protein binding sites (Peters *et al.* 1996). After this, the method has been used in a variety of binding site identification programs and a couple of improvements have been developed for this method such as the vdW radius of atoms weighted Delaunay triangulation (Edelsbrunner 1995, Liang *et al.* 1998a) and the discrete flow method (Edelsbrunner *et al.* 1996). The advantage of Alpha-shape based methods is their ability to distinguish holes and surface shapes of objects.

2.1.3.1 APROPOS

APROPOS (Automatic PROtein POcket Search) (Peters *et al.* 1996) was the first program, which used alpha shape theory (Edelsbrunner and Mücke 1994) for the prediction of the binding sites. At the same time, it was the first fully automatic geometric pocket detection program (Peters *et al.* 1996). APROPOS makes the prediction of binding sites by comparison between the enveloping surface area of protein (ESA), which is actually Convex hull (Fig. 10B) and detailed description of the surface area of protein (DSA). Different alpha values

have been used for the creation of these surfaces, whereupon the separation accuracy of the surface shapes is also different. Peters and coworkers discovered that pocket detection fails with proteins that contain less than 50 amino acids, and also if the ligand binds covalently.

2.1.3.2 CAST and CASTp

In 1998, the development team of Alfa-form theory built their own binding sites identification program named CAST (Computed Atlas of Surface Topography) (Liang *et al.* 1998b, Binkowski *et al.* 2003). CAST identifies the protein pockets and cavities, as well as the volumes and areas. It also detects the mouths of the pockets and identifies atoms on the edges of the pocket as well calculates the edge area and diameter of the mouth. Unlike APROPOS program, CAST uses only one alpha shape for the identification of binding sites and this alpha shape is formed by weighted Delaunay triangulation. For the better identification and measurement of pockets, CAST has been extended with the discrete flow method. The discrete flow is defined only for those tetrahedrons (formed by the Delaunay triangles) that are not part of alpha-shape. Obtuse angled empty tetrahedrons flow to the neighboring tetrahedrons, whereas acute-angled empty tetrahedrons are “sinks” that collect excess flow from neighboring empty tetrahedrons (Fig. 12). The sizes of the pockets mouths are determined by using the discrete flow method. CAST is better defining the deep and narrow pockets than APROPOS.

Upgraded version, CASTp (Computed Atlas of Surface Topography of proteins, released 2003), includes a graphical interface, flexible interactive visualization, as well as on-the-fly calculation for user-uploaded structures and web browser interface (Binkowski *et al.* 2003). The update of year 2006 incorporated functional information of annotated residues from PDB, Swiss-Prot and Online Mendelian Inheritance in Man (OMIM) (Dundas *et al.* 2006). CASTp is able to use on the web browsers in address: http://sts.bioengr.uic.edu/castp/ (Dundas *et al.* 2006).

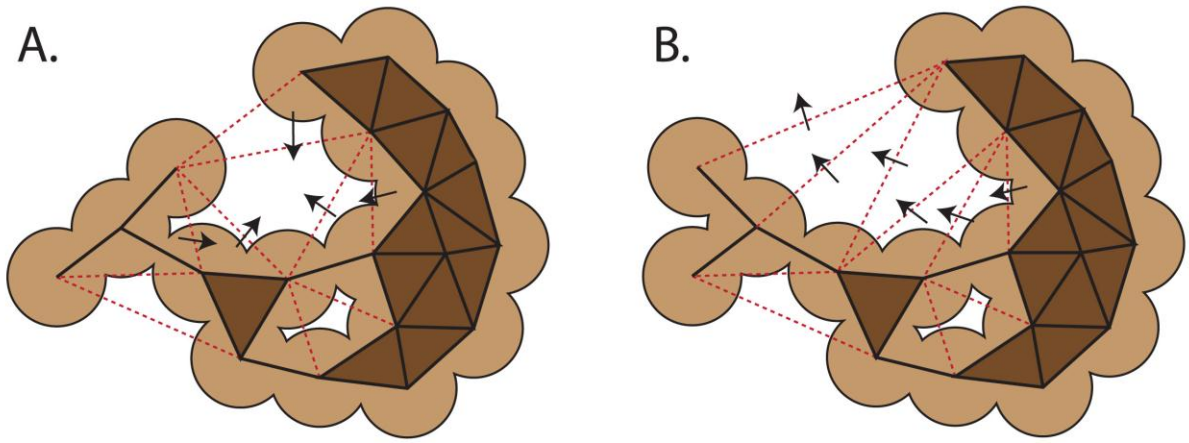

Figure 12 Simplified 2D models of discrete flows **A.** One empty Delaunay triangle acts as a sink, where neighboring empty triangles flow (to direction arrows). **B.** When cavity is too wide open Delaunay triangles flow to infinity. CAST not defines these type cavities as pocket. (Liang *et al.* 1998b)

2.1.3.3 Travel Depth

Coleman and Sharp (2006) developed an algorithm called Travel Depth that can measure the depth of the cavities. The starting point was to develop a method for the quantitative determination of the depth of cavities, which facilitates the analysis of protein structure (Coleman and Sharp 2006). Like APROPOS, Travel Depth determine the depths and volumes of the cavities between two surfaces created by different methods. The first surface is the protein surface, which is created by the alpha-shape method. The second surface is simplified surface of protein created around the protein by rolling a probe with infinite radius. The algorithm measures depth with the assistance of a 1 Å spacing grid. Each grid cell is classified into one of the four categories based on location (Molecular **S**urface, **I**nside molecular surface, **B**etween convex hull and molecular surface, and **O**utside convex hull and molecular surface). Every cell that has been classified to outside convex hull gets the value 0 and all other cells get the value of the shortest distance between the simplified surfaces and closest outside cells. This will show the depths of the surface. Travel depth is also able to determine the depths of tunnels.

2.1.3.4 Method by Xie and Bourne

Xie and Bourne (2007) developed a method, which only needs the location of Cα atoms to present the protein structure. In addition to the alpha-shape their method uses to geometrical potential to describe the shape of the structure. The geometric potential is analogous to the hydrophobicity or electrostatics potential so that it is dependent on both the whole surface shape of the protein as well as the surrounding residue. Finally, the method creates a virtual ligand, which allows to predict the binding site. Usage of Cα atoms in the protein structure determination significantly reduces the calculation time. This method is capable of identifying the binding sites also elsewhere than in the cavities. Geometric potential can distinguish the convex and flat surfaces, even though they have the same depth.

2.1.3.5 CLIPPERS

Travel Depth (Coleman and Sharp 2006) does not predict any binding site by itself and it is also unable to determine the cavities, which are completely within the protein. Therefore Coleman and Sharp (2010) created a new binding site identification method CLIPPERS, which is based on the Travel Depth algorithm. CLIPPPERS analysis the whole protein surface and then calculates the various features of each grid cell. This information is collected to facilitate the collating, filtering, comparing, and clustering of pockets. Each pocket is organized into a hierarchical tree of sub-pockets (Fig. 13). Finally, the binding site prediction

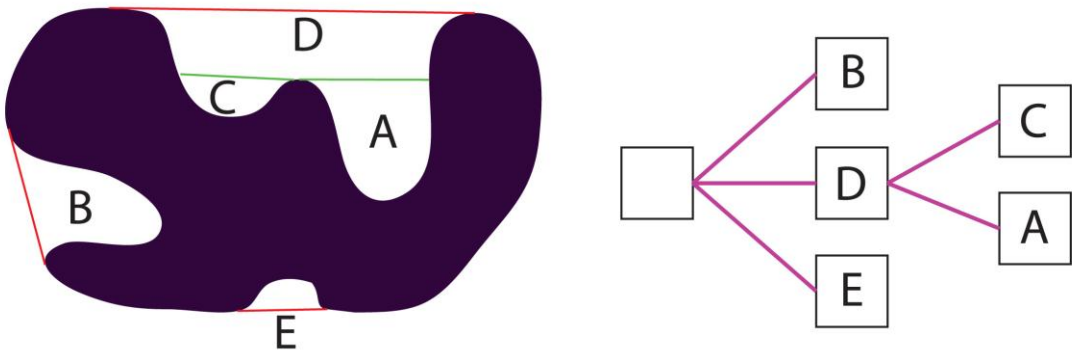

Figure 13 CLIPPERS organize pockets and sub-pockets into a hierarchical tree. Pocket D is divided into two sub-pockets, A and C. (Coleman and Sharp 2006)

is simply made by computing Taminoto-type overlap score and the pocket that maximizes the Tanimoto overlap score is picked. Cataloguing is done without any tunable parameters or user intervention.

2.2 Energy-based methods

Energy-based binding site prediction methods are based on the assumption that the potential binding sites can be identified on the basis of interaction energies. The first energy-based method, GRID (Goodford 1985), uses three different energy functions to calculate non-bonded interaction energies, but most of the energy-based methods are based on vdW interaction energies (Bliznyuk and Gready 1999, Kortvelyesi *et al.* 2003, An *et al.* 2004 & 2005, Laurie and Jackson 2005, Morita *et al.* 2008, Ghersi and Sanchez 2009a, Volkamer 2010). Usually, methods use a small probe and calculate interaction energies between the probe and a protein. Methyl probe is used in most cases, but carbon (An *et al.* 2004 & 2005, Volkamer *et al.* 2010) and other small probes (Ghersi and Sanchez 2009a) are also used and some methods are using several different probe types (Ruppert *et al.* 1997, Kortvelyesi *et al.* 2003).

2.2.1 GRID

GRID was the first method, which was able to identify a potential binding site (Goodford 1985). At the same time, it was also the first energy-based method. GRID calculates non-bonded interaction energies (E_{xyz}) between the probe molecule and the potential binding sites. Calculation of E_{xyz} is used in three different energy functions: the Lennard-Jones Function (E_{lj}), the Electrostatic function (E_{el}), and the Directional hydrogen bond function. The full parameterization of water is the weakness of this method. In 1985, the calculation of complete water interaction energies was too computationally expensive, and therefore GRID uses a very simplified model of water. Due to this, the structural waters can easily falsify results of the interaction energies. On the other hand, e.g. Minke and co-workers

(1999) have reported that GRID was able to predict, which waters are structural in order to propose a correct binding mode for m-nitrophenyl-alpha-galactoside.

2.2.2 Surflex-Protomol

In 1997, Ruppert and co-workers presented their binding site prediction method called Surflex-Protomol (Pérot *et al.* 2010), which consist of three main steps: probe placement, 'Sticky spot' identification, and pocket accretion (Ruppert *et al.* 1997). At first, method places three type probes densely around the protein: a steric (hydrophobic) probe, a hydrogen bond donor probe, and a hydrogen bond acceptor probe. The set of probes are screened and only those are preserved, which have the strongest interaction with the protein. Then 'Sticky spots' are identified by selecting those probe subsets, which are able to make strong cumulative interaction with protein. This interaction is defined from the density of local score, which are produced by Jain's (1996) scoring function. Because the algorithm does not take into account the structure of the protein, the method may produce 'Sticky spots', which are not connected correctly, and thus pockets will not be formed properly (Ruppert *et al.* 1997). To remedy this problem a special 'accretion' step is used to analyze the connectivity of the sticky spots. Finally, the pockets are scored.

2.2.3 vdW-FFT

Bliznyuk and Gready (1999) developed method called vdW–FFT (van der Waals–fast Fourier transform) to find possible binding sites. At first the protein is projected onto two different 3D grids. The interaction energies of the vdW are calculated in each of these grid points – the first grid has repulsive energy and the other grid has attractive energy. These energies are calculated in both AMBER-94 (Cornell *et al.* 1995) and OPLS (Jorgensen and Tirado-Rives 1988) force fields (Bliznyuk and Gready 1999). Then the method finds the best geometrical match, which correspond to the minimum interactions energy by fast Fourier transformation (Harrison *et al.* 1994, Bliznyuk and Gready 1999). Bliznyuk and Gready have also expanded their vdW-FFT method (Bliznyuk and Gready 1998). Extension saves the 100 best results of FFT-based search (MM energy optimized). Resulted sites are grouped together if the RMS is

less than 0,2 Å. Finally, the PRFM (Parameterized Reaction Field Multipoles) (Bliznyuk and Gready 1995) and FDPB (Finite Difference solution of the Poisson-Boltzman equation) (Sharp and Honig 1990, Honig *et al.* 1993) calculations of solvation energies were performed on ten sites with the lowest MM energies (Bliznyuk and Gready 1998).

2.2.4 Method by Elcock

Adrian Elcock (2001) uses continuum electrostatics method (Honig and Nicholls 1995) to identify functionally important residues in otherwise uncharacterized protein. The method calculates electrostatic free energy (G_{elec}) to each amino acid side chain (Elcock 2001). Then functional active sites are located on the basis of electrostatic free energy. It is assumed that the less stable residues are functionally more active. The method itself does not predict any binding sites, but it is probably safe to assume that these active areas are the most likely to also be potential binding sites. According to Elcock there is one weakness in the method: the process is difficult to fully automate (Elcock 2001), thus slowing down the processing of large datasets.

2.2.5 CS-Map

CS-Map (Computational Solvent Mapping) algorithm identifies the most favorable binding positions by mapping protein surfaces using small organic molecules as molecular probes (Kortvelyesi *et al.* 2003). First, an algorithm searches and scores the regions with favorable electrostatics potential and desolvation. Then the algorithm refines each probe-protein complex by minimization and re-scoring based on electrostatics, van der Waals, and desolvation. The minimized probes are grouped into clusters and ranked on the basis of average free energies. After this, the algorithm divides clusters with the minimum average free energy into the sub-clusters based on free energies and the orientations of probes. The sub-clusters are ranked on the basis of the probabilities (p_{ij}, the ratio of sum of sub-clusters Boltzman factors and the sum of the entire cluster's Boltzman factors). Finally, the algorithm

determines consensus sites by finding the positions at which most probes of different types are overlapping.

2.2.6 DrugSite, PocketFinder and DSite

An and co-workers have presented two binding sites prediction algorithms, DrugSite (in 2004) and PocketFinder (in 2005), which are, in fact, identical (An *et al.* 2004 & 2005). Both methods use a transformation of Lennard-Jones potential calculated from a three-dimensional protein structure (An *et al.* 2004 & 2005). Both binding sites prediction algorithms follow four steps. First, it creates a grid potential map of the vdW force field by surrounding protein with aliphatic carbon probes (radius is 1,7 Å) and calculating the vdW interaction energies between the probes and the protein atoms. This potential is calculated in accordance with Lennard-Jones formula. Then the algorithm smoothens the potential map by applying a moving average filter 10 times to reduce density fragmentation and emphasizing to regions with larger cumulative values. In the third step, the algorithm creates the putative ligand envelopes by contouring the threshold of the potential map. In the final step, DrugSite and PocketFinder sort created envelopes by their volumes and filter out those, which are smaller than 100 $Å^3$. These methods are tested successfully also with the APO structures.

In 2010, Volkamer and co-workers presented DSite, the modified version of DrugSite algorithm, which uses a Gaussian filter in place of the moving average filter in second step. This choice is viable, because the iterative application of the moving average filter approximates a Gaussian filter. Furthermore, Gaussian filter is more efficient and introduces fewer artefacts.

2.2.7 Q-SiteFinder

Laurie and Jackson (2005) developed energy-based, Q-SiteFinder called method that uses a methyl molecule as a probe. First, the Q-SiteFinder minimizes the volume of the protein and the program called Liggrid calculates the non-bound interaction energies between the

methyl probe (-CH_3) and protein. Calculations are made on GRID (Jackson 2002) energy field (Laurie and Jackson 2005). Then the probe coordinates are saved and the most favorable interaction energy is set as a threshold. After this, Q-SiteFinder rotates the coordinates back to match the original orientation of the protein. Next, the individual energetically favorable probe coordinates are clustered according to their spatial proximity and the total interaction energies of probes within each cluster is calculated. Finally, the probe clusters are ranked according to their total interaction energies, when the highest interaction energies have been assumed to correspond with the first predicted binding site.

Laurie and Jackson have compared their method with Pocket-Finder algorithm using their own 35 unbound/bound structures data set. It should be noted that this Pocket-Finder is not the same method as An and co-workers (2005) developed Pocket Finder, which is also known as DrugSite (An *et al.* 2004). Pocket-Finder is actually the pocket detection algorithm in LIGSITE method made by Hendlich and co-workers (1997) and it based on Pocket algorithm (Levitt and Banaszak 1992), but Laurie and Jackson called it for some reason with this name. Q-SiteFinder seems to be better than Pocket-Finder and it gets 51 % success rate for unbound state and 80 % success rate for the bound state, when taken only the first predicted site into account. Laurie and Jackson have noticed that Q-SiteFinder is not suitable for detecting the binding sites of small solvent molecules.

2.2.8 DPA and Fast DPA

In 2005, Ming and Wall developed an innovative theoretical tool, called the Dynamics Perturbation Analysis (DPA), to quantify the influence of protein–ligand interactions on probability distributions of reaction rates and protein conformations. Later they used DPA to predict binding sites in protein structure (Ming and Wall 2006). The prediction is based on the identification of sites at which interactions have a large allosteric potential D_x, which is the Kullback–Leibler divergence (Kullback and Leibler 1951) between protein conformational distributions with and without an interaction. Ming and Well compared the DPA algorithm to SURFNET algorithm and the comparison proved DPA have more statistically significant overlaps with the true binding sites than did SURFNET (Ming and Wall 2006).

However, DPA is computationally very expensive and in 2008, Ming and co-workers presented improved version of DPA, which reduced significantly calculation time (Ming *et al.* 2008). This Fast DPA was applied first-order perturbation theory and replaced matrix diagonalization calculations by matrix-vector multiplication calculations for each test point. The comparison between original DPA and Fast DPA proved the performance of Fast DPA is comparable to that of original DPA but the calculations that took hours using DPA could be performed now in a matter of minutes.

2.2.9 Binding Response

Zhong and MacKerrell (2007) developed a method called Binding Response to determine the potential binding site by using collection of already developed methods and programs. First, protein surface is defined by SAS, which is calculated with Connolly's (Connolly 1983) algorithm (Zhong and MacKerrell 2007). Then the putative binding regions are defined using the sphere-based method that was developed in the context of the program DOCK (Kuntz *et al.* 1982). These created probe spheres that are re-clustered by CHARM program (Brooks *et al.* 1983) and for further analysis the top 10 clusters are selected on the basis of the number of the constituent spheres in each cluster (Zhong and MacKerrell 2007). The geometric properties are determined for each of the selected potential pocket. After this, the binding energy of each potential binding site is determined by docking a test set of 1000 compounds into each potential binding site. This test set is selected from the group consisting of three million druglike compounds (Huang *et al.* 2004, Pan *et al.* 2003) that are based on Lipinski's *rule-of-five* (Lipinski *et al.* 1993). Finally, the 'binding response' is calculated for each binding site, which is determined by the sum of the binding energies between the pocket, and for each ligand from the test group minus the descriptor of ligand distance from the pocket (Zhong and MacKerrell 2007). The binding sites are ranked according to the binding response. The success rate of Binding Response is a good (Top1 90 % and Top3 100 % with 29 proteins test set), but the method is rather slow (in 2007, computational time was an average of 60 hours per protein).

2.2.10 AutoLigand

Harris and coworker (2008) presented a method called AutoLigand that predicts the binding site by searching the space surrounding the protein and finding the contiguous 'envelope' with the specified volume of atoms, which has the largest possible interaction energy with the protein. First, user needs to choose two pre-parameters: the particular force field that will be used and the number of atoms of ligands (the number of envelopes points). Then AutoLigand pre-calculates the potential affinity map of protein into 3D grid by using the same researchers group developed AutoGrid (Morris *et al.* 1998) application with default settings (Harris *et al.* 2008). Then method identifies the optimal envelopes within the 3D grid maps in three steps. In the first step, AutoLigand creates ten best solutions of envelopes by flood fill technique, which matches the size of the user's pre-selection ligand. In the second step, these ten envelopes are optimized by local migration. This occurs by adding the neighbor point with the best energy/volume value and removing the worst point from the current envelope list. In the third step, AutoLigand searches for higher affinity sites by extending a set of linear rays of up to 10 grid points away from the edge points of the migrated envelope. If these kinds of sites are found, the envelope is extended to this region and correspondingly the same number of the envelope points is removed from elsewhere because the total number of envelope points does not change. AutoLigand repeats second and third step until the envelope converges on a consistent low energy solution. The best envelope is found by overlapping the envelopes with each other.

The method has been tested on a set of 96 protein-ligand complexes (Gunasekaran and Nussinov 2007) with the crystallographic structures of APO (unbound) and HOLO (bound) forms (Harris *et al.* 2008). AutoLigand was able to predict the binding site in 80% of the apo structures, but in all these cases the sizes of binding ligands were already known. So, the main weakness of this method is that the user needs to know, or estimates, the potential size of the ligand.

2.2.11 Method by Morita *et al.*

In 2008, Morita and co-workers presented the unnamed variation of Q-SiteFinder. Q-SiteFinder is extended in three ways. First, to differentiate from Q-SiteFinder, this method coats the surface of protein with nine layers of methyl probes (Morita *et al.* 2008). These nine layers are made by the iteration of the Double Cubic Lattice Method (DCLM) (Eisenhaber *et al.* 1995). This provides better probe distribution, which is equivalent to the finer grained energy grid (Morita *et al.* 2008). Second, interaction calculations are made by using the AMBER-94 force field (Cornell *et al.* 1995). Third, method uses the two-level clustering technique with two different thresholds (Morita *et al.* 2008). Low threshold value is used for cluster seeds and higher value for the extensions of clusters. This helps to filter out meaningless clusters. Morita and co-workers used the same 35 unbound/bound data set that Laurie and Jackson used with testing Q-SiteFinder (Laurie and Jackson 2005) and they have compared their method with Q-SiteFinder and Pocket-Finder, which is not to be confused with PocketFinder (See above in Section 2.2.7). The comparison proved that their method seems to succeed slightly better than Q-SiteFinder or Pocket-Finder.

2.2.12 SiteHound

The SiteHound algorithm identifies the potential binding sites by recognizing regions characterized by favorable non-bonded interaction with a chemical probe (Ghersi and Sanchez 2009a, 2009b, Hernandez *et al.* 2009). So, the basic idea of SiteHound algorithm is quite similar to Q-SiteFinder and method by Morita *et al.*, but SiteHound can use the phosphate probe in addition to the methyl probe. This enables it to determine the different types of binding sites by changing the probes type (Ghersi and Sanchez 2009a, Hernandez *et al.* 2009). At first algorithm defines the interaction between the probe and the protein and describes Affinity Maps (also called Molecular Interaction Fields, MIF) (Ghersi and Sanchez 2009a, 2009b, Hernandez *et al.* 2009). MIF of methyl probe is calculated using AutoGrid program (Morris *et al.* 1998) and MIF of phosphate probe is calculated using EasyMIF program (Ghersi and Sanchez 2009b). After determining MIF, SiteHound filters out the map points with unfavorable affinity (Ghersi and Sanchez 2009a, Hernandez *et al.* 2009). The algorithm clusters the remaining points according to their spatial proximity using the

agglomerative hierarchical clustering algorithm. Finally, the clusters are ranked by reflecting the putative binding sites to the Total Interaction Energy (TIE, sum of the energy values of all points that belong in the same cluster) and the top ten clusters will be displayed. SiteHound was capable to identify the correct binding sites among the top three in 76 % of APO structures cases (Ghersi and Sanchez 2009a). Used dataset is based on Astex Diverse Set (Hartshorn *et al.* 2007). The algorithm can also be found in SITEHOUND-web version (http://sitehound.sanchezlab.org) that can be used with the internet browser and the results can be downloaded in various formats (Hernandez *et al.* 2009).

2.3 Evolutionary-based

Evolution-based methods, also known named Sequence based methods, are based on the observation that the protein functional areas usually hit a certain fraction of the amino acid sequence (Casari *et al.* 1995) and these functional areas are mostly the binding sites of druglike ligands. Based on this, various methods have been developed to identify the functional sections of unknown proteins by comparing their amino acid sequences to the already known amino acid sequences of proteins. Many different evolutionary methodologies are used to make prediction e.g.: SequenceSpace (SS) (Casari *et al.* 1995), Evolutionary Trace (ET) (e.g. Lichtarge *et al.* 1996), Maximum Likehood (ML) (e.g. Dean and Goldning 2000), Maximum Parsimony (MP) (e.g. Armon *et al.* 2001), Conserved Functional Group (CFG) (Innis *et al.* 2004, Innis 2007), and Neighbor-Joining (NJ) (Saitou and Nei 1987).

The evolutionary-based methods are fast and robust, and mostly the prediction needs only a protein sequence (Except method by de Rinaldis *et al.* and siteFINDER|3D). Usually the input file is PDB file or some sequence format file (e.g. FASTA). Most methods create a Multiple Sequence Alignment (MSA) file, which is the basic format used to estimate a conservation rate of amino acids with the evolutionary-based methods, but some methods are able to receive the user made MSA file. However, it should be noted that these methods do not work if there is not at least one similar protein, which function and structure are already known. The binding site prediction can be also limited for two reasons (Lichtarge and Sowa 2002). First, functional areas of proteins can be large, when functionally important amino

acid residues may be located many different short parts in sequence. Therefore, may be more difficult to identify such areas than those, wherein functional area is located to short exactly local sequence area. Second, when sequence identification falls below 40-50 %, the functional analogies can be fallacious (Casari *et al.* 1995, Lichtarge and Sowa 2002).

2.3.1 Sequence Space

In 1995, Casari and co-workers introduced the first evolutionary-based binding site detection method that called 'Sequence Space'. The method is based on analysis on protein MSA (Doolittle and Feng 1990) and each analyzed sequence is represented as a vector-point in a 3D space (Sequence Space), whose basic dimensions are residue positions and residue types (Casari *et al.* 1995). The direction of vectors describes proteins subfamily and the lengths represented the degree of conservation. Then method detects those residues, which have a tendency to be conserved within a subfamily of proteins, but differ between subfamilies (Tree-determinant positions).

2.3.2 Evolutionary Trace

Lichtarge and co-workers (1996) developed method called Evolutionary Trace (ET). This method builds a phylogenetic tree from the MSA (Lichtarge *et al.* 1996) by the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (Armon *et al.* 2001, Pupko *et al.* 2002). The resulting tree is used to cluster closely related sequences and to find a consensus sequence for each cluster and each position. After this ET compares the consensus sequences and labels each position with status of either 'neutral', 'class-specific' or 'conserved'. These are described onto the 3D-structure of the protein by color coding. ET also ranks residues according to the consensus so that residues with lower numbered ranks are considered more important than those with higher numbered ranks. All invariant (conserved) residues get the ranking value of 1, because they are considered the most important for the functioning of the enzyme and all other residues a get ranking value of 2 further. The method has been able to identify functional significant regions of proteins

successfully with 80% of the tested proteins (Lichtarge and Sowa 2002). The main weakness of the method can be considered that it is difficult to automate (Pettit *et al.* 2007). Although ET method works also with protein-ligand complexes, it was mainly used in identification protein-protein and protein-DNA functional sites (Lichtarge *et al.* 1996, Lichtarge and Sowa 2002). ET method was first attempted to take into considered the evolutionary history of a protein family, when determining the functionally important residues (Armon *et al.* 2001).

2.3.3 Variations of ET method by Landgraf *et al.*

There are also couples of variations of ET method. In 1999 Landgraf and coworkers introduced modified ET method by adding discriminating power to analysis by a quantitative measure of residue variation at each position (Landgraf *et al.* 1999). The same researcher group have also designed another extension of ET method by adding three-dimensional cluster analysis, so that it searches for Tree-determinant positions in a multiple sequence alignment whose mutational behavior is similar to the mutational behavior of the whole family (Landgraf *et al.* 2001). There were two goals of this extension: The first was to improve the sensitivity with the identification of functional residue clusters. Second was the ability to identify functionally important residue clusters without reliance on a phylogenetic tree as input data.

2.3.3 Method by de Rinaldis *et al.*

De Rinaldis and co-workers method (1998) is composed of two independent modules. The first one forms a 3D grid around a three-dimensional structure of the protein and the grid cells are assigned to the residue whose atoms occupy the highest volume (de Rinaldis *et al.* 1998). The cells of the 3D grid correspond to single positions in the protein sequences. Then method can create 3D multiple alignment grid with different transformed template proteins. The 3D multiple alignment grid corresponds to columns of a protein sequence multiple alignment. Each cell of the 3D multiple alignment grid is then associated with a profile row, calculated using the Dayhoff matrix (Schwartz and Dayhoff 1979). The profile row allows the

identification of conserved residues and cells, called ‘heavy cells’ (de Rinaldis *et al.* 1998). The second module arranges ‘heavy’ cells in all possible combinations of three cells that are triangles in 3D profile grid. These triads are used as search keys when comparing the protein with the protein structure database. Then second module estimates similarity with special scores and finds the most similar protein. Still in fact, this method is not able to automatically predict any binding sites.

2.3.4 Method by Dean and Golding

Dean and Golding (2000) developed Maximum Likehood (ML) method for discriminating between the slowly and rapidly evolving regions of a protein. The calculations of likelihoods are made by the method of Felsenstein (1981) and then method evaluates empirical the likelihood of observing replacement rates in small regions throughout the molecule (Dean and Golding 2000). The development of this method was only focused on the enzyme family of eubacterial isocitrate dehydrogenases, so the functionality of the method in general is quite difficult to evaluate. However, ML approach was used later to functional site detection in other methods, so also this method could work with other proteins.

2.3.5 Method by Aloy *et al.*

Aloy and coworkers (2001) developed method, which is based on the idea that the protein functional areas can be identified by finding the clustered invariant polar residues. At first, method uses a MSA to identify invariant polar residues and maps those residues onto the structure of examined protein (Aloy *et al.* 2001). Then it forms the spatial clusters of these invariant polar residues based on their spatial proximity. Finally, method predicts functional sites by overlapping spatial clusters to the observed functional site, which is defined a single sphere that encloses those residues, which are in PDB showed to be involved in the activity of protein. Aloy and coworkers have benchmarked their method with the ET method, as the approaches are similar in both methods. The success rate of this method is approximately

80 %, like in the case of ET method, and it was mainly used, as well ET method, for the identification of protein-protein and protein-DNA complexes functional sites.

2.3.6 ConSurf and ConSurf 3.0

In 2001, Armon and coworkers introduced conservation surface-mapping (ConSurf) method for mapping evolutionarily conserved regions on the surface of proteins of known 3D-structure. First, the method constructs the Maximum Parsimony (MP) tree from the MSA (Armon *et al.* 2001). Then the ConSurf attempts to rebuild the tree, minimizing the number of replacements, and reconstructing the ancestral sequences. Replacements of amino-acids are mapped onto the tree. Finally, ConSurf defines the conservation score, which consist of the total number of replacements weighted by the distance between each pair of amino acids. Armon and coworkers (2001) have compared ConSurf with ET method and they have found that Consurf is much more sensitive and better than the ET method, because the algorithm of ConSurf takes into account the fact that amino acids differ in frequency as well as the uncertainty of the rebuilt ancestral sequences (Pupko *et al.* 2002).

In 2005, ConSurf 3.0 was published and it uses an empirical Bayesian method for scoring conservation (Landau *et al.* 2005) being more accurate than ML or MP methods (Mayrose *et al.* 2004). There is also way to control various additional steps in the calculation by a number of advance options, which enables further improving the accuracy of the calculation (Landau *et al.* 2005). ConSurf 3.0 is available via web browsers in the address: http://consurftest.tau.ac.il/ (Glaser *et al.* 2003, Ashkenazy *et al.* 2010).

2.3.7 Rate4Site

Pupko and coworkers (2002) introduced Rate4Site method, which maps the rate of evolution among homologous proteins onto the molecular surface of one of the homologous whose 3D-structure is already known. Rate4Site use MSA for input, like ET and ConSurf, but unaligned input sequence can be also used in which case Rate4Site generates the MSA automatically using CLUSTAL W (Thompson *et al.* 1994) with default parameters (Pupko *et*

al. 2002). At first method reconstructs a phylogenetic tree from inputted sequences, but there is also possibility to use the whole advance phylogenetic tree for input. If the number of sequences is less than 20, Rate4Site uses ML method (e.g. Felsenstein 1981, Dean and Golding 2000, Friedman *et al.* 2002) to find the most likely tree. In other cases Rate4Site uses the Neighbor-Joining (NJ) (Saitou and Nei 1987) algorithm (Pupko *et al.* 2002). Finally, Rate4Site estimates of the degree of the evolution of amino acid sequences by ML method. Rate4Site and ConSurf are very similar, but Rate4Site will take into account the length of a branch of the phylogenetic tree, which corresponds to the expected substitutions of amino acid at each site. Pupko and co-workers have estimated that ConSurf is more susceptible to make mistakes in terms of the variable component and gapped positions than the Rate4Site, because ConSurf takes into account to those too precisely. Rate4Site treats gapped positions as missing data and uses only a subset of the sequences for determining the conservation score. Even so, this is not a complete solution, because the main problems of calculation conservation grades are distinguishing between amino acids that are conserved due their functionality and those than appear to be conserved due to the shortness of evolutional time.

2.3.8 Method by Mayrose *et al.*

The same team of researchers that developed Rate4Site, ConSurf, and PatchFinder (Nimrod *et al.* 2005, 2008) introduced the variation of Rate4Site derivatives using Bayesian approach that presupposes a prior distribution of evolutionary rates (Mayrose *et al.* 2004). The results indicate that this kind of approach is better than the ML method (Rate4Site), especially when the number of available homologous sequences is small. Mayrose and co-workers also demonstrated that it is better to start estimating branch length and proceed to only after this to estimating site-specific rates when branch length is unknown.

2.3.9 Method by del Sol Mesa *et al.*

In 2003, del Sol Mesa and co-workers presented a method combining of three separate methods that represent the range of available approaches. The first of three method is 'The Level Entropy method' (so called S-method) that is based on the automatic search for the optimal division of the family to the subfamilies, which was used to estimate to the number of Tree-determinants involved in the function of the protein family (del Sol Mesa *et al.* 2003). The general model of S-method is developed by Hannenhalli & Russel (2000). Second method is the 'Mutational Behavior Method' (MB-method), which is actually the implementation of the extension of ET by Landgraf *et al.* (2001). The third method, 'Sequence Space Automatization Method' (SS-method), is based on above introduced Sequence Space method (Casari *et al.* 1995).

The benchmarking of the method has been assessed on their own dSM (del Sol Mesa) (Nimrod *et al.* 2008) test set, which contains 112 non-redundant protein families with annotations for functionally important sites (SITE records in PDB) (del Sol Mesa *et al.* 2003). The combination of two or three these different methods provides more accurate tool for the prediction of more significant set of functionally important residues than any of these methods alone. Based on the results of these methods Del Sol Mesa and co-workers presented attention at evolutionary-based methods fail more often to predict the binding sites of small molecules (e.g. ion) as large molecules (e.g. nucleotides). This trend can also be seen in other evolutionary-based methods.

2.3.10 Method by Innis *et al.*

Innis and co-worker developed a functional site prediction method, which use Conserved Functional Group (CFG) analysis (Innis *et al.* 2004). CFG is akin to ET and it is derivatives, but it focuses on the relative extent of functional/chemical group conservation throughout a protein structure. The idea of the approaches of CFG has been simplifying the Venn diagram (Taylor 1986) for amino acids residues (Innis *et al.* 2004). The underlying assumption behind CFG analysis is that functionally important regions of protein display a higher degree of conservation compared with other parts of the protein. Prediction of functional active sites

is based on a map of local CFG "density", which consists in a protein structure and a number of its sequence homologous. The region of the highest value of CFG density map indicates the most likely functional site. Innis and co-worker have tested the prediction ability of CFG method with 470 proteins and it can provide the overall success rate for even 95,7 %, but if only the top predictions were considered, 48,5 % correct hits and 35,1 % partially correct hits were obtained.

2.3.11 PatchFinder (2005 & 2008)

In 2005, developer team of ConSurf, Rate4Site, and variations of those introduced PachtFinder called method, which has many similarities with its predecessors (Nimrod *et al.* 2005). PatchFinder uses the Rate4Site algorithm to count an evolutionary conservation score to each amino acids position. Then method identifies the 'buried' and 'expose' residues by calculating ASA with Surface Race program (Tsodikov *et al.* 2002). Identification of functional patches is based on assumption that the more larger and conserved a patch is, the more likely it is to be a functional region (Nimrod *et al.* 2005). Identification of the most significant cluster of conserved residues on the protein's surface is based on the ML method. PatchFinder can also identify secondary, non-overlapping patches, which may represent secondary functional regions. For benchmarking the method Nimrod and co-workers have used dSM test set.

In 2008, Nimrod and co-workers published an enhanced version of PatchFinder. In this version evolutionary conservation is computed by using the Bayesian version of Rate4Site (Mayrose *et al.* 2004), because it is evidently superior to the ML version, especially when the number of available homologous sequences is small. Second improvement was the using Delaunay triangulation (Barber *et al.* 1996, de Berg *et al.* 2000) to describe the neighborhood and accessibility to the solvent of each residue (Nimrod *et al.* 2008). This version is also available in the address: http://patchfinder.tau.ac.il/. Also this upgraded version of PatchFinder is tested with dSM test set. It was also compared with PatchFinder 2005, HotPatch (Pettit *et al.* 2007), siteFinder|3D, and ET viewer (Morgan *et al.* 2006). According to comparison results, PatchFinder version 2005 and 2008 is equally effectual

(2005 version even little bit better) and both are much better than HotPatch, siteFinder|3D or ET viewer.

2.3.12 HotPatch

Pettit and co-workers (2007) developed a method, HotPatch, that finds surface patches of unusual physicochemical properties on protein structures, and estimates the patches' probability of overlapping functional sites. The prediction process can be divided into three main steps (Pettit *et al.* 2007). First, HotPatch evaluates the property of interest for all atoms in the protein. Second, it clusters atoms with the high values of the property together. Third, HotPatch assigns a statistical score called Functional Confidence (FC) to each of the clustered patches. This FC describes probability that the patch to overlaps a functional site. HotPatch is intended for more like the identification of macromolecules (protein-protein and protein-DNA) binding sites. Pettit and co-workers have noticed that the binding sites of small molecules are difficult to predict with HotPatch.

2.3.13 siteFINDER|3D

siteFINDER|3D is an online tool, which uses Conserved Functional Group (CFG) analysis (Innis *et al.* 2004) for predicting the location of functionally important regions within a protein of known structure (Innis 2007). It is basically the method by Innis *et al.* but implemented as web-based platform. siteFINDER|3D is available on the http://sitefinder3d.mbb.yale.edu/ and requires, at a minimum, the atomic coordinates of query protein in PDB format.

2.3.14 ConSurf 2010

In 2010, researcher group e.g. of ConSurf and Rate4Site, published a new version of the ConSurf web server that combines two independent methods, ConSurf 3.0 (*Landau et al.* 2005) and ConSeq (Berezin *et al.* 2004), providing an easier and more intuitive step-by-step interface, while offering more user-adjustable settings during the prediction process

(Ashkenazy *et al.* 2010). In addition, this version uses Rate4Site algorithm (Pupko *et al.* 2002) to calculate the evolutionary rates for nucleic acid sequences. For input, a sequence of protein could be given straight or it could be extracted from the 3D-structure (PDB file). ConSurf 2010 collects homologous sequences from the selected database, e.g. SWISS-PROT (Boutet *et al.* 2007), UNIREF-90 (Suzek *et al.* 2007), or UniProt (UniProt Consortium 2007), using BLAST (Altschul *et al.* 2005) or PSI-BLAST (Altschul *et al.* 1997). The user may specify criteria for defining homologous, but it is also possible to manually select the desired sequences from the BLAST results. At the next stage ConSurf 2010 removes redundant and unrelated sequences and constructs a MSA and reconstructs a phylogenetic tree using NJ algorithm. It is also possible to provide by user MSA and/or phylogenetic tree. After this, ConSurf 2010 uses Rate4Site (Bayesian or ML) algorithm to compute position-specific conservation scores. Finally, the scores are projected on the protein, nucleic acid structure, or sequence. ConSurf 2010 can be used in the address: http://consurftest.tau.ac.il/ (Ashkenazy *et al.* 2010).

2.4 Blind Docking and Molecular Dynamics methods

The huge growth of computers computational power is allowed to press computationally heavier techniques for finding the binding site of unknown protein. Molecular dynamics (MD) calculations and docking algorithms can take into account much more details from the environment of protein than other methods. In Blind Docking methods, test ligands are docked without any pre-setting of binding site location finding the right binding site. MD calculations have used with two different approaches. Other one based on mobility of water molecules. Other one uses long scale MD simulations to finding the right binding site of already known active ligand. These approaches are the most useful in a situation where the ligand that binds in the target protein is known, but the binding site is unknown. These methods have always needed to know the 3D-structure of proteins. The main weakness in these methods is their computational expensiveness.

2.4.1 Blind docking

In 2002, Hetényi and van der Spoel were among the first that used docking program to finding binding site of unknown protein without any prior knowledge of their location and conformation. They named the method to 'Blind Docking' (BD), because the docking algorithm is not able to "see" the binding site, but despite this it can find it (Hetényi and van der Spoel 2002). Docking was carried out by using AutoDock program (Morris *et al.* 1998) but also others docking programs, e.g EADock (Grosdidier *et al.* 2007), can be used (Hetényi and van der Spoel 2002, 2011). Peptides were used to test ligands, because they have various functional groups, they have several possible torsional degrees of freedom, and they are composed of amino acids, so the same force field parameters can be used as for the target molecule (Hetényi and van der Spoel 2002). Docking was made in AMBER-94 force field (Cornell *et al.* 1995).

Ability to distinguish the real binding site on the protein from nonspecific and/or energetically unfavourable ones is the most important requirement of a blind docking calculation (Hetényi and van der Spoel 2002). For this reason, the docking parameters are very important role and Hetényi and van der Spoel have striven to find the best docking parameters for BD. The success of docking depends critically also on the quality of target protein structure. Hetényi and van der Spoel have also found some evidence that solvent molecules can be important in the finding of binding sites. The presence of water increases the computational time of docking for the already computationally expensive method, which is one of the main weaknesses of blind docking. One docking process is taken 8 to 89 hours with rigid ligand and 21 to 1017 hours with flexible ligand, depending on a much the size of the target protein and, and available hardware and the number of trials and estimates. Therefore, for some case it may be necessary to use explicit water molecules to explore the binding site to reducing computational time (Minke *et al.* 1999, Hetényi and van der Spoel 2002). BD method was successfully identifying 79 % of the used 43 proteins test set (Hetényi and van der Spoel 2006).

2.4.2 Optimized MD simulations

In 2004, Bhinge and co-workers developed Optimized MD simulations (OMD) for detecting possible ligand binding sites in protein. Method is optimized to detect and quantify interior cavities as well as surface pockets (Bhinge *et al.* 2004). The detection is based on the observation that the mobility of water in such pockets is significantly lower than that of bulkwater. OMD calculates Voroinoi volumes of each detected pockets from solvated protein structures derived from an MD simulation. The pocket ranking is based on the assumption that the binding site is usually the largest pocket. The comparison between standard MD simulation and OMD showed that OMD made prediction 20-30 times faster than a standard MD simulation. Bhinge and co-workers compared OMD with CAST and MC procedure (Monte Carlo procedure -based approach that measures the Voronoi volume of a cluster of overlapping spheres that map the cavity, Chakravarty *et al.* 2002) and the comparison showed OMD succeed a little better than the CAST or MC procedure.

2.4.3 Method by Aita *et al.*

Aita and co-workers (2010) presented a new method to predict binding sites of peptides. They wanted to use blind docking methods at low computational costs (Aita *et al.* 2010). Method based on the ideas of Delaunay tessellation of a coarse-grained protein–peptide complex and a four-body statistical pseudo-potential. First, the algorithm constructs a tessellated protein structure. For this, each amino acid residue is represented by a single fixed point, which is defined using the “side-chain center” as the geometric center of the coordinates of all atoms of side-chains. Then on the basis of all these points together simplified protein is formed by Delaunay tessellation. Before starting blind docking, the globular shell-like grid surrounding a tessellated protein structure is formed by placing grid-points at intervals of 1,0 Å around the surface of the entire protein target. Then blind docking system describes the variable coordinations of peptides, which includes the binding site on the protein surface, the conformation and the orientation of the peptide, and based on this it defines so called ‘fitness function’. After this, method describes a process of finding candidates for the correctly docked conformation of a peptide based on fitness of peptide-coordinations. Finally, the binding site is predicted by blind docking using Hill-climbing

optimization with 1000 trials. This method is much faster than original BD method. One prediction is taken 0,4 to 12,1 hours (Intel Itanium2 dual core processor × 6 (= 12 cores); 6,4 GFLOPS/core; 32GB of RAM), depending on a lot of number of grid points. Of course, there should be noted that the used computational power is much more powerful than the almost ten years earlier.

2.4.4 MolSite

Japanese Fukunishi and Nakamura (2011) developed method called Molecular-docking binding-site finding (MolSite) to predict the binding site. The basic idea was the assumption that the true biding site of protein indicates a greater affinity for random ligands as the rest of the protein (Fukunishi and Nakamura 2011). They also assumed that if the true binding site of a ligand was correctly predicted, its affinity would correlate to the docking scores of the random compounds. So basically, MolSite docks onto the whole protein surface various compounds from a random compound library, which did not even necessarily include the ligand corresponding to the true binding site. Then the docking score indicate the position, which is supposedly the true ligand binding site.

MolSite is benchmarked with four datasets and dataset A was used to compare Molsite with FINDSITE (Bryliński and Skolnick 2008), LIGSITE (Hendlich *et al.* 1997), PASS (Brady and Stouten 2000), Q-SiteFinder (Laurie and Jackson 2005), SURFNET (Glaser *et al. 2006*), and MetaPocket (Huang 2009). When only the single top-ranked site was adopted, predicting accuracy of MolSite was 80-99 % of the 89 proteins test set, dependent the distance limits and how this distance are measured. Dataset A contains only HOLO structures so it is not the most favorable for the estimation of the binding site prediction ability. Anyway, MolSite scores better than other method for predicting the binding site. Fukunishi and Nakamura have noticed that when the binding site had small volume, the correlation between docking score and experimental affinity was weak, with a correlation coefficient of 0,44. So docking scoring cannot be used for evaluation of binding affinity. The main drawback of MolSite is computational expensiveness. The docking calculation of one protein, which is divided into 50 scoring grids, requires 500000 calculations with the 10000 compounds library. This took about 280 hours by one processor with available hardware.

2.4.5 Long scale MD

Shan and co-workers (2011) investigated how molecules find their binding sites by using the unguided long scale molecular dynamics (MD) simulations of protein-ligand binding and they noticed at this kind method could be used to identify the previous unknown binding sites. The MD simulations were conducted with cancer drug dasatinib and kinase inhibitor PP1 to Src kinase enzyme (Shan *et al.* 2011). Dasatinib found the correct binding site in one of the four separate simulations (total time 35 µs) and PP1 found the correct binding site in three of the seven separate simulations (total time 115 µs). In the true binding sites, which long scale MD has found, the ligands are virtually identical posed to those captured by X-ray crystallography. The method could also provide a promising tool for identifying the allosteric binding sites of protein, because the researcher noticed that the PP1 repeatedly grabbed with several previously known Src kinase allosteric sites during the MD simulations. This method has also obtained good results for predicting the critical water placement of the pocket. The main drawback of method is huge computational cost so probably this method will not be generally a part of the drug design process in the near future but some special cases, e.g. with flexible protein, this method could probably provide a much better success rate than the traditional methods. The illustrative video to the dastinib's binding site discovery process are found in the address (Shan *et al.* 2011): http://pubs.acs.org/doi/suppl/10.1021/ja202726y/suppl_file/ja202726y_si_001.avi

2.5 Combined approaches

In the 2000s, it was clear that none of the methods described above are good enough to universally detect the right cavities of proteins and to rank the detected cavities according to druggability. The methods have been developed so computational as fast as possible leading to the situation where the methods often fail in certain types of cases, in which the algorithms are not able to take into account correctly or adequately all properties of the target site. Furthermore, it is noticed that the endless improvement of old algorithms does not necessarily produce much better results. The thought that extending the methods with other types of methods, which aims to patch up the weaknesses of other algorithm, was

born. This solution seems to be clear and simple, because a combination of two or more methods is not a problem, but which results are interpreted as "correct" it is a trick of this approach.

Del Sol Mesa and co-workers (2003) were among the first on the trail of the combined approaches, but their method was the combination of three separate evolutionary-based methods – not a combination of different type of methods. Glaser and co-workers (2006) and Huang and Schroeder (2006) were the first to extend the old methods by adding a different type of approach. At first, their methods aims to find potential pockets using geometry-based algorithms, followed by re-ranking those pockets according to the degree of conservation of the involved surface residues (Huang and Schroeder 2006, Glaser *et al.* 2006).

2.5.1 SURFNET-ConSurf

Glaser and co-workers (2006) presented the first method, which was the combination of two different type approaches. As its name suggests, SURFNET-ConSurf consists of both SURFNET (Laskowski 1995) that ConSurf (Armon *et al.* 2001) methods (Glaser *et al.* 2006). The first part of method finds the four largest clefts by using SURFNET algorithm. Then the second part trims each clefts volume according to the degree of conservation of residues into that cleft area. This combination of methods aims to determine which of these four pockets would be the most potent binding site. The trimming process of second part based on a residue conservation score that are obtained from the ConSurf-HSSP database version 1.0 (Glaser *et al.* 2005). The ConSurf-HSSP database provides estimates for the rate of evolution of each amino acid in a PDB structure. The residue conservation scores are calculated using MSA of a homology-derived secondary structure of proteins (HSSP) (Sander and Schneider 1991), which corresponds to the input of the Rate4Site algorithm (Glaser *et al.* 2006). SURFNET-ConSurf provides a promising tool for identifying important areas and residues in binding sites, but as Glaser and co-workers have noted their method fails in cases where the same binding site includes more than one ligand molecule.

2.5.2 LIGSITE^{CSC}

LIGSITE^{CSC} (Connolly surface and conservation) is the extension of Pocket, LIGSITE and LIGSITE^{CS} methods, which are presented in the Sections 2.1.1.5 and 2.1.1.10. Difference to LIGSITE^{CS}, in LIGISTE^{CSC} the prediction of binding sites is not based only on protein geometry, but more like combined SURFNET-ConSurf approach. At first, LIGSITE^{CSC} determines the possible biding site using LIGSITE^{CS} and then those possible pockets are re-ranked according to the degree of conservation of the involved surface residues into the area of the 'surface-solvent-surface' events. The conservation score for each residue in proteins is obtained from the ConSurf-HSSP database (Glaser *et al.* 2005), where the alignments of all proteins to PDB structures are made according to HSSP curve. In ConSurf-HSSP database, all proteins are re-weighted using a reconstructed phylo-genetic tree and by estimated evolutionary rates of each amino acid position. This conservation step can be carried out only on proteins, for which conservation score from the ConSurf-HSSP database can be found (Huang and Schroeder 2006).

2.5.3 FINDSITE and FINSITE^{LHM}

Bryliński and Skolnick (2008) presented method called FINDSITE, which identifies possible binding sites by comparing the superimposed structures of set of distantly homologous proteins and the target protein. The distantly homologous means that proteins homologous to threading templates with a sequence identity to the target sequence >35 % are rejected. At first, FINDSITE recognizes a set of ligand bound template structures for a given target sequence from distantly homologous proteins by the PROSPECTOR_3 (Skolnick *et al.* 2004) threading approach (Bryliński and Skolnick 2008, 2009). Then method uses TM-align (Zhang and Skolnick 2005) approaches to superimpose HOLO structure of each the ligand bound threading template onto the target structure. The clustered mass centers of bounded ligands into superimposed structures indicate the expected binding sites and they will be ranked according to the number of threading templates that share the same binding site.

Bryliński and Skolnick (2009) have also developed FINDSITE to further. FINDSITE^{LHM} (Ligand docking by Homology Modeling) is a very simple, robust and rapid approach to predict the

binding pose of the already known ligand (Bryliński and Skolnick 2009). This method uses FINDSITE algorithm to find the most likely, so-called anchor substructure. Then, the some known ligand is docked onto this anchor area and FINDSITE^{LHM} defines ligand binding pose using all-atom minimization by AMBER (Pearlman *et al.* 1995) program.

2.5.4 SiteMap

In 2007, Halgren presented a tool named SiteMap for identifying the potential binding sites and for the characterizing binding sites and for predicting their druggability in lead-discovery applications. SiteMap is a part of Schrödinger's Maestro application (Halgren 2007, 2009). The identification of potential binding sites is carried in four steps. First, SiteMap creates 3D grid around to protein and classifies each grid point as being "inside" or "outside" the protein by comparing the distance to nearby protein atoms. In the second step, SiteMap calculates the vdW interactions energies and the value of enclosures for each 'outside' grid point and discards those, which are not inside of threshold limits. In the third step, the remaining grid points (site points) are grouped if they are located within a given distance (default value is 1,67 Å) and the created groups with less than three site points (that is default value) are discarded. In the fourth step, the site point groups are merged when the gap between them is less or equal to a user specified distance (default value 6,5 Å). Halgren has tested SiteMap first with a set of 230 proteins (Halgren 2007) and later with an extensive set of 538 proteins (Halgren 2009) and both sets are taken from the PDBbind database (Wang *et al.* 2004). The binding site identification of SiteMap is provided 96,5 % success rate with 230 proteins test set (Halgren 2007) and 85,9 % success rate with 538 proteins test set. SiteMap seems to be effective methods of identification and the computational speed of SiteMap is also very fast. Unfortunately, neither of the test sets does appear to contain any unbound protein structures.

2.5.5 Focused Docking

Ghersi and Sanchez (2009a) developed an improved version of BD method that was called Focused Docking (FD). Their method restricts the search space for the vicinity of the top three binding sites predicted by the SiteHound program (Ghersi and Sanchez 2009a). Actual BD runs are made independently in each of three boxes, which size is 23 Å x 23 Å x 23 Å, using the same program (AutoDock) and the same docking parameters as in the original BD study (Hetényi and van der Spoel 2002). FD proves better prediction accurate (84 %) than BD (71 %) with 77 proteins test set from Astex Diverse Set (Hartshorn *et al.* 2007). When the comparison was made with 19 APO structures, FD achieved 58 % success rate than BD achieved only 32 % (Ghersi and Sanchez 2008). Focused Docking is not so computational expensive than BD and it provides better prediction accurate. Still it should be noted that the SiteHound algorithm cannot promise with 100% certainty that the true pocket is one of the top three predicted by SiteHound.

2.5.6 MetaPocket and MetaPocket 2.0

In 2009, developer e.g. of LIGSITE^{CS} and LIGISTE^{CSC} presented MetaPocket (MPK1), which is a the combination of three geometric based methods (LIGSITE^{CS}, PASS, and SURFNET) and one energy-based method (Q-SiteFinder) (Huang 2009, Zhang *et al.* 2011). At first, the geometric based methods identify potential binding sites and each identified binding site is presented as a single probe, which has also ranking score (Huang 2009). Then the algorithm of MPK1 automatically sends the protein structure to the Q-SiteFinder server. The predicted binding sites are retrieved automatically and they are represented as probes, which are already clustered. MPK1 calculates a mass center of the probes for each cluster and clusters are ranked by their sizes. Scorings are calculated separately for each site in different methods, because these four methods have different ranking scoring functions, which makes it hard to compare and evaluate the predicted pocket sites directly. Therefore, only the top three pocket sites in each method are taken into further consideration. Finally, these 12 pocket sites are clustered using a simple hierarchical clustering algorithm according to their spatial similarity (distance based) and each cluster is ranked by a special combinational scoring function metaZScore.

MetaPocket 2.0 (MPK2) (Zhang *et al.* 2011) is the extended version of MPK1 and it includes four more freely available binding site identification tools: Fpocket (Le Guilloux *et al.* 2009), GHECOM, ConCavity, and POCASA. The operating principle of MPK2 is exactly the same as MPK1, but now there are 8 used methods, resulting up to 24 different predictions for a possible binding site.

MPK1 is benchmarked with LIGSITE^{CS}, PASS, Q-SiteFinder, and SURFNET, and MPK2 is benchmarked in addition to these with Fpocket, GHECOM, ConCavity, and POCASA. Benchmarking was made on a dataset of 48 unbound/bound structures and a nonredundant dataset of 210 ligand-bound only structures (Huang and Schroeder 2006) and benchmarking of MPK2 was used also new a data set of 198 drug-target (Zhang *et al.* 2011). The first two test sets of these were the same as Huang and Schroeder developed for the evaluating of LIGSITE^{CS} and LIGSITE^{CSC} (Huang and Schroeder 2006). The benchmarking results indicate that MPK2 is able to predict the binding site better than any of these programs alone (Zhang 2011) and MPK1 is also better than LIGSITE^{CS}, PASS, Q-SiteFinder or SURFNET (Huang 2009). MetaPocket 2.0 is freely available at http://projects.biotec.tu-dresden.de/metapocket/

2.5.7 SiteIdentify

Bray and co-workers (2009) presented SitesIdentify for binding site prediction. SiteIdentify can uses two separate approaches, method by Bate and Warwicker (2004) and method by Greaves and Warwicker (2005), which both are developed by the same research group (BrayI *et al.* 2009). Method by Bate and Warwicker calculates the electrostatic potential of protein using Finite Difference Poisson-Boltzmann (FDPB) calculations and the peak potential is predicted as the binding site (Bate and Warwicker 2004). Method by Greaves and Warwicker combines method by Bate and Warwicker with sequence conservation information. That method finds close homologous by running the sequence through PSI-BLAST (Altschul *et al.* 1997) and calculates a normalized conservation score for each residue based on the amino acid and stereochemical diversity and the gap occurrence at that position (Greaves and Warwicker 2005). Then method calculates the peak potential in the same way as the previous method, but now single central atom in each amino acid is weighted with the conservation scores. User selects which method will be used to prediction.

The benchmarking was made with seven methods, ConSurf 3.0, PASS, Q-SiteFinder, Crescendo (Chelliah *et al.* 2004), FOD (Bryliński *et al.* 2007a, 2007b), PDBSiteScan (Ivanisenko *et al.* 2005), and THEMATICS (Wei *et al.* 2007), on a non-redundant set of 237 enzymes with annotated active sites. Method by Greaves and Warwicker seems to be better (74,7 %) than any other method except ConSurf 3.0, which has the slightly better success rate (78,2 %), while the method by Bate and Warwicker (2004) is the fourth best (63,0 %). The open source code of SitesIdentify is available for download on the BMC Bioinformatics website: http://www.biomedcentral.com/1471-2105/10/379

2.5.8 Fpocket

In 2009, Le Guilloux and co-workers introduced Fpocket method that relies on the concept of alpha spheres, introduced by Liang *et al.* (1998b). Fpocket is a part of an open source package, which contains three main programs: Fpocket, Tpocket (organises pocket detection benchmarking), and Dpocket (collects pocket descriptors values), but in this thesis only Fpocket is analyzed. The binding site detection process of Fpocket can be divided into three major steps. The first step determines protein structure by alpha spheres and pre-filtered created spheres (Le Guilloux *et al.* 2009). The second step clusters spheres that lie close to each other, identifies potential pockets, and removes clusters of poor interest. Third step calculates properties from the atoms of identified pockets, in order to score to each pocket. Fpocket is benchmarked against SURFNET, PASS, CAST, LIGSITE, LIGSITE^{CS}, LIGSITE^{CSC}, and PocketPicker with the dataset of 48 unbound/bound protein structures developed by Huang and Schroeder (2006). Benchmarking proved that Fpocket have as good success accuracy as LIGSITE^{CSC} and PocketPicker, while the other methods have weaker performance. Le Guilloux and co-workers are also compared Fpocket with PocketPicker using the datasets of Cheng *et al.* (2007) and Astex Diverse set (Hartshorn *et al.* 2010). The comparison proved a better success rate to Fpocket with both test sets than to the other methods (Le Guilloux *et al.* 2009). Fpocket is also freely available (Schmidtke *et al.* 2010): http://fpocket.sourceforge.net/

2.5.9 ConCavity

Capra and co-workers (2009) developed a binding site detection method called ConCavity that is also able to identify the individual ligand binding residues. ConCavity's algorithm directly integrates the estimation of evolutionary sequence conservation with structure-based surface pocket prediction in the three-step process (Capra *et al.* 2009). In the first step, ConCavity integrates the conservation of surface residues with structural attributes. The structure-based binding site identification uses some existing algorithms e.g. LIGSITE, SURFNET or PocketFinder. Evolutionary conservation scoring is made by using Jensen-Shannon divergence (JSD) (Capra and Singh 2007) method. In the second step, ConCavity extracts potential pockets by searching for the grid threshold such that pockets have reasonable shapes. In the third step, method blurs each pocket grid values and scores every protein residue with an estimate of how likely it is to bind to a ligand based on overlapping value to each residue. Capra and co-workers have compared ConCavity's ability of binding site identification against LIGSITE, SURFNET, and PocketFinder and the comparison indicates that ConCavity's ability of binding site identification is better with both APO (unbound) and HOLO (bound) structures. ConCavity's data, source code, and prediction visualizations are freely available on the web site: http://compbio.cs.princeton.edu/concavity/

2.5.10 DEPTH

In 2011, Pern Tan and co-workers presented a binding site prediction method named DEPTH, which runs on a web server, to calculate the depth of the residues and to identify the caves. In addition, it is able to predict pK_a values (Pern Tan *et al.* 2011). The program assumes that the most likely binding site is the one having the maximum depth and the maximum 'solvent accessible surface area' (SASA) (Pern Tan *et al.* 2011). The server input-format is PDB format and there is an option to adjust the values of four parameters associated with the computation of residue depth and the prediction of binding cavities. Depths of residues are computed using the pre-equilibrated box of SPC216 model water (Berendsen *et al.* 1981 and 1987). Then method removes waters in cavities and all those waters that clash with atoms of protein. The accurately estimating of depth was made by repeatedly solvating the protein, which mimics bulk solvent dynamics. The solvation is repeated sufficient number (default is

25 repeats) and each time in a different orientation when water molecules can explore all regions accessible to bulk solvent water. The depth is reported as the average depth over all solvation iterations. DEPTH uses the 'rolling-ball' algorithm to compute SASA of the residues (Shrake and Rupley 1973) and the accessibility of each residue is normalized against theoretically calculated values of accessible surface area for an extended conformation of an Ala-X-Ala tripeptide (Hubbard and Blundell 1987). DETPH is benchmarked against LIGSITE, PocketFinder, SURFNET, and ConCavity with the dataset of 225 HOLO structures of single and multi-chain proteins, which are taken from LigASite v7.0 (Dessailly *et al.* 2008). The success rates of all other methods are almost the same, except ConCavity that uses the combined evolutionary-based and geometry-based methods for the binding site identification. ConCavity is superior in comparison with the other. DEPTH is freely available on the web site: http://mspc.bii.a-star.edu.sg/tankp/run_depth.html

2.5.11 Method by Gu *et al.*

Gu and co-workers (2012) developed method where the binding site prediction is based on the size and the amino acid composition of pocket. First, the method defines possible pockets using Pocket method (Levitt and Banaszak 1992) and retains for the following analysis those pockets with a volume more than 16 $Å^3$ (volume of 1,5 water molecules) (Gu *et al.* 2012). Then method computes the amino acid preference using an atom-based method (Qui and Wang 2009), which is developed by the same researcher group (not presented in this paper) (Gu *et al.* 2012). The method is analyzing only amino acids, which are located on the binding surface and the size of accessible surface area (ASA) of them is more significant than their number. These so called hotspot regions are identified and retained for the following analysis if the size of the regions is higher than 100 $Å^2$. The method calculates 'Score' value for each hotspot, which describes how well and how much the active groups of the amino acids are accessible onto these hotspot regions. Because the larger hotspot region is able to containing more potential amino acid residues than the smaller region and so it is able to get much higher 'Score', the method adjusts scores by multiplying them with the common logarithm (Log_{10}) of size of hotspot region when formed

'FinalScore'. This enables better comparison between the hotspot regions and finally, the method sorts these pockets in decreasing order according to theirs 'FinalScore'.

Gu and co-workers have been validating their method with two test sets, which were selected from published studies. Data set I is the collection of 100 non-redundant protein-ligand complexes made by Nayal and Honig (2006) and the Dataset II is 35 unbound/bound structures data set developed by Laurie and Jackson (2005). The prediction results with Dataset I are compared with the prediction results, which are based only on amino acid composition or pocket size and the comparison proved that this combination method can provide 10 % better success rate than either of size based or amino acid composition based methods alone. Method by Gu *et al.* is also compared with Q-SiteFinder, SCREEN, and method by Morita *et al.* with Dataset II. Comparison proved that according to the rules of top one and top three, this method achieved the same accuracy as Morita's method and SCREEN, and better than Q-SiteFinder.

3 Test sets

Developers have benchmarked ability of their methodologies to identify binding sites with different protein-ligand complexes, for which 3D complex structures are known. Tests are usually made with 20 to 50 ligand-protein complexes, but some developers have reported tests with an up to more than 11,000 binding sites collected from APO structures from the PDB (An *et al.* 2004, 2005).

Initially, benchmarkings were made with ligand bound structures, but later it was realized to give too optimistic results. The structure of flexible protein may change in the context of the binding process – even so much so that the binding site is not clearly visible before the structure of the protein has changed during the binding process. Currently researches are well conscious of that possibility (e.g. Nicholsan *et al.* 1995, Shao *et al.* 1997, Ishima *et al.* 1999, Freedberg *et al.* 2002, Katoh *et al.* 2003, Tóth and Borics 2006b). Because the main goal is to predict a potential binding site from an uncomplexed structure, it is critical to compile a benchmark of unliganded pocket sites. It can be said that the binding site

prediction methods, which are not capable to detect binding sites from unbound state of protein, are as a matter of fact useless. Therefore the use of unbound (APO) structures for benchmarking is of primary importance. In the following chapters are presented some most interesting dataset, which have been used for benchmarking purposes.

3.1 Dataset of 20 unbound/bound structures by Brady and Stouten

Brady and Stouten (2000) were among the first to use APO structures for benchmarking, so that they would be able to obtain a much more realistic picture of their program's (PASS) ability to identify the binding site. Their test set includes 30 HOLO structures, of which 20 have also APO structures and all structures are taken from PDB.

3.2 dSM dataset

Del Sol Mesa and co-workers (2003) developed a new dataset for testing their own method. This dataset, which later was called dSM dataset (Nimrod *et al.* 2008), contains 112 non-redundant protein families with annotations for functionally important sites (SITE records in PDB) (del Sol Mesa *et al.* 2003). dSM dataset was used to benchmarking with Sequence Space, variation of ET method by Landgraf *et al.*, method by Hannenhallli and Russell (2000), and the different combinations of these three (del Sol Mesa *et al.* 2003). Nimrod and co-workers (2008) have also used dSM dataset for benchmarking PatchFinder (version 2005 and 2008) with HotPatch, siteFinder|3D, and ET viewer (Morgan *et al.* 2006).

3.3 Dataset of Perola *et al.*

Perola and co-workers (2004) collected dataset of 99 nonredundant, comprehensive protein-ligand complexes from the PDB. This dataset is designed for docking methods testing, but it is also used for testing of binding sites identification with SCREEN, CASTp, and the method by Gu *et al.* (Nayal and Honig 2006, Gu *et al.* 2012).

3.4 Dataset of 35 unbound/bound structures by Laurie and Jackson

Laurie and Jackson (2005) created a dataset of 35 structurally distinct proteins in the unbound state, which share structural similarity with 35 proteins in the ligand-bound dataset. They collected dataset by examination of the Structural Classification Of Proteins (SCOP) database (Murzin *et al.* 1995) for the 305 proteins described by Nissink *et al.* (2002). That dataset was used to benchmark Q-SiteFinder, Pocket-Finder, SCREEN, Method by Morita *et al.*, and method by Gu *et al.* (Laurie and Jackson 2005, Morita *et al.* 2008, Gu *et al.* 2012). Results of comparison are presented in Table 2.

Table 2 Comparison of 5 methods with dataset of 35 unbound/bound structures.

	Method	**Top 1**		**Top 3**	
		Unbound (%)	Bound (%)	Unbound (%)	Bound (%)
1.	Method by Morita *et al. (a)*	77	80	86	100
2.	Method by Gu *et al.* (b)	74	-	86	-
3.	SCREEN (b)	71	-	86	-
4.	Q-SiteFinder (a)	51	74	83	94
5.	Pocket-Finder (a)	51	72	66	77

(a) The data for Method by Morita *et al.*, Q-SiteFinder, and Pocket-Finder are taken from Morita *et al.* 2008
(b) The data for Method by Gu *et al.* and SCREEN are taken from Gu *et al.* 2012

3.5 Dataset of 48 unbound/bound structures by Huang and Schroeder

In 2006, Huang and Schroeder unified datasets of 20 unbound/bound and 35 unbound/bound structures together. These datasets have five same proteins and one structure was ignored since no ligand was found in the PDB entry of that time. There remained 48 unbound/bound structures, which used to Benchmarking LIGSITE^{CS} and LIGSITE^{CSC}. Huang and Schroeder used dataset also to compare LIGSITE, LIGSITE^{CS}, LIGSITE^{CSC}, CAST, PASS, and SURFNET methods. This dataset is most widely used for comparing methods (Weisell *et al.* 2007, Kalidas and Chandra 2008, Li *et al.* 2008, Huang 2009, Le Guilloux *et al.*

2009, Tripathi and Kellogg 2010, Volkamer *et al.* 2010, Yu *et al.* 2010, Zhang *et al.* 2011, Zhu and Pisabarro 2011). Results of citated studies are presented in Table 3.

Table 3 Comparison of 20 methods with dataset of 48 unbound/bound structures.

	Method	Top 1		Top 3	
		Unbound (%)	Bound (%)	Unbound (%)	Bound (%)
1.	VICE (c)	83	85	90	94
2.	MetaPocket 2.0 (k)	80	85	94	96
3.	MSPocket (a)	75	77	92	90
4.	MetaPocket 1.0 (j)	75	83	90	96
5.	POCASA (b)	75 (73)	77 (77)	88 (85)	94 (94)
6.	Lsite (LigSite) (g)	75	75	85	88
7.	DoGSite (g)	71	83	92	92
8.	LIGSITE^{CSC} (e)	71	79	-	-
9.	VisGrid (i)	71	67	85	79
10.	PocketPicker (f)	69	72	85	85
11.	Fpocket (d)	69	83	94	92
12.	Dsite (DrugSite) (g)	65	69	77	79
13.	LIGSITE^{CS} (e)(j)	60 (71)	69 (81)	77 (85)	87 (92)
14.	PASS (e)(i)(j)	60 (56) (58)	63 (67) (58)	71 (71) (75)	81 (88) (85)
15.	LIGSITE (e)(i)	58 (75)	69 (83)	75 (79)	87 (92)
16.	CAST (e)(i)	58 (64)	67 (66)	75 (77)	83 (79)
17.	PocketDepth (h)	53	-	87	-
18.	Q-SiteFinder (j)	52	75	75	90
19.	SURFNET (e)(i)(j)	52 (40) (42)	54 (48) (42)	75 (60) (62)	78 (71) (60)
20.	MolSite (l)	-	88 (98)	-	-

(a) The data for MSPocket are taken from Zhu and Pisabarro (2011)
(b) POCASA results are taken from Yu et al. (2010) and are calculated using Dc and other are calculated using Mc. Dc = Depth center Mc = Mass center
(c) The data of VICE are taken from Tripathi and Kellogg (2010) and are calculated using the 'center-of-gravity' of pockets.
(d) Fpocket results are taken from Le Guilloux *et al.* (2009).
(e) The data of LIGSITEcs, CAST, PASS and SURFNET were first reported by Huang *et al.* (2005)
(f) The data of PocketPicker were first reported by Weisel *et al.* (2007)
(g) The data of DoGSite, LSite, and DSite are taken from Volkamer *et al.* (2010)
(h) The data of PocketDepth are taken from Kalidas and Chandra (2008)
(i) The data of VisGrid and (LIGSITE, CAST, PASS and SURFNET are remaked) from Li *et al.* (2008)
(j) The data of MetaPocket and (Q-SiteFinder, SURFNET, PASS and LIGSITEcs) from Huang (2009)
(k) The data of MetaPocket 2.0 from Zhang *et al.* (2011)
(l) The data of MolSite from Fukunishi and Nakamura (2011). Top1 bound structres using D_C and other are calculated using pocket Dmin = the minimum distance between the center of the predicted pocket and any atom of the bound ligand.

3.6 210 complexes from Protein Ligand Database (PLD)

Huang and Schroeder (2006) created also another dataset named 210 PLD containing 210 bound structures that are collected from Protein Ligand Database (PLD, version 1.3). Version 1.3 of PLD contained 485 protein-ligand complexes, all available in the PDB. Huang and Schroeder used this dataset to compare LIGSITE^{CSC}, LIGSITE^{CS}, LIGSITE, PASS, and SURFNET with each other. This dataset was also used against MetaPocket, MetaPocket 2.0, Q-SiteFinder, GHECOM, ConCavity, Fpocket, POCASA, method by Dai *et al.*, PASS, and PocketPicker (Huang 2009, Dai *et al. 2011,* Zhang *et al.* 2011).

3.7 Dataset of 98 unbound/bound structures

Gunasekaran and Nussinov (2007) constructed dataset of 98 unbound/bound protein-ligand complexes for estimating how different structurally flexible and rigid binding sites are. They get structures from PDB and divided dataset into three classes (Gunasekaran and Nussinov 2007). In the class I there have 41 proteins, which have not conformational change upon ligand binding. The class II contains 35 proteins, which have a moderate conformational change (greater than or equal to 0,5 Å but less than or equal to 2,0 Å). In the class III have 22 proteins, which have a large conformational change (greater 2,0 Å) initiated by ligand binding. This dataset, especially unbound structures in classes II and III, provide a good circumstance for testing and estimating the binding site detection ability. All in all, this dataset is very similar to that of 48 unbound/bound but twice as much extensive. AutoLigand is benchmarked with this dataset, but only 96 unbound/bound structures and the results proved robustness of AutoLigand (Table 4) (Harris *et al.* 2008).

Table 4 Comparison of AutoLigand with 96 unbound/bound structures in each class. Success rate, (number of correct predicted structures / number of all structures in class).

Method		Class I	Class II	Class III
AutoLigand	Unbound	83 % (33/40)	86 % (30/35)	67 % (14/21)
	Bound	90 % (36/40)	89 % (31/35)	95 % (20/21)

3.8 Astex Diverse Set

Astex Diverse Set is collection of 85 protein-ligand complexes, which have been specially prepared in a format suitable for docking by Hartshorn *et al.* (2010). This dataset does not include the unbound structures. Astex Diverse Set was used to comparisons with Blind Docking, Focused Docking, FINDSITE, FINDSITE^{LHM}, PocketPicker, and Fpocket (Bryliński and Skolnick 2009, Ghersi and Sanchez 2009a, Le Guilloux *et al.* 2009).

3.9 Datasets by Fukunishi and Nakamura (A, B, C, and D)

Fukunishi and Nakamura (2011) compiled dataset for benchmarking their MolSite method. Their dataset contains 89 known protein-ligand structures and is divided to four groups. Each of group is collected in the emphasis on certain properties and groups can contained some same structures, but total number of structures is 89. Dataset A is same than bound structures of 48 unbound/bound datasets and it is for ligand-binding bite prediction and Binding Free Energy Estimation. Dataset B contains 50 structures for binding free energy estimation and Dataset C is so called small set and it contains 18 different structures. Dataset D contains 16 structures for ligand binding site prediction using only one ligand included in the bound complex crystal. Fukunishi and Nakamura have used Dataset A set to compare prediction accuracy of MolSite with LIGSITE^{CS}, PASS, FINDSITE, MetaPocket, Q-SiteFinder, and SURFNET (Bryliński and Skolnick 2008, Huang 2009, Fukunishi and Nakamura 2011).

3.10 198 drug-target complexes (DT198)

Zhang and co-workers collected a new dataset of 198 drug-target (called this paper to DT198) complexes for comparing MetaPocket 2.0 and other methods with the real drug binding sites. These real drug binding sites are derived from the DrugPort database (http://www.ebi.ac.uk/thornton-srv/databases/drugport/), which is based on DrugBank database (Wishart *et al.* 2006, 2008, Knox *et al.* 2011) and only those structures were accepted that were found also in PDB and contained both protein target and ligand. Zhang

and co-workers selected only one complex structure for each drug-target pair and they only kept the single chain where ligands bind. Zhang and co-workers used this dataset for benchmarking MetaPocket 2.0 with LIGSITE^{CS}, ConCavity, POCASA, Q-SiteFinder, GHECOM, PASS, Fpocket, and SURFNET (Zhang *et al.* 2011) and results are presented in Table 5.

Table 5 Comparison of 9 methods with dataset of 198 drug-target complexes.

Method	Top 1	Top 2	Top 3
MPK2	61	70	74
LIGSITE^{CS}	48	57	61
ConCavity	47	53	56
POCASA	43	54	56
Q-SiteFinder	40	54	62
GHECOM	39	51	56
PASS	35	50	56
Fpocket	31	48	57
SURFNET	24	30	34

4 Recapitulation

The drug targets locations and the physicochemical properties of binding sites are the most important knowledge, which creates the good basis for drug development process. For this reason, the binding site prediction has become a common practice for the identification of unknown proteins binding sites. The first identification methods developed in the 80's, but not until 2000's these have been systematically developed (Table 1). Different identification techniques have been developed tremendously and many interesting approaches have been tried with the varying degrees of success. However, none of these methods can offer a complete assurance to the correctness of predictions, but some methods have achieved higher than 80 % success rate with the Top1 predictions and higher than 90 % success rate with the Top3 predictions (Table 3).

Validation of the methods is significant for estimating the true prediction ability. The best way for estimating the goodness of a method is to use 3D-structures of the unbound state of

selected proteins, because the conformational change can be large in the case of flexible proteins. In these cases, structures of bound state do not correspond to the real circumstance. The effect of unbound forms may well be seen in Table 4. When the difference of AutoLigandin's ability to predict true binding sites between the bound and unbound structures is 7 percentage points with class 1 (no conformational change), the difference is four times bigger, 28 percentage points with class 3 structures (large conformational change). This proves the importance of the unbound structures for estimating the prediction ability of the methods.

For this reason, several datasets has been developed, among which the dataset of 48 unbound/bound structures collected by Huang and Schroeder (Table 3) is the most widely used. A couple of years later collected dataset by Gunasekaran and Nussinov includes 98 unbound/bound structures but in addition it includes classification based on the magnitude of protein conformational change upon binding (Table 4). The idea of the DT198 dataset is very good (Table 5), because these structures are true drug-targets, which offer realistic experiment circumstances. Unfortunately, the unbound structures are missing. If the DT198 dataset would also include the unbound structures, it would be one of the best datasets with the datasets of 48 and 98 unbound/bound structures for estimating the binding site prediction ability.

The dataset of 48 unbound/bound structures offers a great opportunity for comparing 20 different methods with each other (Table 3). In addition, four other methods can be included in the assessment because the dataset of 48 unbound/bound structures contains all of the structures from the data set of 35 unbound/bound structures (Table 2, Huang and Schroeder 2006) and Q-SiteFinder has been evaluated with both datasets. Since Q-SiteFinder may be the roughly same success rate with both datasets, it can be assumed that the success rate of other methods would also be equal with the dataset of 48 unbound/bound structures.

VICE and MPK2 showed the powerful binding site prediction ability (Table 3). While VICE have only one effectual algorithm, MPK2 combines 8 different types open source methods to one effective and freely available tool. MPK2 has also the best success rate with 198 drug-target dataset (Table 5). MolSite seems to give a good success rate (Table 3) - up to 98 %, but unfortunately the evaluation has been made only with bound forms, so the true

performance may be something else if APO-structures are used as criteria. AutoLigand also appears to be a good method and it has a really interesting approach to prediction. Unfortunately AutoLigand has not validated against any other methods. The success rates of methods by Morita *et al.* and Gu *et al.* seem to be the same degree as MSPocket, MPK1, POCASA, and Lsite. SCREEN seems to be the same degree as DoGSite, LIGSITE^{CSC}, and VisGrid.

Although the highly sophisticated algorithm of VICE has shown its strength for the predicting of binding sites, it can generally be concluded that combination of several types prediction techniques brings clearly better success rate (Tables 3 and 5). The evolutionary-based methods may provide additional strength for the combined approaches in the case of small molecule ligand bound protein, but alone those methods failed in most of the cases with small molecules (del Sol Mesa *et al.* 2003, Pettit *et al.* 2007). On the contrary, the evolutionary-based methods have shown the robustness ability for predicting the binding sites with protein-protein and protein-DNA complexes, outperforming both geometry and energy-based methods (Lichtarge *et al.* 1996, Alloy *et al.* 2001, Lichtarge and Sowa 2002, del Sol Mesa *et al.* 2003, Pettit *et al.* 2007).

Blind docking methods provide a reasonably good success rate, but compared with the time spent and the fact that the docked ligand should be known a priori, those can be recommended only use in exceptional cases (e.g. some true ligand is already know). Focused docking can make prediction much faster than original BD, but it may lose the correct binding site before the docking. This problem of FD method could be solved by focus the docking in the pre-defined molecular path areas (Lindow *et al.* 2011) instead of the pre-predicted pockets. Thus the docking area would be more comprehensive, but not the entire surface of the target protein.

The long scale MD simulations are computationally highly expensive (especially with large complexes) and the results of predictions may not be as unambiguous as with the other methods. In addition, this approach requires a known ligand or some estimated model of that, which limit its uses. Therefore, it does not make sense to generally use this method, but with some special cases it can be irreplaceable useful. One such special case could be very flexible target proteins. Another special case could be the identification of allosteric binding sites of proteins, because Shan and co-workers (2011) have noticed that the PP1

repeatedly grabbed with several previously known Src kinase allosteric sites during the long scale MD simulations. This method has also obtained good results for predicting the critical water placement of the pocket (Shan *et al.* 2011). The technique, where accelerated molecular dynamics (aMD) (Voter 1997, Hammelberg *et al.* 2004) was combined with the inherent power of graphics processor units (GPU), could significantly reduce the long scale MD required computation time (Pierce *et al.* 2012), allowing more widespread use of this method.

In summary, from these 77 methods can be found in a number of interesting approaches to predict the binding site. These methods offer a good tool palette for a variety prediction need. From the perspective of the small molecules binding site prediction the most interesting methods could bring up MetaPocket 2.0, VICE, AutoLigand, ConCavity, and MolSite. Hopefully, all these methods are being improved and new ones also developed further, since none of them is able to provide 100 % success rate. The continuous growth of the computational power will be offering a better chance for using long scale MD simulations and other computationally heavy technologies. In the future, the methods should be benchmarked with good datasets in order to improve the comparability between different methods. For this purpose the datasets of 48 and 98 unbound/bound structures and DT198 can be recommended.

Classification of binding cavities: What are the differences between good and not so good binding pockets?

5 Introduction

The purpose of this study was to determine why some molecules are better binders than others. Is it possible to find any connective single factor or multiple factors, which would explain the difference in affinities of binding molecules? In this context, it can be assumed that the binding affinity depends not only on the properties of the bound molecule. So, the other big question is: What are the differences between good and not so good binding pockets?

The aim of study was primarily to draw attention to the size of molecules and on the other hand assumption that the binding capacity of molecules do not improve significantly after a certain molecular size (Kuntz *et al.* 1999). Although the size of the molecule plays an important role in terms of binding, it cannot alone explain the large variation on the binding affinity. Lipophilicity of ligand and the ability to form hydrogen and ionic bonds also affect to binding affinity and this has been known for a number of years (Lipinski *et al.* 1997). In 1997 Christopher Lipinski and coworkers presented "rule of five" (RO5), which describes the properties of drug-like molecule, based on these aforementioned factors (Lipinski *et al.* 1997). Although it is well known that the RO5 is not an all-inclusive guideline, it has remained for years a cornerstone of orally bioavailable drug development (Keller *et al.* 2006, Zhang and Wilkinson 2007). Since then the rules have also spawned many extensions (Ghose *et al.* 1999, Congreve *et al.* 2003).

There are still many factors like a number of rotatable bonds and a number of ligand's tautomer, unconventional dihydrogen bonds and the C-H···π interactions that can play a highly important role for the formation of the binding affinity. So that the issue would not be too simple, it should be noted that the ligand-protein complex formation is not only depending on the properties of the ligand, but also on properties of the target protein. This

study aims to identify possible trends that can be found in between well- and poorly-bound molecules.

6 Material and methods

6.1 Databases

In this study, it was decided to utilize some the already existing binding affinity database, which can be found on the Internet. When selecting the database a special attention was drawn to its quality, which must be line with Wallach and Lillien's (2009) criteria for a good database. In addition, the perspective of study sets some additional requirements for database and its format. For this study the desired features of the database is comparable values of binding affinity, the ability to download the crystallographic structure of protein-ligand complexes, and the molecular weight (M_w) of the ligand or a structure data, from which could easily calculated M_w.

Table 6 Freely available binding databases.

	Database:	Internet address:	Last update:	Binding Affinity:	Approx. Number of structures:
1.	The PDBbind Database	http://sw16.im.med.umich.edu/databases/pdbbind/index.jsp	06/2007	Yes	3214
2.	Binding MOAD	http://www.bindingmoad.org/	2011	Yes	# 18764 Protein-Ligand Structures # 6311 Structures with Binding Data # 9048 Different Ligands
3.	PDBselect	http://bioinfo.tg.fh-giessen.de/pdbselect/	02/2011		
4.	SitesBase	http://www.modelling.leeds.ac.uk/sb/	08/2006		
5.	The Binding Database	http://www.bindingdb.org/bind/index.jsp	02/2011	Yes	# 910836 Binding data # 5630 protein targets # 378980 small molecules
6.	AffinDB	http://www.agklebe.de/affinity	?		748
7.	Het-PDB Navi	http://hetpdbnavi.nagahama-i-bio.ac.jp/	4.4.2011		72104
8.	Protein Ligand Interaction Database	http://203.199.182.73/gnsmmg/databases/plid/	?		25844
9.	LigBase	http://modbase.compbio.ucsf.edu/ligbase/	?		
10.	RSCB PDB (Protein Data Bank)	http://www.rcsb.org/pdb/home/home.do	weekly	Often	89393
11.	PSMDB (The Protein - Small-Molecule DataBase)	http://compbio.cs.toronto.edu/psmdb/	12/2010	Yes	
12.	ChEMBL	https://www.ebi.ac.uk/chembl/	?		758082

In this study a dozen molecular affinity databases (Table 6), which freely available via Internet, were analyzed in more details. Each of these databases is compiled with emphasis on certain features that creators have seen the most useful on their own terms. Protein Data Bank is oldest, created 1971 (Bernstein *et al.* 1977), and one of the biggest of those and is probably a benchmark for all others. Different databases working also together and share data for each other. Binding MOAD (Hu *et al.* 2005) was the most appropriate database for this study. The crystallographic structures of protein-ligand complexes were applied from the Protein Data Bank (RSCB PDB).

6.1.1 Binding MOAD

Binding MOAD (Mother of All Databases) has been developed at University of Michigan and the aim of developers was to make Binding MOAD the largest possible collection of high-quality protein–ligand complexes, which also contained the binding data (Hu *et al.* 2005, Benson *et al.* 2008). The focus of Binding MOAD is on protein binding sites and protein flexibility. The researcher group of Binding MOAD is co-operating with the researcher group of the PDBbind database. Both groups are sharing collected binding data with each other, which allows double check all of the data. This kind of quality control is unique and unusually high level for datasets of this size.

Binding MOAD (Mother of All Databases) contains at the time of writing (May 2013) 18764 complexes with binding data, which are collected from various published studies. Each entry has a resolution better than 2,5 Å and contains valid ligand, PDB id, enzyme classification (EC) number, binding-affinity data (IC_{50}, K_i, K_d or K_a), and SMILES code of ligand structure. All binding data and complexes are available and downloadable with a comma-separated values (CSV) file at the Binding MOAD website, www.bindingmoad.org. In this study has used only those entries that contain K_i affinity values have been used.

6.1.2 PDB

Protein Data Bank is the largest online archive of structural data of biological macromolecules such as proteins, viruses, DNA and RNA fragments (Berman *et al.* 2000, Furnham *et al.* 2012). Brookhaven National Laboratories (BNL) created Protein Data Bank (PDB) in 1971 (Bernstein *et al.* 1977), when there were 7 structures (Berman *et al.* 2000). The database was growing slowly as 1977 there were only 47 structures in PDB (Bernstein *et al.* 1977). Since October 1989, the Research Collaboratory for Structural Bioinformatics (RCSB) has been responsible for the management of PDB (Berman *et al.* 2000). At the moment (May 2013) PDB includes over 90000 protein structures and it grows on average with 20 new structures per day. PDB is updated on a weekly basis.

Today, PDB is a basic tool macromolecule research, as PDB combines the biological and structural information from different sources. Each PBD entry contains the information e.g. the EC number, the binding-affinity data (if available), the biological function (if known), the experimental details of 3D-structure, possible bounded ligands, the gene names, and the sequences data. With the latest extension of PDB, in each entry was added data from UniprotKB (The UniProt Knowledgebase, Magrene *et al.* 2011), MSA by Pfam (Punta *et al.* 2012), calculated hydropathy of the residue by BioJava (Prlić *et al.* 2012), the protein predicted disorder by JRONN (Yang *et al.* 2005a), to the sequence combined ranges of Homology Models from SBKB (http://www.sbkb.org) and the Protein Model Portal (Arnold *et al.* 2009). One of the future aims is to discover and collect the functions of all proteins (Furnham *et al.* 2012).

6.2 Molecular modeling

6.2.1 Molecular Operating Environment (MOE)

The Molecular Operating Environment (MOE, v. 2012.10) is fully integrated drug discovery software package, which is developed by Chemical Computing Group Inc. MOE can use to structure- and fragment-based drug design, pharmacophore discovery, protein and antibody modeling, docking, and the molecular modeling and simulations. It also contains many

medicinal chemistry and biologics applications and it able to build QSAR and QSPR models. Molecules structures can be imported and exported in the most common file types. MOE is also very the powerful and flexible software to make high quality presentation pictures.

In this study, MOE have been used to the minimization of the complexes, the visual examination, creating the ligand-interaction maps of the interactions between the ligand and the protein, and to make the presentation pictures. MOE was also used to calculate variety measurement parameters of the ligand properties, as well as interactions between the ligand and the protein.

6.2.2 Maestro and SiteMap

Schrödinger's Maestro (version 9.2.112) is a powerful, all-purpose molecular modeling environment. Maestro is linchpin of Schrödinger's computational technology and almost all Schrödinger's software, which is developed over 25 different, is used via Maestro. Maestro offers the flexible visualization, 3D realism, model generation, 2D ligand interaction diagrams, and the ability to customize scripts. Today, Maestro is one most used molecular modeling environment.

Schrödinger's SiteMap program (version 2.5) is a promising tool for the analysis and discovery of the binding sites, as well as lead optimization and virtual-hit assessment (Halgren 2009). The best feature of the program is its speed. One-average protein-ligand complex (5000 atoms) calculation requires about 2-3 minutes, when using a single CPU of a 2,4 GHz Intel Pentium 4 workstation.

Sitemap is not very widely used, at least not in published studies, but Vidler and co-workers (2012) have used SiteMap to assess the druggability of diverse members of the bromodomain family druggability. It is interesting to know how good the results of this program can be reached and that's why we decided to compare how well the SiteMap can estimate the druggability. In this study was also estimated whether SiteMap capable to predict the binding affinity. If the results are good enough, the program would fit well to be used routinely in drug-discovery studies.

6.3 Other used software

Microsoft Excel® (Microsoft Office Excel 2007, Microsoft® Excel® for Mac 2011 v. 14.0.0) was used to collecting and organization of data. The automated trimming and modification of data was made by Visual Basic (Microsoft Visual Basic v. 14.0 Mac/ Microsoft Visual Basic 6.5 version 1053). The SMILES strings conversion to M_w was made in Excel with ChemDraw add-in (Chem & Bio Draw 12.0 / Excel).

6.4 Methods

6.4.1 Pre-processing of the data

Binding MOAD affinity database (Hu *et al.* 2005) was the most appropriate database for this study. It included the data of measured binding affinities, SMILES strings, and PDB entries, and covalently attached molecules not considered valid ligands, which is important for this study. Only those data that contained binding data was downloaded from the server (http://www.bindingmoad.org/moad/downloadMoad.do) 18.4.2011 in CSV file format. The CSV file is imported in Microsoft Excel® (Microsoft Office Excel 2007, Microsoft® Excel® for Mac 2011 v. 14.0.0) so that each column contained only one kind of information and a new file was created. In that file contains 4851 valid protein-ligand complexes with affinity data and each of these are in its own row. If the protein has multiple valid ligands, each of them has its own row.

The downloaded file contained affinity data in IC_{50}, K_i, K_d or K_a values. In this study only K_i values were used in order to calculate the free energy, if necessary. Accordingly those rows without K_i data removed. As K_i data was expressed in several formats all data points were unified using automated Visual Basic script (Microsoft Visual Basic v. 14.0 Mac/ Microsoft Visual Basic 6.5 version 1053). The entire scripts (DoItAll) are attached at the end of the thesis (Appendix I.). At the end there were 1826 protein-ligand complexes included in the data set.

At the next stage SMILES codes were converted to molecular weights (M_w) in Excel by ChemDraw add-in (Chem & Bio Draw 12.0 / Excel) and the numeric values of calculated Mw

were copied to other column, because only Excel with ChemDraw add-in is able to shown converted M_w. This copying was made by running VBA script that called to MWpaste (Appendix I). 31 complexes did not include SMILES strings and 29 complexes had corrupted or unsuitable SMILES string. At this stage there were 1766 complexes successfully converted. Most of complexes were molecular weight between 100-800 Da (Fig. 14).

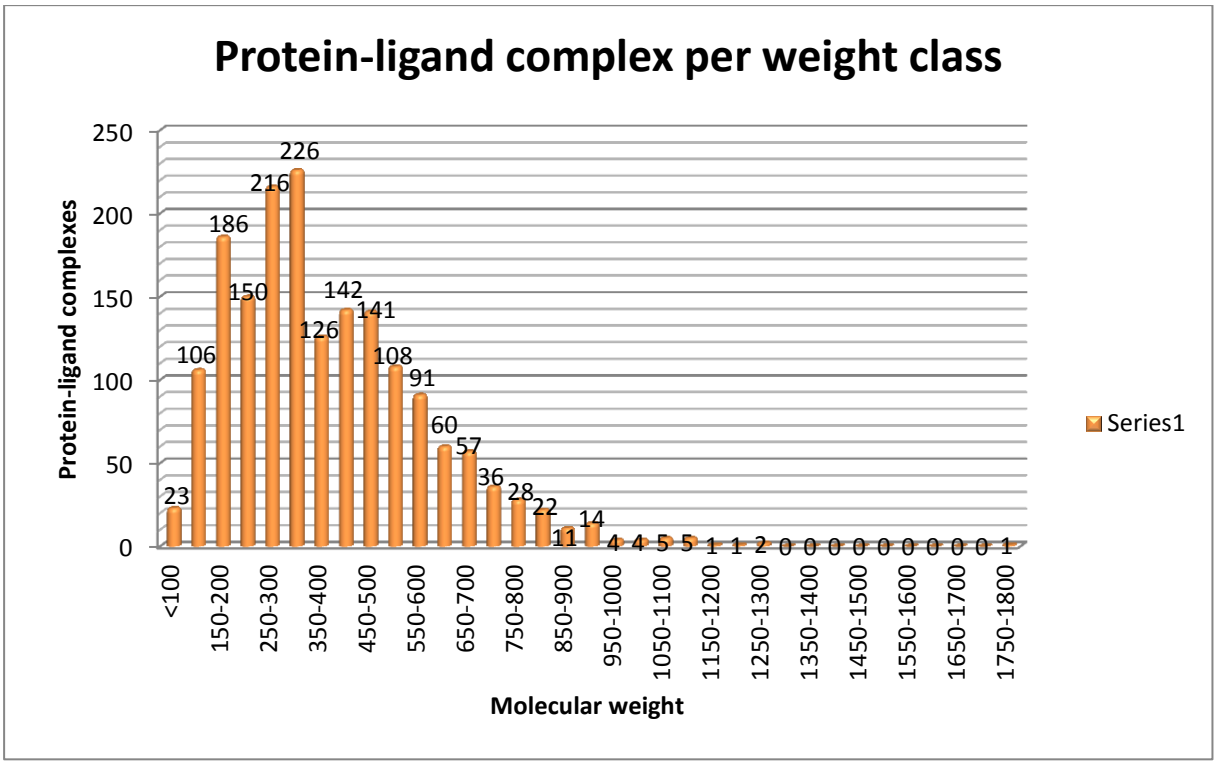

Figure 14 Number of the protein-ligand complexes in different molecular weight ranges.

This study is focused on only those protein-ligand complexes whose molecular weight of ligands is between 100-600 Da, because average affinity values do not increase beyond 600 Da (Fig. 15). 1492 protein-ligand complexes were in this weight range. In order not to be the too extensive study, the sampling was limited to approximately 100 protein-ligand complexes. The samples were randomly selected from the table, but in some cases were made knowingly the exception for randomness. In these cases, the same ligand was selected more than once if the values of binding affinities differed significantly. For this reason some weight ranges have more samples than average a random sample rate should be (Table 7). 127 complexes were chose for detailed analysis.

Table 7 Sample rate in different molecular weight ranges.

Weight range	Samples in chart	Theoretical sample rate	Theoretical Sample %	Chosen sample rate	Sample %
100-150	106	7	6,60	17	16,04
150-200	186	13	6,99	14	7,53
200-250	150	10	6,67	11	7,33
250-300	216	15	6,94	19	8,80
300-350	226	15	6,64	16	7,08
350-400	126	9	7,14	10	7,94
400-450	142	10	7,04	11	7,75
450-500	141	10	7,09	11	7,80
500-550	108	7	6,48	9	8,33
550-600	91	6	6,59	9	9,89
Summary =	1492	102		127	

6.4.2 Processing of structures data collecting

The crystal structures chosen for closer analysis were downloaded from PDB by MOE (version 2012.10). Each crystal structure was loaded one at a time and structure was examined manually in screen. All downloads have been carried out 21.11.2011-30.1.2012. At first, the crystal structure of each ligand was checked in order that they are accurate. If not, then the structure was repaired manually to corresponding the ligand structure in PDB. At the following step the protein-ligand complexes were minimized using MOE's tool LigX. LigX's settings were used:

PROTONATE:	***yes***	*Use Protonate 3D for Protonation*				
	yes	*Allow ASN/GLN/HIS "Flips" in protonate 3D*				
DELETE:	***yes***	*Water molecule farther than 4,5 Å from Ligand or Receptor*				
TETHER:	***yes***	*Receptor*	*Strength:* ***10***	*Buffer:* ***0,25***	***no***	*Hydrogen*
	yes	*Ligand*	*Strength:* ***10***	*Buffer:* ***0,25***	***no***	*Hydrogen*
	yes	*Solvent*	*Strength:* ***10***	*Buffer:* ***0,25***	***no***	*Hydrogen*
FIX:	***yes***	*Atoms farther than* ***8*** *Å from Ligand*				
	yes	*Hydrogen Close to Ligands will not be Fixed*				
REFINE:	***yes***	*Complex to an RMS Gradient of* ***0,1*** *kcal/mol/Å*				

The ligand interaction maps were created for each minimized complexes and saved to picture. The number of different interactions between the ligand and the protein can be manually calculated from this kind map. The following properties were especially recorded from interaction maps: number of hydrogen bonds, the number of unconventional hydrogen bonds and their types, the number of water bridged hydrogen bonds, ion-interactions and their types. Information on these interactions was recorded in an Excel spread sheet. Also the amounts of ligand's tautomers were recorded. The three-dimensional crystal structure of the complex and the molecular interactions between protein and ligand were also examined visually. All kinds of interesting observation on the interactions, the protein structure, the location of the ligand and the surrounding circumstances were also recorded in an Excel spreadsheet. Special emphasis was given to the type of interaction pocket, its type, occupancy in the sense of volume and whether it was lipophilic or hydrophilic. The table has one column, which was written noteworthy things in words, which were identified by visual inspection. MOE is able to calculate the variety computational measurement parameters of the ligand properties, as well as interactions between the ligand and the protein and those parameters added to the Excel spreadsheet.

The protein-ligand complexes that minimized by MOE were saved to Schrödinger Maestro file (.mae), to allow SiteMap program (version 2.5) to be run. All protein structures were minimized and checked with Schrödinger Maestro's (version 9.2.112) tool named PrepWiz (Schrödinger Suite 2011). Default settings were used. After that, every complex was analyzed using the Sitemap program and the default settings were used:

TASK: ***yes*** *Evaluate a single binding site region:*

Region about selected atoms plus ***6 Å*** *buffer will be examined*

Select non-receptor atoms defining region to evaluate

ASL: ***mol. n 3***

yes *Pick Molecule* ***yes*** *show markers*

SETTINGS:

Require at least ***15*** *site points per reported site*

Use ***more restrictive*** *definition of hydrophobicity*

Use ***standard*** *grid*

Crop site maps at ***4 Å*** *from nearest site point*

Use the ***OPLS_2005*** *force field*

In five cases program was unable to create a site map correctly. In these cases, SiteMap's estimated binding site was not at the same place as ligand and the distance between results and a real binding place might be greater than 10 Å. These five results were ignored. Obtained results were recorded into a new database and this database was added to previous Excel spreadsheet.

7 Results and Conclusion

7.1 Binding Affinity vs. Molecular weight

The assumption was that the ligand binding affinity value does not increase significantly longer after a certain molecular weight (idea originally proposed by Kuntz *et al.* 1999). As Figure 15 shows, although the binding affinity of molecules is greatly varied, there is a clear plateau on binding affinity when the molecular weight is above 600 Da.

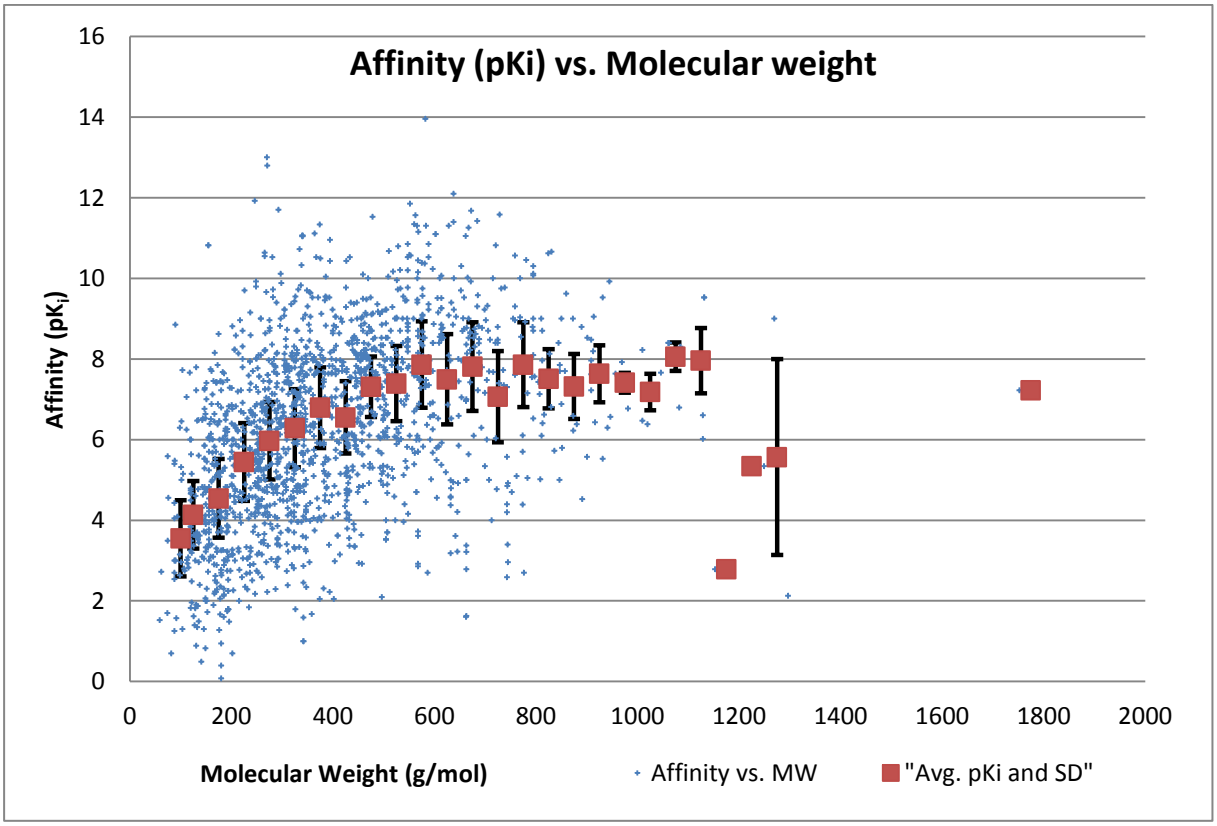

Figure 15 The relationship between affinity (pK_i) values and molecular weights of all 1766 suitable complexes from Binding MOAD.

This result was in line with expectations, but reasons for this behavior are not clear. Ligand binding to the protein never depends on only the properties of the ligand, but also on ensuring properties of the target protein.

7.2 LogP

In this study, the logP values are described by using computational SlogP (Wildman and Crippen 1999) values and later in this study with the logP value are meant to this computational SlogP value. When the relationship between logP values of the sample complexes and the binding affinity are reviewed it can be seen to indicate growing lipophilicity trend together with affinity increase (Fig. 16). Similar results have been reported previously in various studies (Carlson *et al.* 2008). The correlation is poor, but when the best and worst binders of each weight group are contemplated this tendency is easier to observe.

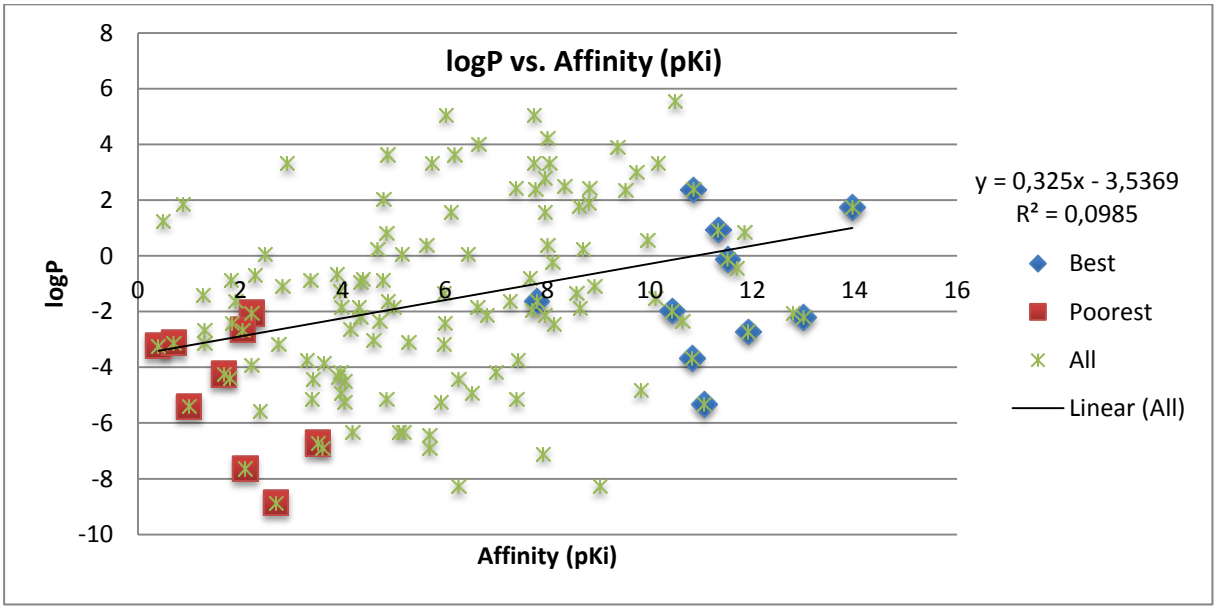

Figure 16 The relationship between logP and affinity (pK_i).

If the relationship between logP and binding affinity in different molecular weight classes is analyzed, an interesting phenomenon is seen (Fig. 17). It seems that the molecules with low molecular weight do not have a clear growing trend in terms of binding affinity but when the molecular weight increases, the tendency is stronger. The correlation is nil, but the tendency

is obvious. This could be due to the fact that larger molecules can more easily get the extreme values. In addition, it should be noted that MOE's SlogP is theoretically calculated value, while the M_w and the binding affinity of are measured values. So these results might contain some errors.

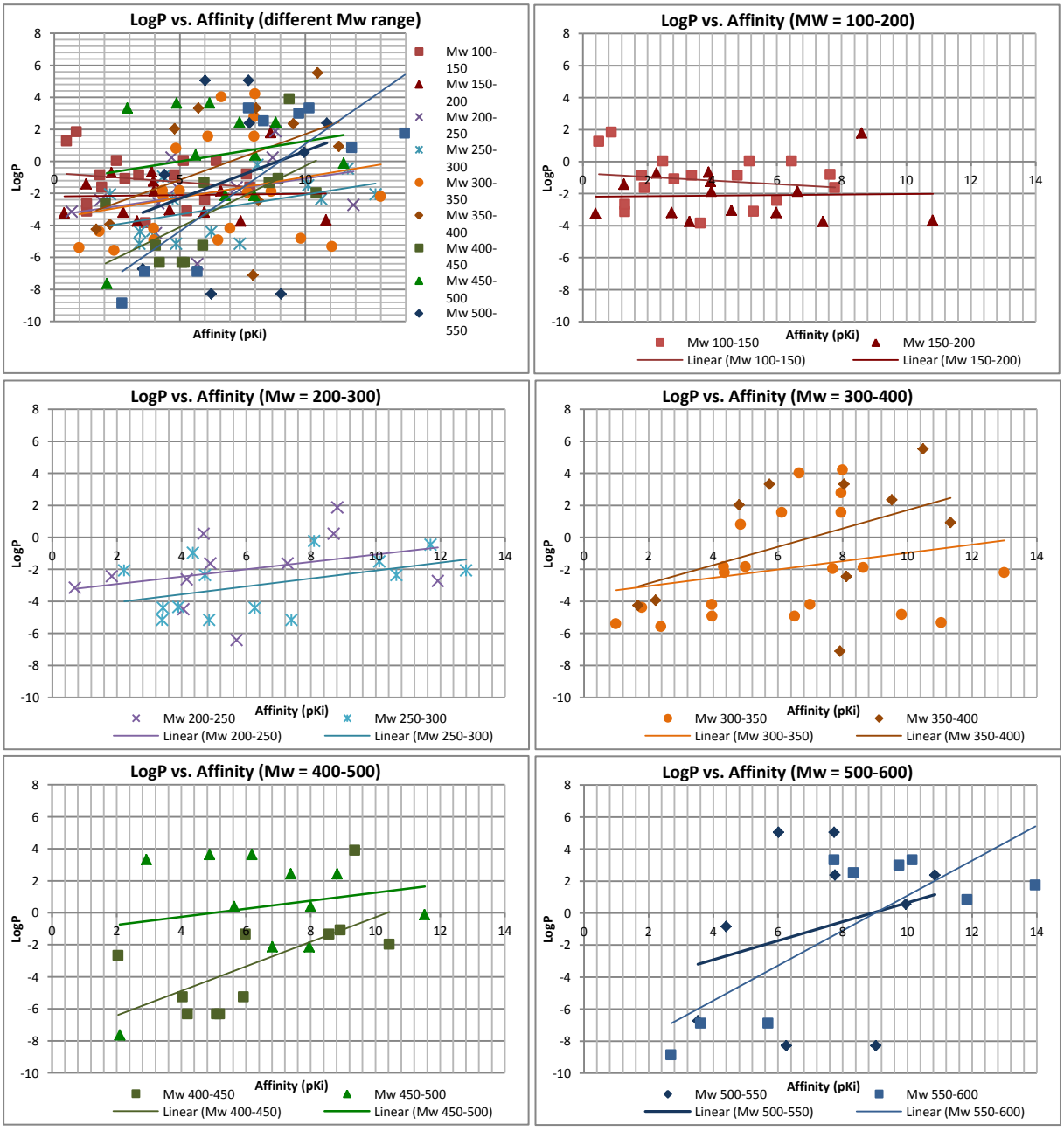

Figure 17 The relationship between logP and affinity in different molecular weight range.

Those three factors (Mw, logP, and affinity) also give another interesting aspect. The composition is almost the same, but now was looked at the relationship between logP and molecular weight in different binding affinity ranges (Fig. 18). The diagram shows that the logP value has no major impact on the binding affinity of the molecule below the molecular weight of 300 Da, but after that the logP value seems to affect dramatically to the binding affinity. The data also supports the view that the lipophilicity of better binding molecules increases when they molecular weight increases (Warring 2009). If only the best and worst binders of each weight group are taken to the examination, this tendency is emphasized more clearly (Fig. 19). Still, it should be noted that correlation is quite poor.

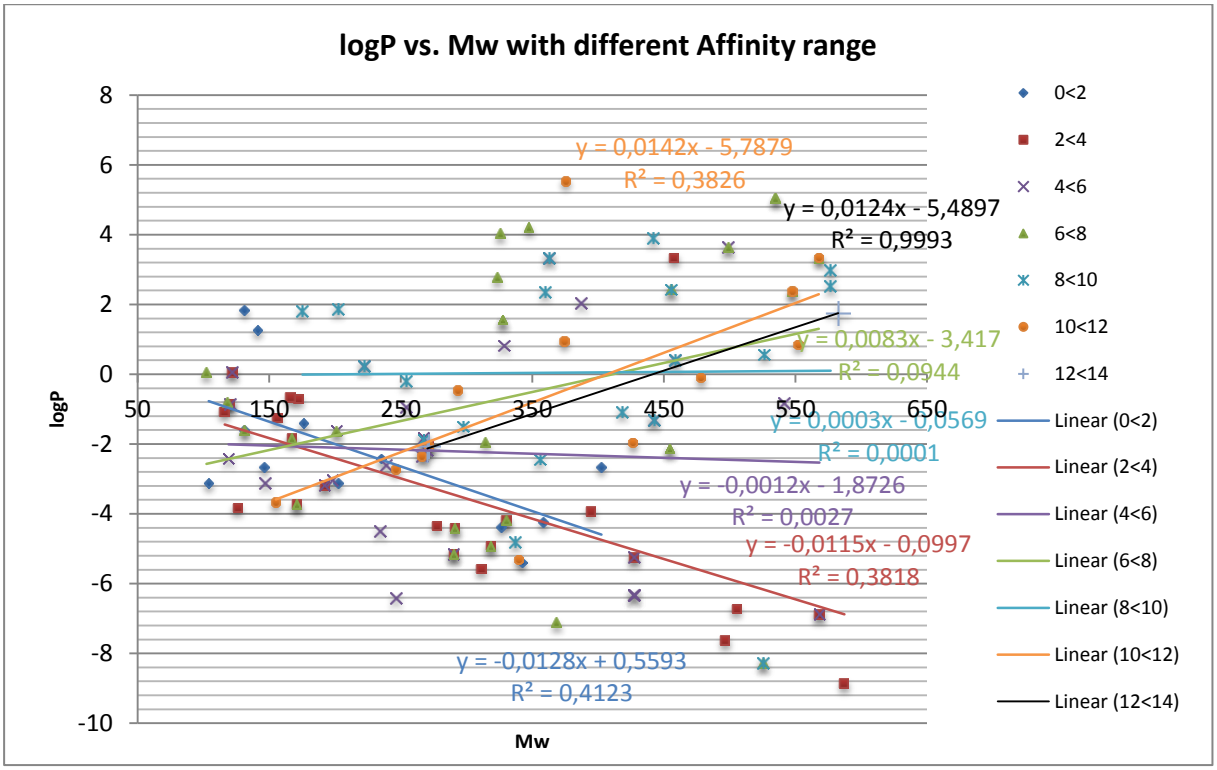

Figure 18 The relationship between logP and Mw in different affinity range.

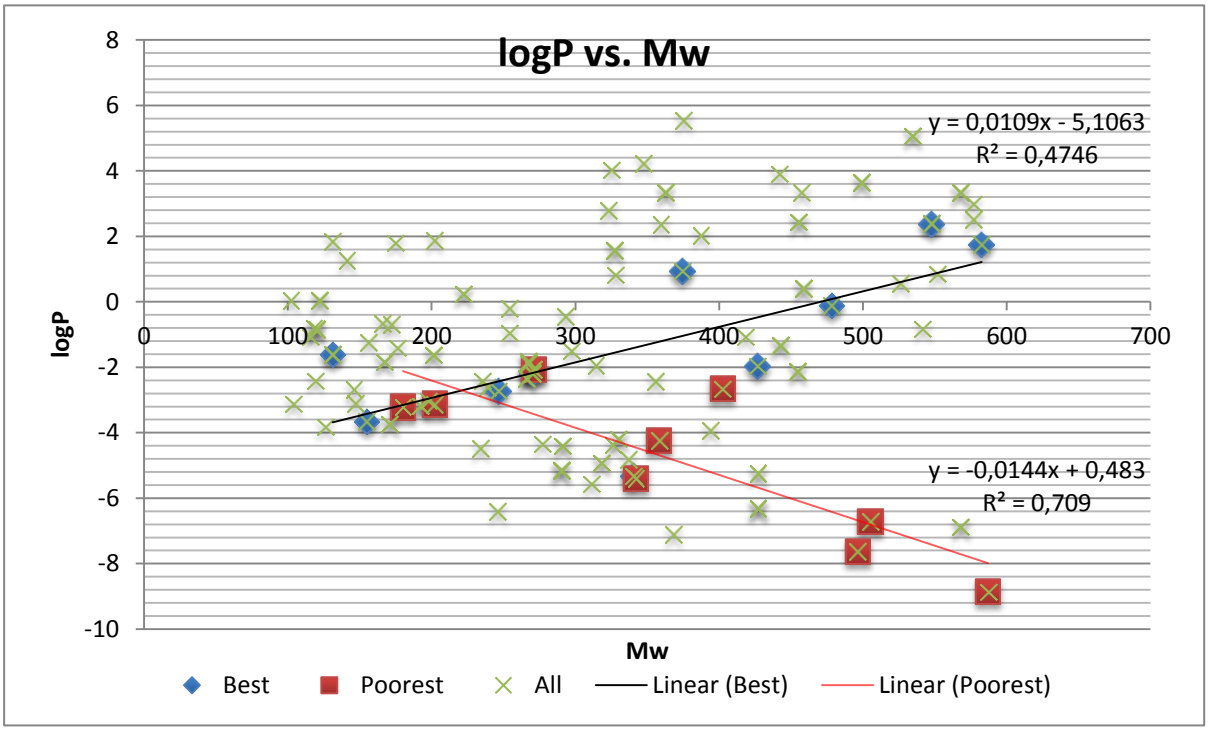

Figure 19 The relationship between logP vs. Mw. The best and poorest binding affinities in each weight range.

7.3 H-bonds

Figure 20 shows that between the amount of hydrogen bonds and the binding affinity is no correlation in general. Böhm and Klebe (1996) have reported similar results in their study. In addition, from Figure 21 may be seen an interesting circumstance. In the sample of this study does not have any complex, wherein the binding affinity achieves a greater value than 7 without any hydrogen bonds. What makes it even more intriguing, is that this same result can be also seen in the result of Böhm and Klebe's study (Fig. 21) (Fig. 2, Böhm and Klebe 1996). It would seem at binding molecules need hydrogen bonds in order to achieve a higher affinity. However, it should be noted that the graph does not take into account any other interactions than hydrogen bonds, so it cannot for sure know, which kind interactions affect the affinity.

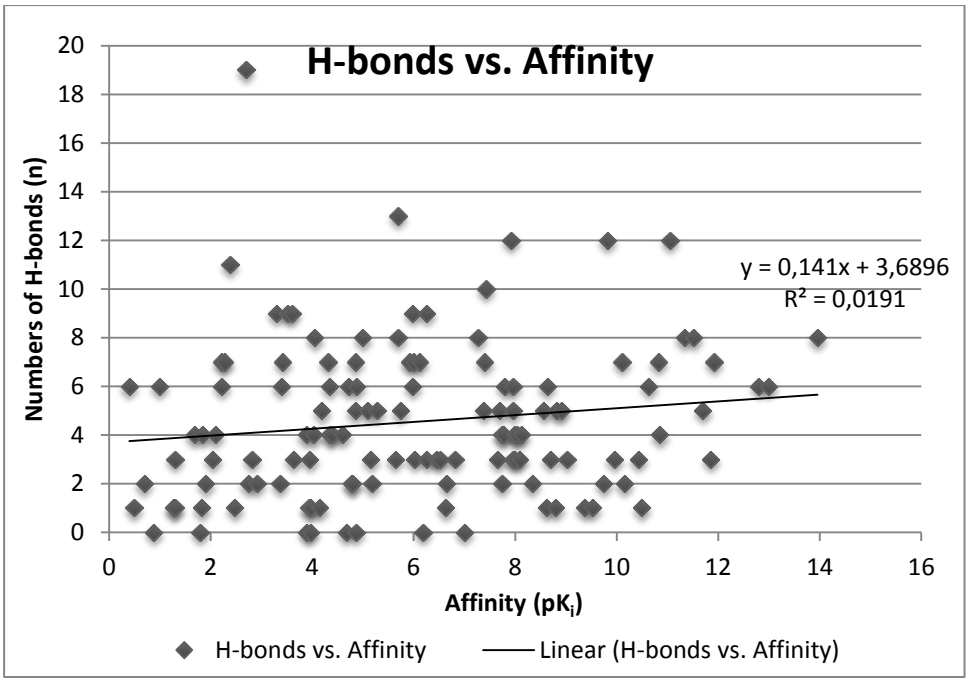

Figure 20 The relationship between the numbers of H-bonds and affinity (pK_i).

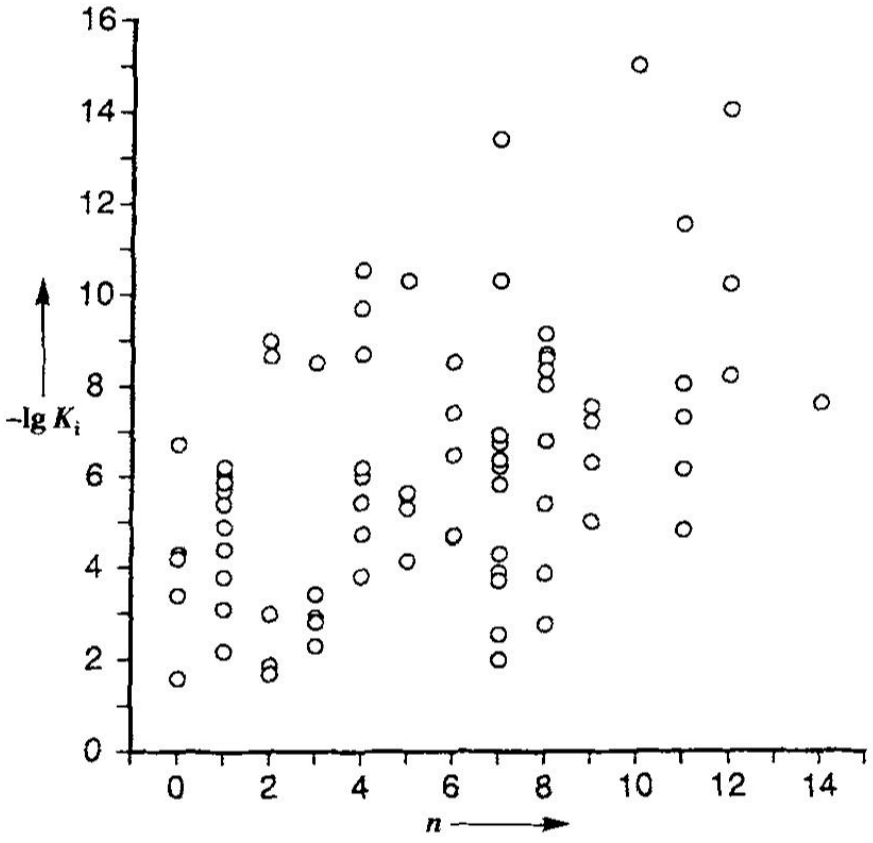

Figure 21 The relationship between the numbers of H-bonds (n) and affinity (-lg K_i = pK_i) from the results of Böhm and Klebe's study (Fig. 2, Böhm and Klebe 1996).

When the number of H-bonds and M_w are compared with different affinity ranges (Fig. 22), it seems to be case that small molecules, which have more hydrogen bonds, get more likely better binding affinity values. In contrast, the larger molecules with many H-bonds seem to get poorer binding affinity values. Although correlation is poor this may indicates that increasing the number of H-bonds does not provided better binding affinity in general.

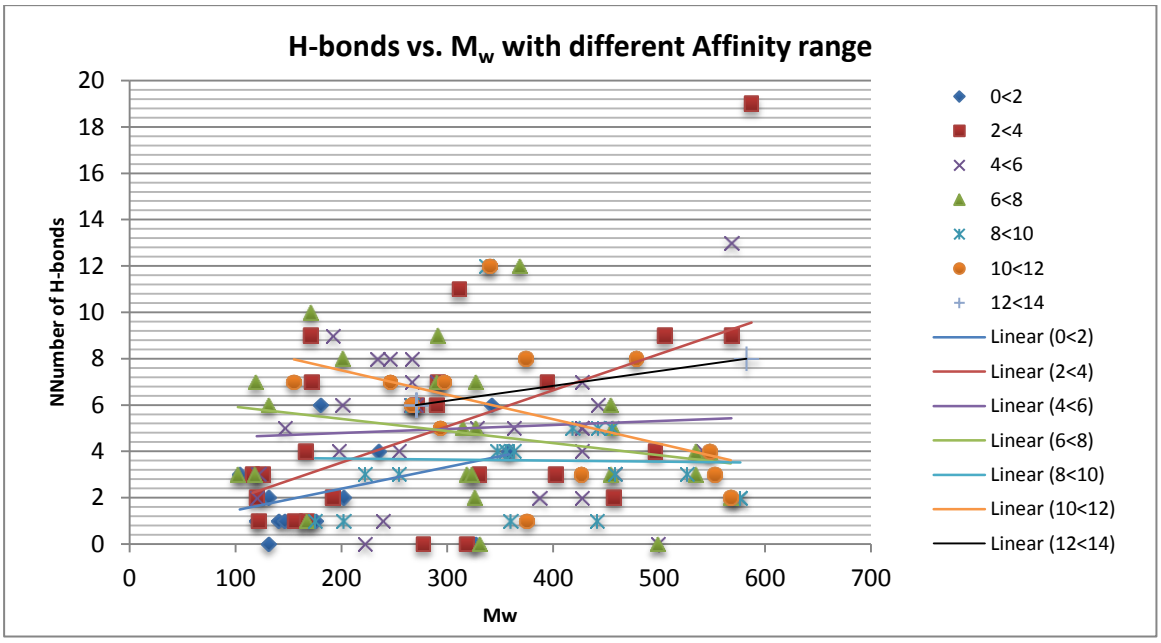

Figure 22 The graph can be deciphered that small molecules, which have more hydrogen bonds, get more likely better binding affinity values whereas the larger molecules with many H-bonds seem to get poorer binding affinity values.

7.4 The number of size points vs. pocket volume

SiteMap is capable to calculate a pocket volume and size, which indicate the number of site points in binding site. As Figure 23 shows binding sites, which have more site points in smaller volume, seems to get a better binding affinity values than those binding sites, which contain less site points in relation to volume. The tendency is observable, even though the correlation is quite poor.

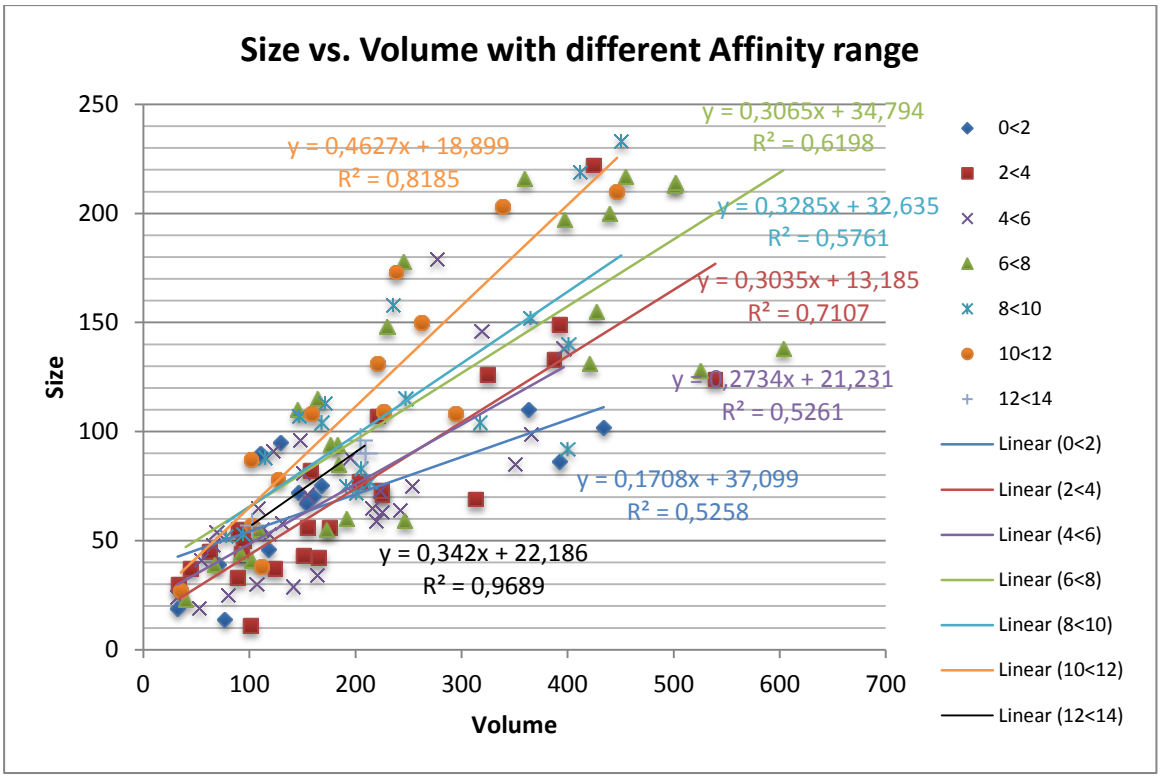

Figure 23 The relationship between sizes of ligands and pocket volumes in different affinity range.

7.5 Pocket types

When complexes were visually examined, it became apparent that it is quite important question where ligand is bound in protein. For this reason, the classification method, which helps to the binding pocket specification, was created. Every complex was categorized into one of the seven binding pocket types. Those are:

Well: This binding pocket type is narrow and clearly inside protein. The pocket is filled with ligand almost fully. In some cases the 'well' is deeper and ligand binds to the bottom of the pocket. This is the most common pocket type in this sample.

Cave: The volume of the 'cave' type pocket is much higher than the 'well' type and the opening is narrower than the pocket inside. There are usually lots of water molecules and sometimes also other small molecules inside.

Tunnel: The ‘tunnel’ type pocket can be narrow like ‘well’ or little bit roomier, but there are always at least two entrances. Usually, those are longer and deeper than the ‘well’ and narrower than the ‘cave’.

Crater: Like to name describe the ‘crater’ type pocket is like a crater on the moon and the ligand binds to the bottom of the crater. This kind pocket is not very deep.

Furrow: This type binding pocket is between the proteins folds and open to sides. The binding pocket is like a canyon.

Closed: This pocket type does not include clear entrance and its look like the ligand is trapped inside protein.

Surface: Ligand is bound on the protein surface. There is no kind of pocket.

The ‘Well’ pocket type includes 45 complexes, 7 complexes are the ‘Cave’ type, 25 complexes are the ‘Tunnel’ type, 26 complexes are The ‘Crater’ type, 15 complexes are the ‘Furrow’ type, only one complex is the ‘Closed’ type and 6 complexes are the ‘Surface’ type. 8 complexes were originally classified to the ‘Surface’ type, but the more accurate analyze has revealed that caspase-7 enzyme complexes crystallography data in PDB file has marked on the binding affinity value to wrong ligands, which binding site is different type and which bound covalently (more information in Section 7.7.3). Because of their accuracy could not verify these two are not considered in the comparison. It is absolutely needless to make any conclusions from the merely closed pocket sample alone, therefore the graph of them were not made.

It should be noted that six of the seven ‘Cave’ type complexes are the same type of nitric-oxide synthase enzyme and in addition to a reported ligand they all have a heme-molecule inside the pocket. Because the too similar complexes form this distribution, it was not used to draw any conclusion.

7.5.1 Ligands M_w in different pocket types

When the molecular weight distribution of the ligands is considered in the various pocket type (Fig. 24), some trends can be recognized. It seems that smaller molecules are favoured in the 'Well' type pockets as seems to be the case also with the 'Cave' type pockets. The latter is little unexpected, because these kinds of pockets are more spacious. When these complexes are examined more closely to seems that all the 'Cave' type pockets also have other larger molecules inside. As noted earlier, the generalizability of the results of 'Cave' type complexes is far from clear and results could also be misleading.

The 'Surface' type binding site seems to favor larger molecules. This seems quite reasonable, because larger molecules have more surfaces to compose interactions with the protein than smaller molecules and larger surface area also increases the vdW forces between the ligand and the protein, which can be decisive. Consequently larger molecules may form sufficiently high interaction energy in order to stay attached to the protein surface. So, it can be said that small molecules are "flushed" away from the surface more easily. The 'Crater', 'Tunnel' and 'Furrow' types do not seem to prefer certain sizes.

7.5.2 Affinity vs. M_w in different pocket types

The 'Well' type pockets seem to be able to achieve a very high binding affinity values - even with very small ligands. The most of the ligands binding into 'Well' type pockets are bound tightly and there are not much free spaces around ligands. This may be thought to reduce and stabilize the movement of ligands hydrogen bonds. This idea is also supported by the results of different studies. As an example, Nilsson and co-workers (2008) have reported in their study the better structural compatibility of FimH-mannoses in the binding pocket to improve a lifetime of a receptor-ligand interaction. The shorter interval hydrogen bond, less mobility and less free space around the ligand will thus increase the binding affinity, because access to the water molecules between the ligand and receptor is so blocked (Schmidtke *et al.* 2011). The increasing number of attacks of water molecules leads to increasing sensitivity of breakage hydrogen bonds between the receptor and the ligand (Lu and Schulten 2000, Craig *et al.* 2001, 2004a, 2004b, Gao *et al.* 2002, Nilsson *et al.* 2008).

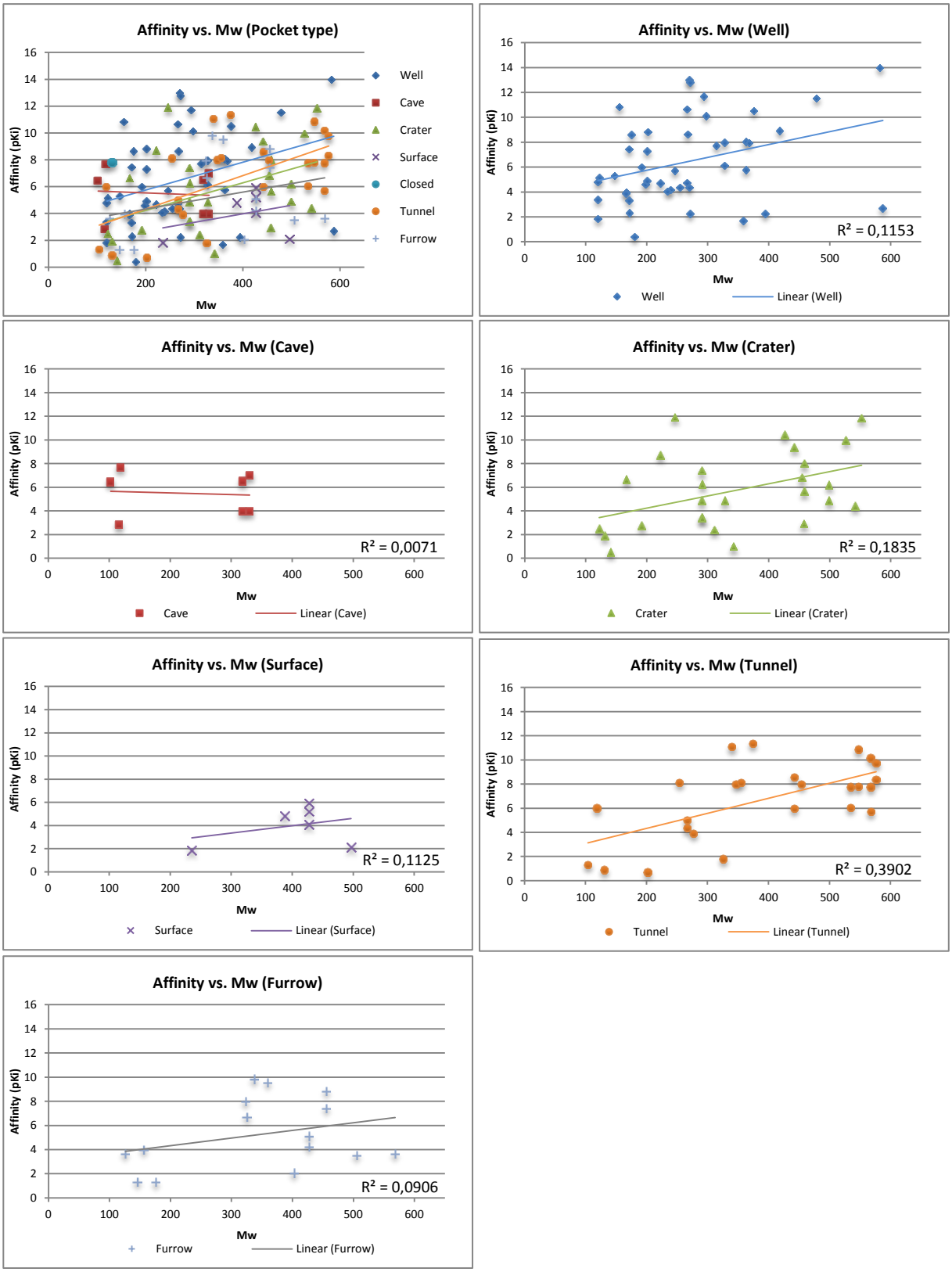

Figure 24 The relationship between affinity and Mw with the different pocket types.

The ‘Crater’, ‘Furrow’ and ‘Tunnel’ types seem to be able to achieve up to 12 pK_i’s binding affinity and can bind to both smaller and larger molecules. The ‘Surface’ type does not seem to reach a very large affinity values even with large molecules. All pocket types seem to follow the basic trend, which has been assumed previously (Fig. 15), that the bigger molecules can get better binding affinity values. However, the slight surprise was the fact that the ‘Crater’ type was able to achieve such a seemingly high binding affinity values. If the molecule fills out the pocket well, it protects the hydrogen bonds from water and so-called almost buried hydrogen bonds (ABPA) (Schmidtke *et al.* 2011) are formed. This kind hydrogen bonding allows a longer lifetime (Nilsson *et al.* 2008) and they may be up to 1.2 kcal / mol stronger (Gao *et al.* 2009).

7.5.3 LogP vs. Affinity in different pocket types

As mentioned in Section 7.2, logP value has an impact on the binding affinity of the molecule and its importance seems to become emphasized when molecular weight increases. When taken into consideration the binding site of the molecule, one of the most interesting phenomena appears (Fig. 25): the surface types binding affinity is not connected with logP values of ligands. On average, all the other binding types seem to favor greater lipophilicity in relation to binding affinity. Here are also the deviations and even large exceptions. The correlation is also quite poor, but on the other hand the sample is very small, and in smaller groups even a small deviation can strongly affect the average. It would be interesting to see these results from all 1492 complexes of the source material.

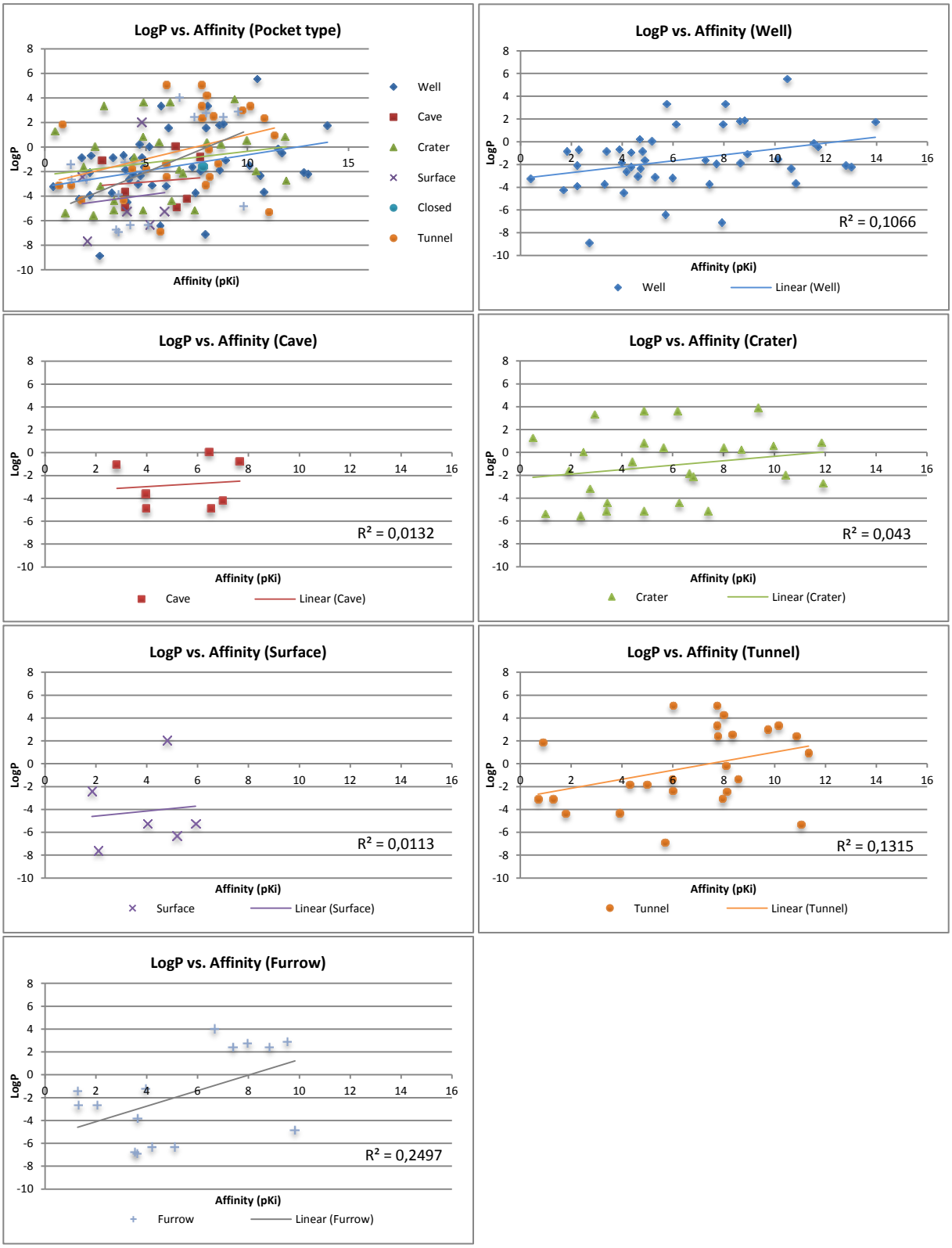

Figure 25 The relationship between logP and affinity with different pocket types.

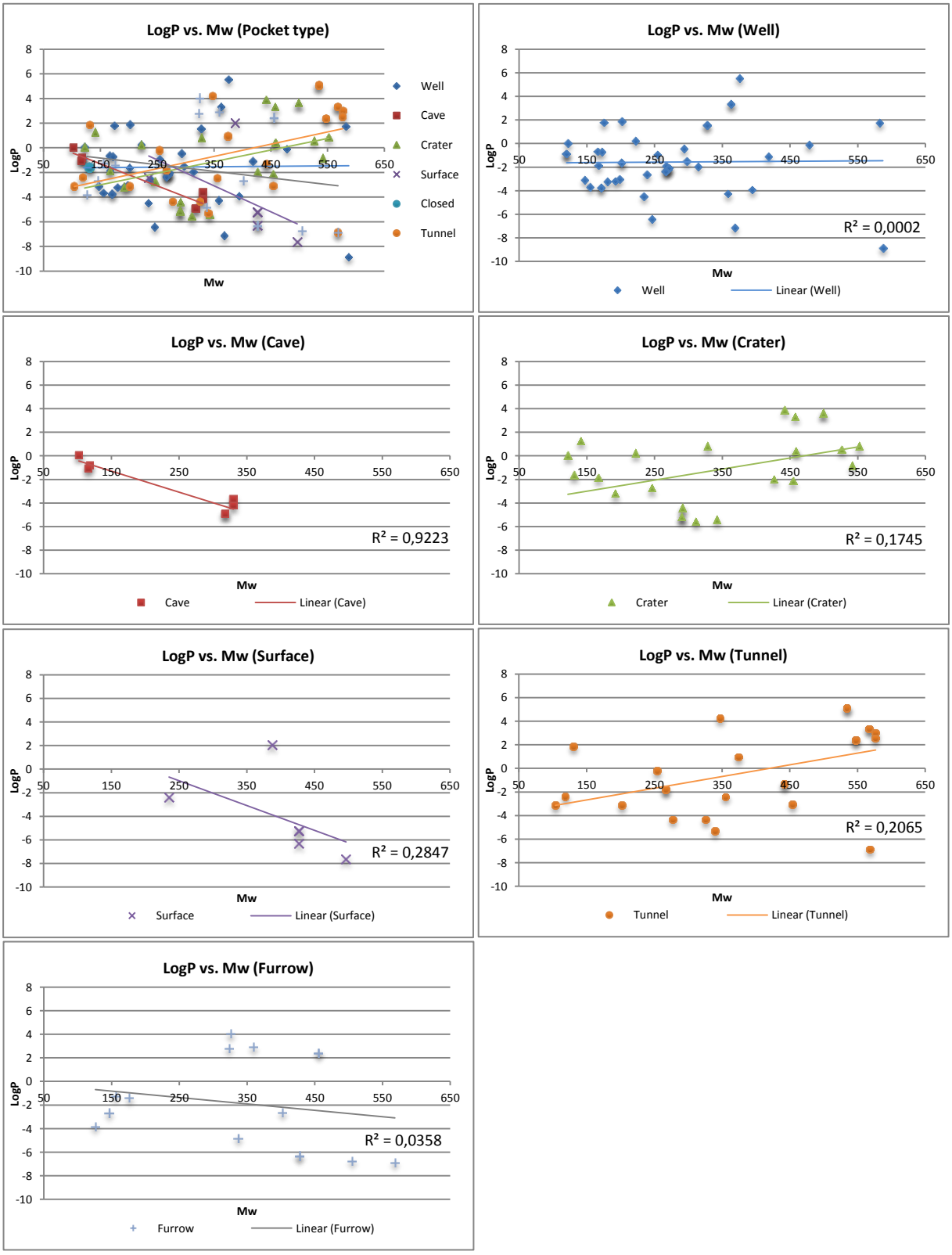

Figure 26 The relationship between logP and Mw with the different pocket types.

7.5.4 LogP vs. M_w in different pocket types

Figure 26 shows that between the molecular weights and lipophilicity of the ligands is no clear correlation, when considering in the different binding site types. However, by comparing at values of molecular weights in Figure 26 and values of binding affinity in Figure 25 in terms of y-axis (i.e. logP value), may be noted that the further assumption (Section 7.2 and Warring 2009) that the lipophilicity of better binding molecules increases when they molecular weight increases, seems carried out in most of the pocket types. The extreme example of this is found in the 'Well' type, wherein highly hydrophilic (logP = -8,86) and a large (Mw = 587) ligand have a poor binding affinity (pK_i = 2,7), while almost the same weight (Mw = 583) but more lipophilic (logP = 1,75) ligand have a really good binding affinity (pK_i = 13,9). Also in the 'Tunnel' and 'Furrow' types are found some large and very hydrophilic ligands, which have poor binding affinity.

The 'Surface' type would behave differently also in this assumption. Molecules that bind into the surface, on the contrary would seem to receive a better affinity when lipophilicity decreases and molecular weight increases. There is one exception (3EB1) in the 'Surface' types where 4-[3-(dibenzylamino)phenyl]-2,4-dioxobutanoic acid bind to the protein-tyrosine-phosphatase enzyme. The ligand is more lipophilic and has a better binding affinity than what one would expect. There are two possible explanations: The first one; when the binding site is considered more closely it can be seen that the ligands more hydrophilic head is positioned in a small pit-like point so that two benzyl ring remains outside the pit. Thereby protein interacts only with a hydrophilic head and one aromatic ring. This could also be classified in the 'Crater' type (Fig. 27), but the pit is so small and half of the ligand is outside the pit, hence this was led to the classification of the 'Surface' type. If it was classified the 'Crater' type, it should be in accordance with the previous assumption. Second one; in any case, here the bonds are protected from water when the pit is filled tightly by the hydrophilic part of the ligand, which improves the binding affinity (Schmidtke *et al.* 2011).

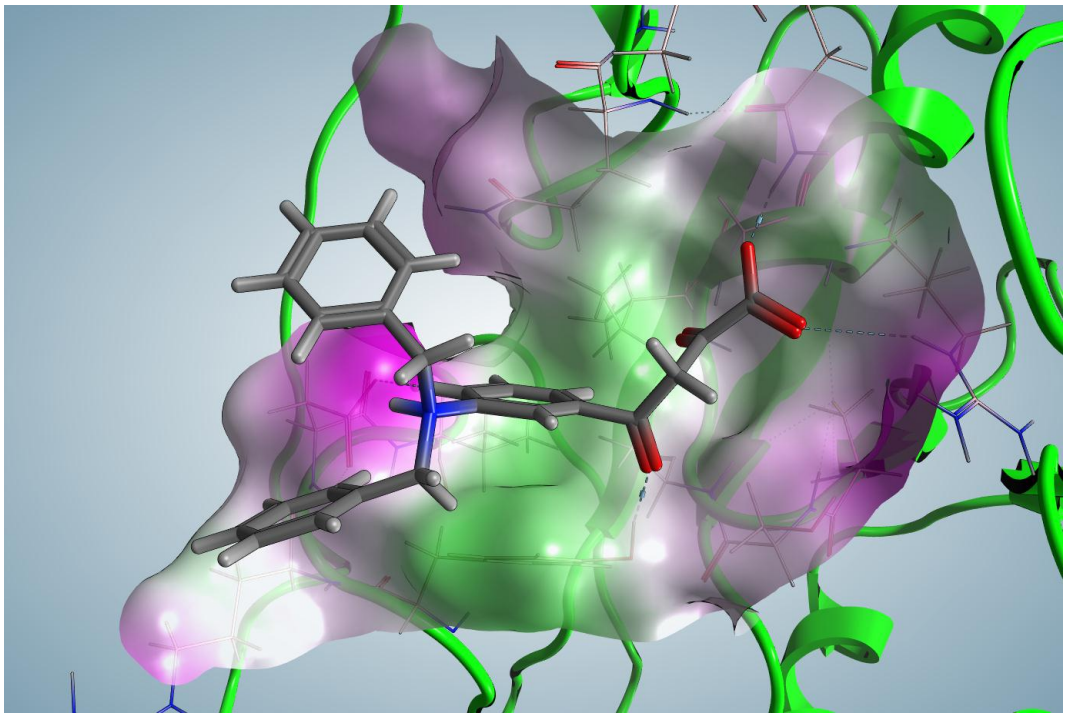

Figure 27 4-[3-(dibenzylamino)phenyl]-2,4-dioxobutanoic acid is bound into the binding site in protein-tyrosine-phosphatase enzyme. Two aromatic rings are remained outside the binding site.

It appears that the 'Surface' type does not follow the same trend as those, which bind inside to the protein. However, same kind of tendency is also visible in the pockets of Cave-types, although those binding site were inside the protein, but it should be noted as mentioned above in Section 7.5, the 'Cave' type is not fully comparable. On this basis, could be said that molecules that bind to well protein on the surface, seem exhibit abnormal chemical properties compared with those that bind better inside the protein. However, it was noted that the correlation is still very poor and on this basis cannot make to the highly generalized conclusion.

7.6 SiteScore and DScore

In this study, all complexes are examined by Scrödinger's SiteMap application. In SiteMap are two scoring functions, SiteScore (Halgren 2007) and Druggability Score (Halgren 2009), which can be evaluated binding sites and those druggabilities. SiteScore and DScore (Druggability Score) are based on a weighted sum of same three properties but different

coefficients. With SiteScore function the scores of greater than 1 suggest a particularly promising drug-binding site and value of 0,8 is considered the threshold between drug-binding and non-drug-binding sites (Halgren 2007). DScore function has been developed so that it would be able to distinguish “difficult” and “undruggable” binding targets from “druggable” ones (Halgren 2009). There was great interest to know, can those functions correlate with binding affinity.

Unfortunately, SiteScore and DScore do not correlate in any way with the binding affinity (Fig. 28 and 29) and with SiteScore almost all of the values fit into the same range of variation. When looking at the results in terms of different binding sites, two interesting things can be seen. The first on; the ‘Surface’ type complexes get much poorer Dscore and SiteScore values than in other binding types. The second on; it appears that the ‘Tunnel’ type complexes get almost without exception the peak values. Even so that the two of the three weakest binding pockets to get the best scores by both scoring functions.

These deviations can be explained by the emphasis of those scoring functions 1 & 2:

$$SiteScore = 0{,}0733\, n^{1/2} + 0{,}6688\, e - 0{,}20\, p \quad (1)$$

$$DScore = 0{,}094\, n^{1/2} + 0{,}60\, e - 0{,}324\, p \quad (2)$$

The scores of those functions are formed a linear combination of only three descriptors: pocket volume encoded by site points (*n*), hydrophobicity (*p*) and enclosure (*e*) (Halgren 2007 & 2009). The enclosure is the descriptor of buriedness and therefore the results of these functions will favour the pockets, which are more buried. Because the ‘Surface’ type binders have not really any kind of pocket and so the enclosure values are automatically smaller, they will get a lot worse SiteScore and DScore values. All things considered it is probably safe to say that SiteScore and DScore are not suitable for predicting the potential binding affinity of a pocket.

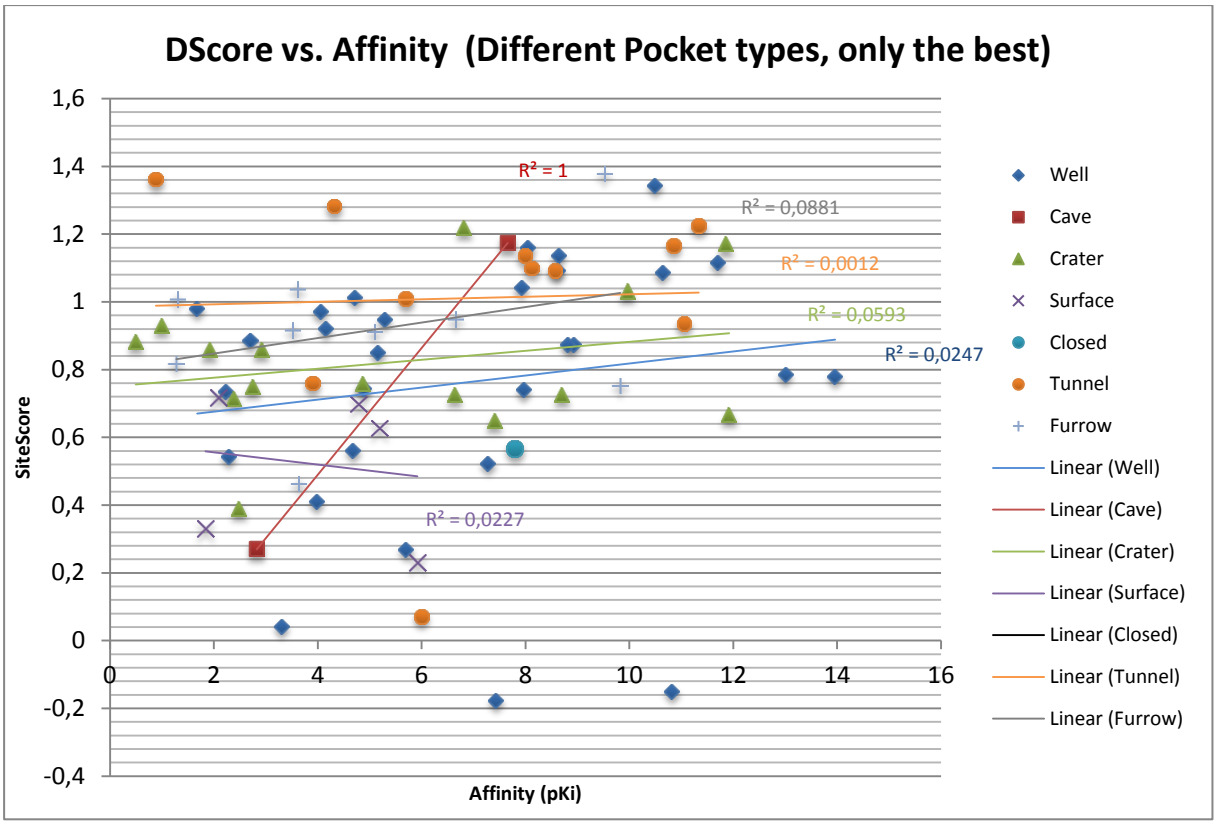

Figure 28 The relationship between DScore and the best affinity of each protein family.

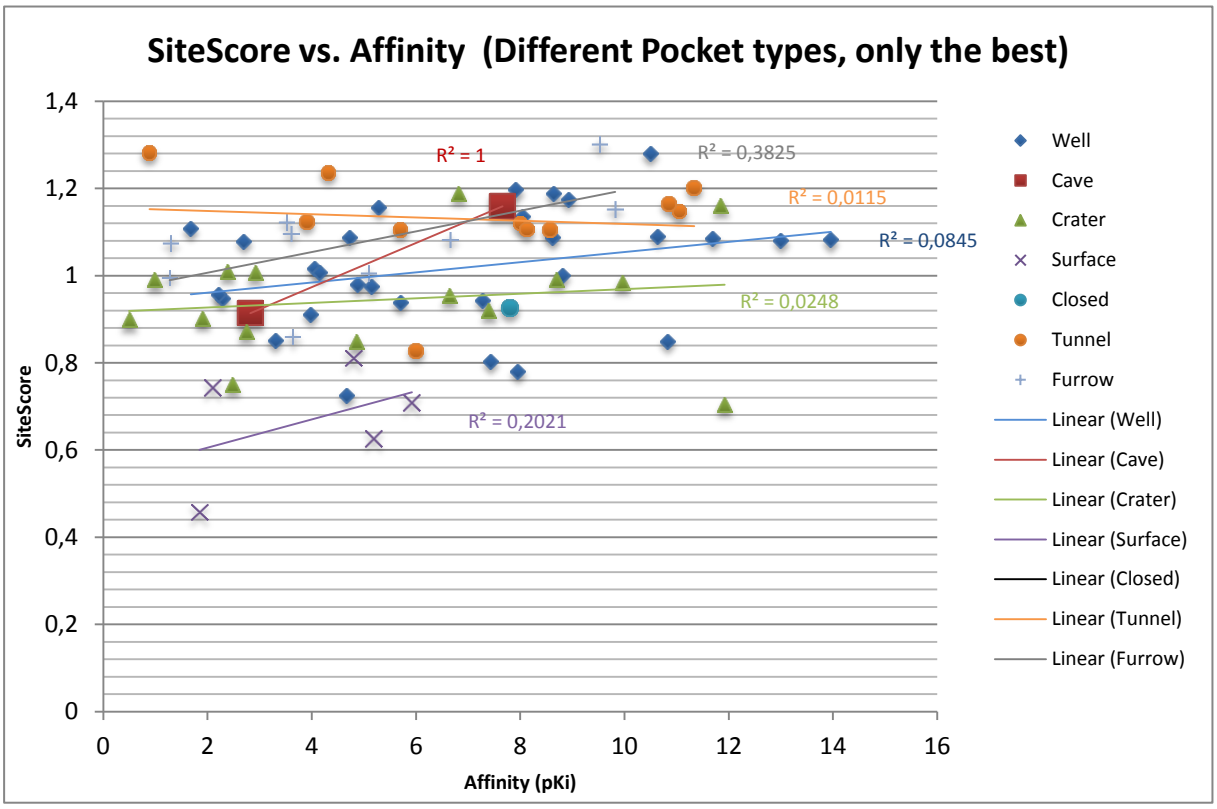

Figure 29 The relationship between SiteScore and the best affinity of each protein family.

SiteMap may still be useful to the predictions of the binding sites and their properties. When Halgren (2009) used SiteMap to the binding site identification, it identifies correctly 86 % of a set of 538 complexes which taken from the PDBbind database when used SiteScore algorithm and score cutoff was 0,8. Similar results have been obtained from other studies. Perola and co-workers (2012) identified correctly 88,3 % of the binding-sites from Drug Target Set with SiteMap when a score cutoff was 0,8. Drug Target Set is their own compilation of the confirmed targets of approved drugs with one or more publicly available crystal structure (Perola *et al.* 2012). Schmidtke and Barril (2010) have compared SiteMap's DScore and Fpocket (Le Guilloux *et al.* 2009) methods abilities to the predicting binding sites but they don't tell exactly how god those are.

Volkamer and co-workers (2012) compared the results of their own binding site identification method DoGSite with SiteMap's results, which were provided by Schmidtke and Barril (2010). They all used the druggability data set (DD) as well the nonredundant version of it (NRDD), which was compiled by Schmidtke and Barril (2010). Volkamer and co-workers (2012) reported on DoGSite algorithm correctly classified 88 % of DD targets and their method perform comparably well to SiteMap's DScore.

Even 89,6 % of 125 complexes of this study get value at least 0,8 with SiteScore function, it can not use to proving the ability of binding site predictions, because it should be noted that in this study SiteMap analyzation was made by selecting the true bounded ligands to starting point. In other words, true binging site was already known and no binding site prediction was made, only the analysis of binding sites.

7.7 Special cases

Here are presented a few exceptional cases that do not support the assumptions given above. They have their own logic, but in such cases the calculated prediction is hopeless. Information on these can be increased by a more detailed study and this underlines the importance to know the target protein binding site as well as possible, so as to achieve successful results in the drug design process.

7.7.1 The case of adenosine deaminase mutation

When Chang and coworkers (1991) discovered two adjacent aspartates in adenosine deaminase gene sequences, the research team thought that they were an important part of the functioning of the enzyme. Soon it was discovered that the Asp295 and Asp296 are located in the active site and they interact with the ligand, the zinc cofactor and the catalytic water (Wilson *et al.* 1991). Asp 295 and three histidines (15, 17 and 214) occupy zinc cofactor together and keep it in place (Wilson *et al.* 1991, Sideraki *et al.* 1996a, Wang and Quiocho 1998). Then Asp 295, zinc cofactor and His 238 could orientate the catalytic water correctly for the deprotonation by His 238 (Sideraki *et al.* 1996a, Wang and Quiocho 1998). Glutamate 217 is also involved in this proton transfer chain together with His 238 and adenosine ring (Wilson *et al.* 1991, Wilson and Quiocho 1993).

Although Asp 296 not takes part in deaminase process it interacts direct to adenosine and helps by orienting it correctly (Sideraki *et al.* 1996a). Sideraki and coworkers (1996a) showed the importance of Asp 295 and Asp 296 to enzyme reaction by mutating the enzyme in such a way that one of the aspartates was changed to alanine. When Asp 296 was changed to Ala 296, 6(R)-hydroxy-1,6-dihydropurine riboside binding affinity (pK_i) plummeted 12,80 to 2,22. As Figure 30 show, the three-dimensional structure of the proteins does not change, so the only explanation for the dramatic deterioration of the binding affinity is the importance of Asp 296 for deaminase process. The crystallize structure has downloaded from PDB (Asp 296 = 1A4M and Ala 296 = 1FKX).

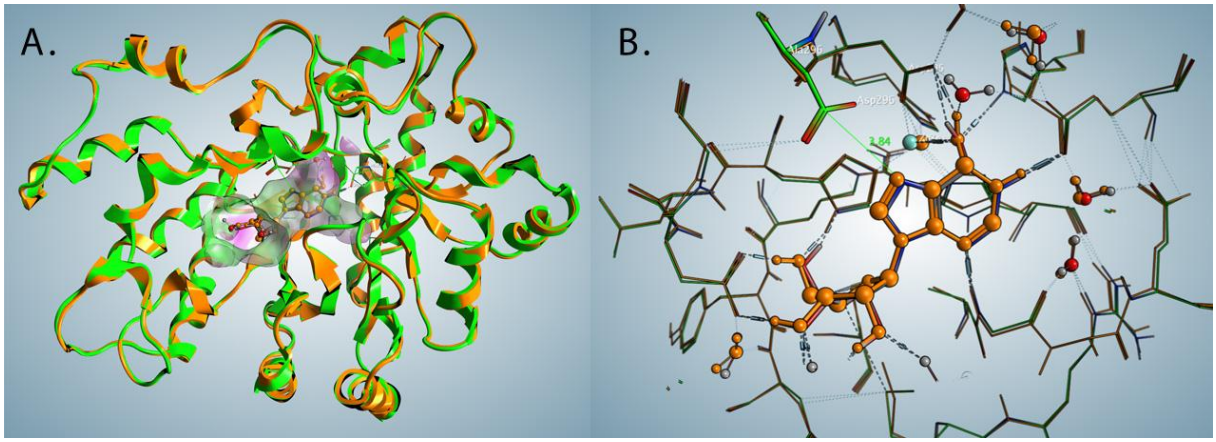

Figure 30 The superimposed structures of adenosine deaminase enzyme (green) and the mutation of this (orange). **A.** The structures of enzymes are very uniform. **B.** The spatial structures are very similar the same binding site also with the water molecules and the ligand. The enzyme structures differ only between amino acids Asp 296 and Ala 296.

Similar dramatic reduction in enzyme activity of the ADA has also been reported with the mutations of His 17, His 214 (Bhaumik *et al.* 1993), Glu 217 (Mohamedali *et al.* 1996, Bhaumik *et al.* 1993) and His 238 (Sideraki *et al.* 1996b, Bhaumik *et al.* 1993). The effect of His 238 mutation to Ala 238 is displayed in the materials of this study. The binding affinity of 6-hydroxy-7,8-dihydro purine nucleoside to His 238 is 13,0 (pK_i) and PDB code is 2ADA while binding affinity to Ala 238 is 4,35 (pK_i) and PDB code is 1UIO. Binding pocket type of adenosine aminase enzyme is the 'Well' and it should be noted that these unusually poor binding affinity values, which caused by the deterioration of the mutated enzyme activity, distorted results which examined the relationship between the binding affinity and the properties of ligand-protein complexes.

7.7.2 Case of HIV-1 retropepsin and one water molecule

The material of this study was 8 HIV-1 protease enzyme. HIV-1 protease is a homodimeric enzyme (Miller *et al.* 1989) and it consists of two identical proteins that have settled against each other in such a way that the tunnel is formed between them and inside this tunnel is the binding site of inhibitors (Wlodawer & Erickson 1993, Lam *et al.* 1994). Soon after the structure of the enzyme was solved (Navia *et al.* 1989 and Wlodawer *et al.* 1989) the importance of the water to protein activity became clear (Wlodaver and Erickson 1993). The protease has a flexible "flap", which is formed from β-strand (residues 43-49) and β-chain (residues 52-58) (Lapato *et al.* 1989, Miller *et al.* 1989, Navia *et al.* 1989, Wlodawer *et al.* 1989), which form the roof of the tunnel (Fig. 31). In these "flaps" Ile50 and Ile50' settled against each other and they are bridged to each other through the single buried water molecule.

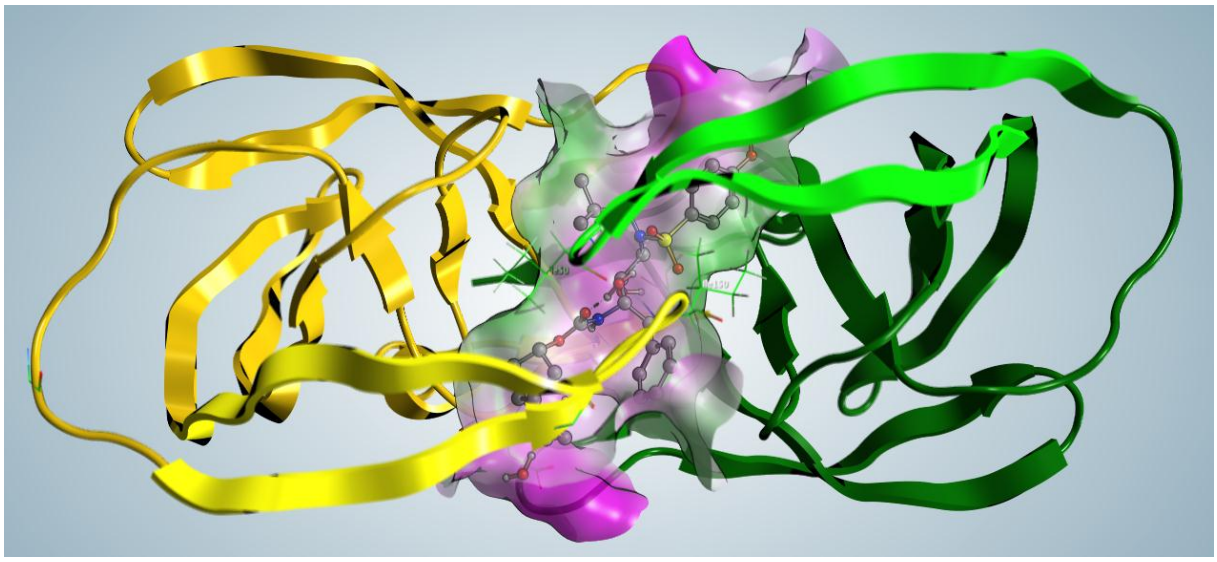

Figure 31 HIV-1 protease enzymes contain two identical proteins (green and yellow) that have settled against each other in such a way that the tunnel is formed between them and inside this tunnel is the binding site of inhibitors. Brighter parts indicate the "flaps" of proteins and W301 is located between heads of these "flaps". Structure's PDB code is 3DK1.

This single water molecule, also known as Water 301 (Baldwin *et al.* 1995) and "flap water" (Singh and Senapati 2008), is displayed in almost all the known crystal structures of the inhibited HIV protein in the same place (Wlodaver and Erickson 1993), as well as in six samples of HIV-1 protease complexes in this study (Abdel-Meguid *et al.* 1994, Kožíšek *et al.* 2007, Liu *et al.* 2008, Wang *et al.* 2007, Ghosh *et al.* 2008, Surleraux *et al.* 2005). Miller and coworkers (1989) discovered that this tetrahedrally hydrogen-bonded water molecule (W301) accepts two H-bonds from the flaps of the protease residues Ile 50 and Ile 50' and donated two H-bonds to the carbonyl oxygen of the inhibitor peptides (Fig. 32). These four hydrogen bonds from one water molecules are something very exceptional and this herald the strong interaction. Several studies (Grzesiek *et al.* 1994, Singh and Senapati 2008) have proved that the H-bonds of buried W301 are unusually long-lived, up more than 9 ns.

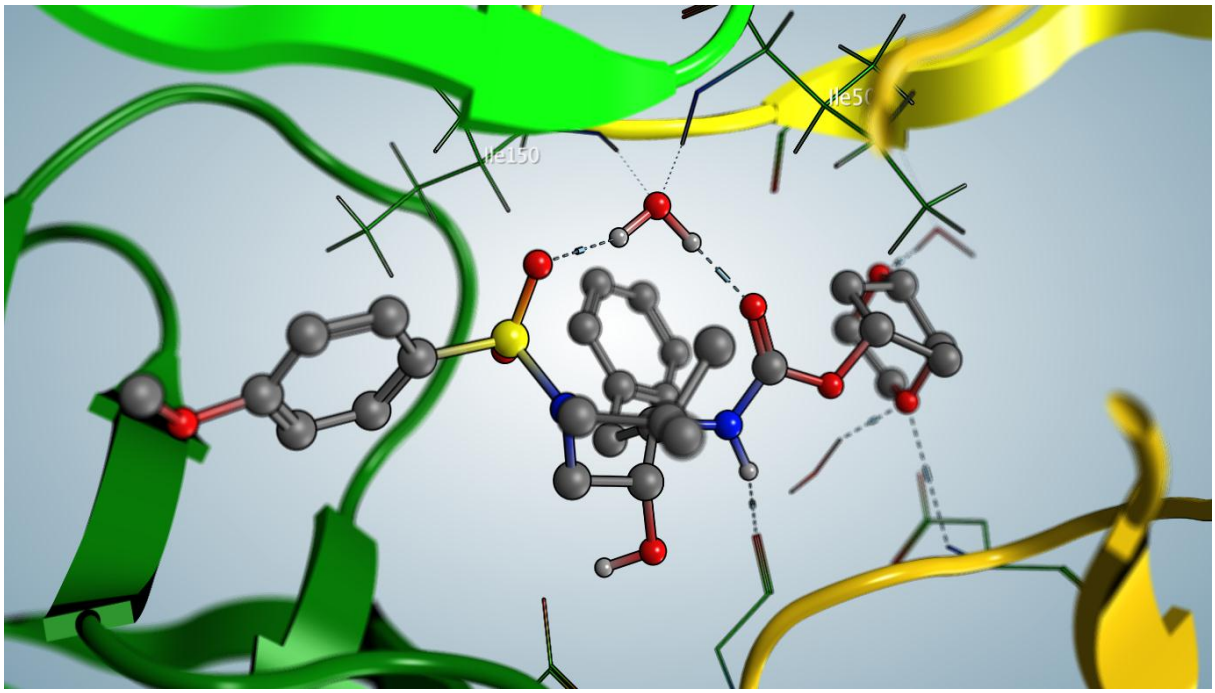

Figure 32 W301 forms tetrahedral hydrogen-bonds between Ile50 and Ile50' of "flaps" of HIV-1 proteases and the carbonyl oxygens of the inhibitor. This is very exceptional and the bonds are unusually long-lived, up more than 9 ns (Grzesiek *et al.* 1994, Singh and Senapati 2008). Structure's PDB code is 3DK1.

The role of water in the activity of HIV-1 protease is not only this. To date, the essential of flaps flexibility for efficient catalytic and activation processes in HIV-1 protease are well-documented (Nicholsan *et al.* 1995, Shao *et al.* 1997, Ishima *et al.* 1999, Freedberg *et al.* 2002, Katoh *et al.* 2003, Tóth and Borics 2006b) and consequently are awakened interest in the possibility of finding the allosteric binding site of the inhibitor, which would affect the flexibility of the flap and thereby disrupt with the HIV-1 protease function (Perryman *et al.* 2006, Hornak and Simmerling 2007). Both earlier experimental (Spinelli *et al.* 1991, Ishima *et al.* 1999, Pillai *et al.* 2001, Freedberg *et al.* 2002, Heaslet *et al.* 2007) and computational (Collins *et al.* 1995, Scott and Schiffer 2000, Tozzini and McCammon 2005, Tóth and Borics 2006a, Hornak *et al.* 2006) studies have demonstrated that free HIV-1 protease may occur in different conformations, which have been classified into three main categories: fully open (Heaslet *et al.* 2007), semiopen (Spinelli *et al.* 1991) and closed (Pillai *et al.* 2001). These conformations occur in proteases catalytic cycle (Tóth and Borics 2006b), which starts when the fully open enzyme grabs the substrate to inside. The closing of flaps places the substrate in the proper position and subsequent enzyme hydrolyzes substrate and then the opening of flaps releases the product. Enzyme inhibitors fit into the same mechanism. Flaps are

therefore an important part of the enzyme and Singh and Senepati (2008) showed that the structural water (W301) is a critical part of the flap closing dynamics.

Thereafter in the development process of the HIV-1 protease inhibitor are taken into account this water molecule. W301 has also succeeded replaced by the molecule that included cyclic urea wherein double bonded oxygen performs the role of W301 (Grzesiek *et al.* 1994, Singh and Senapati 2008), but still buried water has been found to induce a more stable inhibitor-protease complex than the CO-group of cyclic urea (Singh and Senapati 2008).

7.7.3 Caspase-7 and wrong ligands

From the material of this study was also found two erroneous binding data, both Caspase-7 enzyme complexes. Pentapeptide Gln-Gly-His-Gly-Glu, which is bound onto 'Surface', was marked to valid ligand in the Binding MOAD. Other measured binding affinity was pK_i 9,03 (PDB code 2QL9) and other was pK_i 6,26 (PDB code 2QL7). Astonishment is aroused that the crystal structures and interactions did not seem to be any difference, but the difference of binding affinities is almost three pK_i units. When this was revised from the original research article, it was found that the measured values of binding affinities were recorded to the other peptides (Fig. 33). Agniswamy and co-workers (2007) have reported measuring the binding affinity 9,03 with Ace-Asp-Gln-Met-Asj pentapeptide and 6,26 with Ace-Ile-Glu-Pro-Asj pentapeptide. Both of these pentapeptides form a covalent bond with the caspase-7 enzyme, so these two were not suitable for this study and they were ignored.

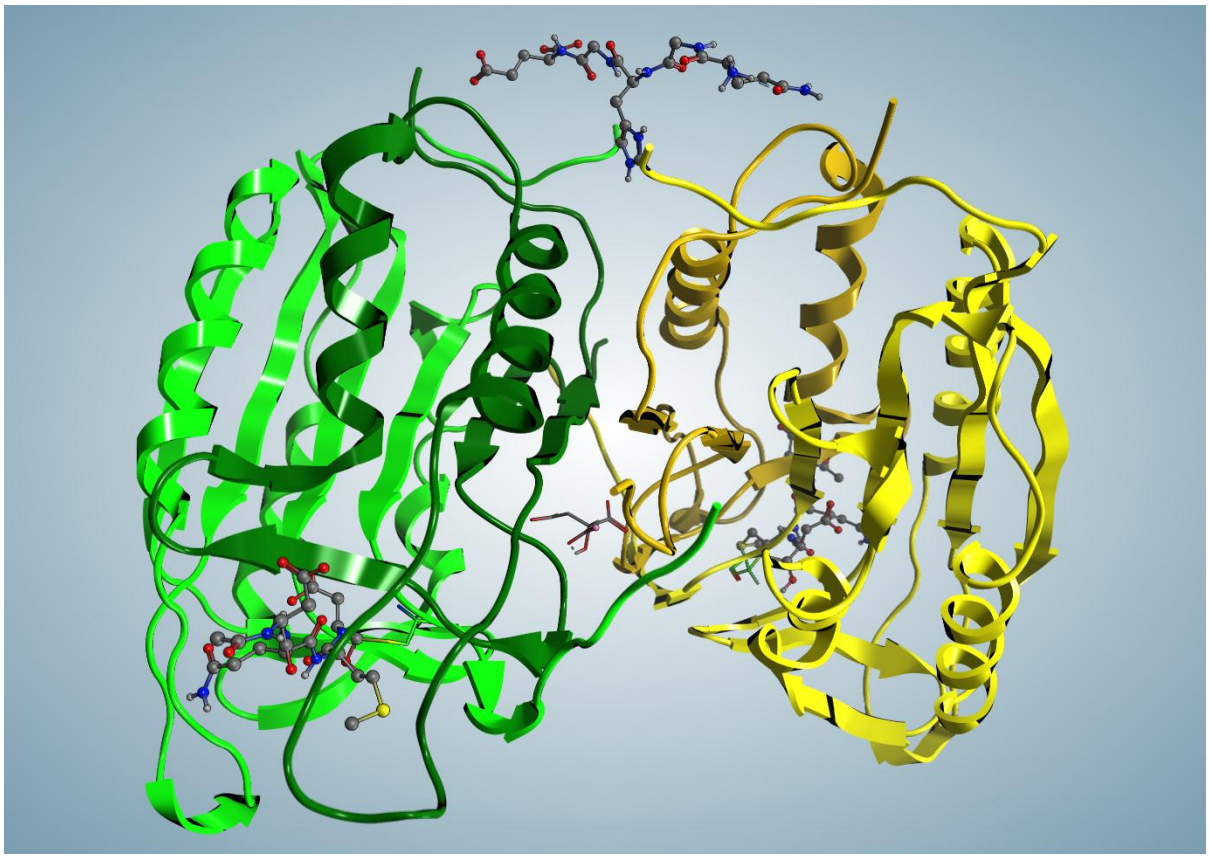

Figure 33 In the BindingMOAD, the measured binding affinity was marked to pentapeptide Gln-Gly-His-Gly-Glu (above enzyme's fourth complex). The correct measured ligand is Ace-Asp-Gln-Met-Asj pentapeptide (Two similar binding sites are located on the opposite sides of protein complex), which is covalently bonded to the protein. Structure's PDB code is 2QL9.

8 Conclusion

The fundamental question was set out the beginning of this study: What are the differences between good and not so good binding pockets? Unfortunately, it is not yet possible to provide an exhaustive answer, but some interesting aspects and thoughts was found, which may lead to future studies closer to the answer.

First, it can be concluded that the binding affinity will not grow any more on average when the molecular weight is above 600 Da (Fig. 15). This result is in line with expectations and the results of study by Kuntz and co-workers (1999).

DScore and SiteScore values do not correlate in any way with the binding affinity. Both scoring functions are strongly weighted by enclosure values, which favors buried binding sites. The extreme example of this can be seen in Figures 28 and 29, where the two of the

three ‘Tunnel’ type weakest binding pockets receive almost the best values of this dataset with both DScore and SiteScore functions.

In this used data, two of the 127 complexes were erroneous, so 1,6 % of the data contained an error. From the PDB loaded structures contained much more the quite small structural defect, which was easy to fix. However, it should be noted that large databases contain very likely also some erroneous, which can affect the results of studies. Based on this study it can be estimated that an average about 1-2 % of the data may be distorted in Binding MOAD. Possibly, there are errors in the same proportion also in other databases.

The increasing of lipophilicity does not correlate with the binding affinity although on average a slight increasing was observed. However, one interesting tendency was observed when compared the lipophilicity with the molecular weight in the different binding affinity ranges (Fig. 18). After the molecular weight of 300 Da, the effect of logP value for the binding affinity seems to be more and more dramatically intensified when the molecular weight of ligand increases. If only the best and worst binders of each weight group are taken to the examination, this tendency is emphasized more clearly (Fig. 19). Although the correlation of this is very poor, it raises the interest that would this remarkable tendency also being realized with the greater dataset or is hits tendency only a great random error?

The numbers of formed H-bonds do not correlate any way with the binding affinity (Fig. 20). This result is line with the results of previous study (Böhm and Klebe 1996). Of course, it is possible that the increasing number of hydrogen bonds may be improved affinity in the certain proteins or protein families, but in general, it cannot be said that the increasing number of H-bonds will increase or decrease the binding affinity. Still, one interesting circumstance may be seen from the results (Fig. 20). In the sample of this study does not have any complex, which achieves a greater value of the binding affinity than 7 without any hydrogen bonds. To this observation will make even more intriguing that this same result can be also seen in the result of Böhm and Klebe’s study (Fig. 21).

Even though the correlation between H-bonds and the binding affinities could not be found, it does not mean that hydrogen bonds would not have an important role in the binding of molecules. Vice versa, in many cases the role of H-bonds can even be crucial, but it should be noted that the binding affinities of ligands depend on many different factors.

The results of this study support the observation that the location and the shape of the binding site may be affected to the biding affinity, because in the different types of binding pockets prevail different conditions. The binding sites were classified in seven different types for making it easier to assess the impact of the shape and location for the binding affinity. Even though there is no strong correlation, some interesting tendency can be seen. The binding sites, which contain more site points in less volume, seem to get higher binding affinity values than those, which contain less site points in relation to volume (Fig. 23). The numbers of site points illustrate pretty much to ligand volume (two to three site points are typically found for each atom of the bound ligand, including hydrogens) (Halgren 2009). Thus based on this result may be hypothesized that ligand, which fills up the binding site more completely, indicates stronger binding affinity. This supported by the fact that the protected and tighter H-bonds are proved to be more stable and more long-lived (Nilsson *et al.* 2008, Schmidtke *et al.* 2011) and the unprotected hydrogen bonds are also more susceptible to attacks by water molecules, which breaks the bonds (Lu and Schulten 2000, Craig *et al.* 2001, 2004a, 2004b; Gao *et al.* 2002, Nilsson *et al.* 2008, Schmidtke *et al.* 2011). So, based on these results support the perception that more protect H-bonds will increase the binding affinity.

The results of this study are insinuated many different tendencies, but any of them could not be attested to the scientifically correct by based on this stud results. The binding process happens in the biochemical environments, which be composed of plenty of factors. In such a complicated system, all the factors interact with each other and some are even directly depend on the other factors. Most probably, it is impossible to find any two or three variables, which could correlate, alone or together, directly with the binding affinity. The tendencies of this study and hypothesis of them would provide an interesting subject for further study.

In some cases, maybe it would be wiser to think about the probabilities of trends. For example, if the tendency, which is discussed in Section 7.2 (Fig. 18), seems to follow these results in the larger dataset, it cannot only explain with a single straight commensurate factor. Figure 18 describes to the depending three factors for each other and any correlation between those cannot find. Still, there seems to be tendency that the logP value affects more to the binding affinity when molecular weight growing.

After this study there is maybe less answers and more detailed questions, so this topic provides still much to further researching. In the future, this study could continue to refine the definition of the binding sites types. The approach of study could be also considered in more detail. One of further studies could be how the long-lived hydrogen bonds and the protected hydrogen bonds affect the binding affinity? Another might be what kind of interactions produces a good affinity in lipophilic and hydrophilic pockets?

References

Abdel-Meguid S.S., Metcalf B.W., Carr T.J., Demarsh P., DesJarlais R.L., Fisher S., Green D.W., Ivanoff L., Lambert D.M., Murthy K.H.M., Petteway Jr. S.R., Pitts W.J., Tomaszek Jr. T.A., Winborne E., Zhao B., Dreyer G.B., Meek T.D.: An orally bioavailable HIV-1 protease inhibitor containing an imidazole-derived peptide bond replacement: crystallographic and pharmacokinetic analysis. *Biochemistry* 33: 11671–11677, 1994.

Agniswamy J., Fang B., Weber I.T.: Plasticity of S2-S4 specificity pockets of executioner caspase-7 revealed by structural and kinetic analysis. *Journal* 274: 4752–4765, 2007.

Aita T., Nishigaki K., Husimi Y.: Toward the fast blind docking of a peptide to a target protein by using a four-body statistical pseudo-potential. *Comput Biol Chem.* 34(1): 53-62, 2010.

Aloy P., Querol E., Aviles F.X., Sternberg M.J.E.: Automated Structure-based Prediction of Functional Sites in Proteins: Applications to Assessing the Validity of Inheriting Protein Function from Homology in Genome Annotation and to Protein Docking. *J. Mol. Biol.* 311(2): 395–408, 2001.

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17): 3389–3402, 1997.

Altschul S.F., Wootton J.C., Gertz E.M., Agarwala R., Morgulis A., Schaffer A.A., Yu Y.K.: Protein database searches using compositionally adjusted substitution matrices. *FEBS J.* 272(20): 5101–5109, 2005.

An J., Totrov M., Abagyan R.: Comprehensive Identification of “Druggable” Protein Ligand Binding Sites. *Genome Informatics* 15(2): 31–41, 2004.

An J., Totrov M., Abagyan R.: Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes. *Mol. Cell. Proteomics* 4(6): 752–761, 2005.

Armon A., Graur D., Ben-Tal N.: ConSurf: An Algorithmic Tool for the Identification of Functional Regions in Proteins by Surface Mapping of Phylogenetic Information. *J. Mol. Biol.* 307(1): 447–463, 2001.

Arnold K., Kiefer F., Kopp J., Battey J.N., Podvinec M., Westbrook J.D., Berman H.M., Bordoli L., Schwede T.: The Protein Model Portal. *J. Struct. Funct. Genomics* 10: 1–8, 2009.

Ashford P., Moss D.S., Alex A., Yeap S.K., Povia A., Nobeli I., Williams M.A.: Visualisation of variable binding pockets on protein surfaces by probabilistic analysis of related structure sets. *BMC Bioinformatics* 13: 39, 2012.

Ashkenazy H., Erez E., Martz E., Pupko T. and Ben-Tal N.: ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* 38(Web Server issue): W529–W533, 2010.

Baldwin E. T., Bhat T. N., Gulnik S., Liu B., Topol I. A., Kiso Y., Minoto T., Mitsuya H., Erickson J. W.: Structure of HIV-1 protease with KNI-272, a tight-binding transition-state analog containing allophenylnorstatine. *Structure* 3(6): 581–590, 1995.

Ballester P.J. and Mitchell J.B.O.: A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* 26(9): 1169–1175, 2010.

Barber C.B., David P.D.,Huhdanpää H.: The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw*. 22: 469–483, 1996.

Bate P. and Warwicker J.: Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. *J. Mol. Biol.* 340(2): 263–276, 2004.

Benson M., Smith R., Khazanov N., Dimcheff B., Beaver J., Dresslar P., Nerothin J., Carlson H.: Binding MOAD, a high-quality protein–ligand database. *Nucleic Acids Res.* 36: 674–678, Oxford Journal 2008.

Berendsen H.J.C., Postma J.P.M., Gunsteren W.F.v. Hermans J.: In Intermolecular Forces. D. Reidel Publishing Company, Dordrecht, 1981.

Berendsen H.J.C., Grigera J.R., Straatsma T.P. The missing term in effective pair potentials. *J. Phys. Chem.* 91: 6269–6271, 1987.

Berezin C., Glaser F., Rosenberg J., Paz I., Pupko T., Fariselli P., Casadio R., Ben-Tal N.: ConSeq: The Identification of Functionally and Structurally Important Residues in Protein Sequences. *Bioinformatics* 20: 1322–1324, 2004.

Berman H., Westbrook J., Feng Z., Gilliland G., Bhat T., Weissig H., Shindyalov I., Bourne P.: The Protein Data Bank. *Nucleic Acids Res.* 28: 235–242, Oxford University Press 2000.

Bernstein F, Koetzle T, Williams G, Meyer E, Brice M, Rodgers J, Kennard O, Shimanouchi T, Tasumi M: The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112: 535–542, 1977

Bhaumik D., Medin J., Gathy K., Coleman M.S.: Mutational analysis of active site residues of human adenosine deaminase. *J. Biol. Chem.* 268(8): 5464–5470, 1993.

Bhinge A., Chakrabarti P., Uthanumallian K., Bajaj K., Chakraborty K., Varadarajan R.: Accurate Detection of Protein: Ligand Binding Sites Using Molecular Dynamics Simulations. *Structure* 12(11): 1989–1999, 2004.

Binkowski T.A., Naghibzadeh S., Liang J.: CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Res.* 31(13): 3352–3355, 2003.

Bliznyuk A.A. and Gready J.E.: A New Approach to Estimation of the Electrostatic Components of the Solvation Energy in Molecular Mechanics Calculations. *J. Phys. Chem.* 99(39): 14506–14513, 1995.

Bliznyuk A.A. and Gready J.E.: Identification and energetic ranking of possible docking sites for pterin on dihydrofolate reductase. *J. Comput. Aided Mol. Des.* 12: 325–333, 1998.

Bliznyuk A.A. and Gready J.E.: Simple Method for Locating Possible Ligand Binding Sites on Protein Surfaces. *J. Comput. Chem. 20:* 983–988, 1999.

Blundell T.L. and Patel S.: High-throughput X-ray crystallography for drug discovery. *Curr. Opin. Pharmacol.* 4(5): 490–496, 2004.

Boutet E., Lieberherr D., Tognolli M., Schneider M., Bairoch A.: UniProtKB/Swiss-Prot. *Methods Mol. Biol.* 406, 89–112, 2007.

Brady G.P. Jr, Stouten P.F.: Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.* 14(4): 383–401, 2000.

Bray T., Chan P., Bougouffa S., Greaves R., Doig. A.J., Warwicker J.: SitesIdentify: a protein functional site prediction tool. *BMC Bioinformatics* 10: 379, 2009.

Breiman L.: Bagging predictors. *Machine Learning* 24:123–140, 1996.

Breiman L.: Random Forests. *Machine Learning* 45: 5–32, 2001.

Brooks B.R., Bruccoleri R.E., Olafson B.D., States D.J., Swaminathan S., Karplus M.: CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4: 187–217, 1983.

Brown D. and Superti-Furga G.: Rediscovering the sweet spot in drug discovery. *Drug Discov. Today* 8(23): 1067–1077, 2003.

Bryliński M., Prymula K., Jurkowski W., Kochańczyk M., Stawowczyk E., Konieczny L., Roterman I.: Prediction of functional sites based on the fuzzy oil drop model. *PLoS Comput. Biol.* 3(5): e94, 2007a.

Bryliński M, Kochańczyk M, Broniatowska E, Roterman I.: Localization of ligand binding site in proteins identified in silico. *J. Mol. Model.* 13(6–7): 665–75, 2007b.

Bryliński M. and Skolnick J.: A threading-based method (FINDSITE) for ligandbinding site prediction and functional annotation. *Proc. Natl. Acad. Sci. U.S.A.* 105(1): 129–134, 2008.

Bryliński M. and Skolnick J.: FINDSITE^{LHM}: A Threading-Based Approach to Ligand Homology Modeling. *PLoS Comput. Biol.* 5(6): e1000405, 2009.

Böhm H.J. and Klebe G.: What Can We Learn from Molecular Recognition in Protein-Ligand Complexes for the Design of New Drugs? *Angew. Chem. Int. Ed.* 35(22): 2588–2614, 1996.

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

Capra J.A., Laskowski R.A., Thornton J.M., Funkhouser T.A.: Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure. *PLoS Comput. Biol.* 5(12): e1000585, 2009.

Carlson H.A., Smith R.D., Khazanov N.A., Kirchhoff P.D., Dunbar J.B. Jr., Benson M.L.: Differences between High- and Low-Affinity Complexes of Enzymes and Nonenzymes. J. Med. Chem. 51: 6432–6441, 2008.

Casari G., Sander C., Valencia A.: A method to predict functional residues in proteins. *Nat. Struct. Biol.* 2(2): 171–178, 1995.

Chakravarty S., Bhinge A., Varadarajan R.: A procedure for detection and quantitation of cavity volumes proteins. Application to measure the strength of the hydrophobic driving force in protein folding. *J. Biol. Chem.* 277(35): 31345–31353, 2002.

Chang Z., Nygaard P., Chinault A.C., Kellems R.: Deduced Amino Acid Sequence of *Escherichia coli* Adenosine Deaminase Reveals Evolutionarily Conserved Amino Acid Residues: Implications for Catalytic Function. *Biochemistry* 30(8): 2273–2280, 1991.

Chang D.T., Oyang Y.J., Lin J.H.: MEDock: a web server for efficient prediction of ligand binding sites based on a novel optimization algorithm. *Nucleic Acids Res.* 33(Web Server issue): W233–W238, 2005.

Chelliah V., Chen L., Blundell T.L., Lovell S.C.: Distinguishing Structural and Functional Restraints in Evolution in Order to Identify Interaction Sites. *J. Mol. Biol.* 342(5): 1487–1504, 2004.

Cheng A.C., Coleman R.G., Smyth K.T., Cao Q., Soulard P., Caffrey D.R., Salzberg A.C., Huang E.S.: Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* 25: 71–75, 2007.

Coleman R.G. and Sharp K.A.: Travel Depth, a New Shape Descriptor for Macromolecules: Application to Ligand Binding. *J. Mol. Biol.* 362: 441–458, 2006.

Coleman R.G. and Sharp K.A.: Protein Pockets: Inventory, Shape, and Comparison. *J. Chem. Inf. Model.* 50: 589–603, 2010.

Collins, J. R., Burt, S. K., and Erickson, J. W.: Flap opening in HIV-1 protease simulated by activated molecular dynamics. *Nat. Struct. Biol.* 2: 334–338, 1995.

Congreve M, Carr R, Murray C, Jhoti H: A 'Rule of Three' for fragment-based lead discovery? *Drug Discov. Today* 8: 876–877, 2003.

Connolly M.L.: Analytical molecular surface calculation. *J. Appl. Cryst.* 16: 548–558, 1983.

Connolly M.L.: Measurement of protein surface shape by solid angles. *J. Mol. Graphics* 4: 4–6, 1986.

Cornell W.D., Cieplak P., Bayly C.I., Gould I.R., Merz K.M., Ferguson D.M., Spellmeyer D.C., Fox T., Caldwell J.W., Kollman P.A.: A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* 117(19): 5179–5197, 1995.

Craig D., Krammer A., Schulten K., Vogel V.: Comparison of the early stages of forced unfolding for fibronectin type III modules. *Proc. Natl. Acad. Sci.* 98: 5590–5595, 2001.

Craig D., Gao M., Schulten K., Vogel V.: Tuning the Mechanical Stability of Fibronectin Type III Modules through Sequence Variations. *Structure* 12(1): 21–30, 2004a.

Craig D., Gao M., Schulten K., Vogel V. Structural Insights into How the MIDAS Ion Stabilizes Integrin Binding to an RGD Peptide under Force. *Structure* 12(11): 2049–2058, 2004b.

Dai T., Liu Q., Gao J., Cao Z., Zhu R.: A new protein-ligand binding sites prediction method based on the integration of protein sequence conservation information. *BMC Bioinformatics* 12(Suppl. 14): S9, 2011.

Dean A.M. and Golding G.B.: Enzyme evolution explained (sort of). *Pac. Symp. Biocomput.* : 6–17, 2000.

de Berg M., van Kreveld M., Overmars M., Schwarzkopf O.: Delaunay Triangulations-Height Interpolation. *In Computational Geometry: Algorithms and Applications.* M. de Berg, M. van Krefeld, M. Overmars, and O. Schwarzkopf, eds. (New York: Springer), pp. 183–210, 2000.

Delaney J.S.: Finding and filling protein cavities using cellular logic operations. *J. Mol. Graph.* 10(3): 174–177, 1992

Del Carpio C.A., Takahashi Y., Sasaki S.: A new approach to the automatic identification of candidates for ligand receptor sites in proteins: (I). Search for pocket regions. *J. Mol. Graph.* 11(1): 23–29, 1993.

del Sol Mesa A., Pazos F., Valencia A.: Automatic Methods for Predicting Functionally Important Residues. *J. Mol. Biol.* 326(4): 1289–1302, 2003.

Drews J.: Genomic sciences and the medicine of tomorrow. *Nat. Biotechnol.* 14(11): 1516–1518, 1996.

Drews J.: Drug discovery: A historical perspective. *Science* 287(5460), 1960–1964, 2000.

DesJarlais R.L., Sheridan R.P., Seibel G.L., Dixon J.S., Kuntz I.D., Venkataraghavan R.: Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem.* 31(4): 722-729, 1988.

Dessailly B.H., Lensink M.F., Orengo C.A., Wodak S.J.: LigASite–a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res.* 36(Database issue): D667–D673, 2008.

Doolittle R.F. and Feng D.F.: Nearest neighbor procedure for relating progressively aligned amino acid sequences. *Meths. Enzymol.* 183: 659–669, 1990.

Dundas J., Ouyang Z., Tseng J., Binkowski A., Turpaz Y., Liang J.: CASTp: computed atas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.* 34(Web Server issue): W116–W118, 2006.

Edelsbrunner H. and Mücke E.P.: Three-dimensional alpha shapes. *ACM Trans. Graph.* 13: 43–72, 1994.

Edelsbrunner H.: The union of balls and its dual shape. *Discrete Comput Geom* 13: 415–440, 1995.

Edelsbrunner H., Facello M., Fu P., Liang J.: Measuring proteins and voids in proteins. *28th Hawaii International Conference on System Sciences (HICSS'95)*: 256–264, 1995.

Edelsbrunner H. and Shah N.R.: Incremental topological flipping works for regular triangulations. *Algorithmica* 15: 223–241, 1996.

Edelsbrunner H., Facello M., Liang J.: On the definition and the construction of pockets in macromolecules. *Pac. Symp. Biocomput.*: 272–287, 1996.

Egner U. and Hillig R.C.: A structural biology view of target drugability. *Expert Opin. Drug Discov.* 3(4): 391–401, 2008.

Eisenhaber F., Lijnzaad P., Argos P., Sander C., Scharf M.: The double cubic lattice methos: efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *J. Comput. Chem.* 16: 273–284, 1995.

Elcock A.H.: Prediction of Functionally Important Residues Based Solely on the Computed Energetics of Protein Structure. *J. Mol. Biol.* 312: 885–896, 2001.

Facello M.A.: Implementation of a randomized algorithm for Delaunay and regular triangulations in three dimensions. *Comput. Aided Geomet. Design* 12: 349–370, 1995.

Felsenstein J.: Evolutionary Trees from DNA Sequences: A Maximum Likehood Approach. *J. Mol. Evol.* 17: 368–376, 1981.

Foley J.D. and Van Dam A., Fundamentals of Interactive Computer Graphics, Addison-Wesley, Reading, MA, 1982.

Freedberg D.I., Ishima R., Jacob J., Wang Y.X., Kustanovich I., Louis J.M., and Torchia D.A.: Rapid structural fluctuations of HIV-1 Protease flaps in solution: Relationship to crystal structures and comparison with predictions of dynamic calculations. *Protein Sci.* 11: 221–232, 2002.

Friedman N., Ninio M., Pe'er I., Pupko T.: A Structural EM Algorithm for Phylogenetic Inference. *J. Comput. Biol.* 9(2): 331–353, 2002.

Fukunishi Y. and Nakamura H.: Prediction of ligand-binding sites of proteins by molecular docking calculation for a random ligand library. *Protein Sci.* 20(1): 95–106, 2011.

Furnham N., Laskowski R.A., Thornton J.M.: Abstracting Knowledge from the Protein Data Bank. *Biopolymers.* 99(3): 183–188, 2012.

Gao J., Bosco D.A., Powers E.T., Kelly J.W.: Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. *Nature Structural & Molecular Biology* 16(7): 684–691, 2009.

Gao M., Wilmanns M., Schulten K.: Steered molecular dynamics studies of titin I1 domain unfolding. *Biophysical journal* 83(6): 3435–3445, 2002

Ghersi D. and Sanchez R.: Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites. *Proteins* 74(2): 417–424, 2009a.

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

Ghose A, Viswanadhan V, Wendoloski J: A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* 1(1): 55–68, 1999.

Ghosh A.K., Gemma S., Takayama J., Baldridge A., Leshchenko-Yashchuk S., Miller H.B., Wang Y.F., Kovalevsky A.Y., Koh Y., Weber I.T., Mitsuya H.: Potent HIV-1 protease inhibitors incorporating meso-bicyclic urethanes as P2-ligands: structure-based design, synthesis, biological evaluation and protein-ligand X-ray studies. *Org. Biomol. Chem.* 6: 3703–3713, 2008.

Glaser F., Pupko T., Paz I., Bell R.E., Bechor-Shental D., Martz E., Ben-Tal N.: ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. *Bioninformatics* 19(1): 163–164, 2003.

Glaser F., Rosenberg Y., Kessel A., Pupko T., Ben-Tal N.: The ConSurf-HSSP Database: The Mapping of Evolutionary Conservation Among Homologs Onto PDB Structures. *Proteins* 58(3): 610–617, 2005.

Glaser F., Morris R.J., Najmanovich R.J., Laskowski R.A., Thorton J.M.: A Method for Localizing Ligand Binding Pockets in Protein Structures. *Proteins* 62(2): 479–488, 2006.

Greaves R. and Warwicker J.: Active site identification through geometry-based and sequence profile-based calculations: burial of catalytic clefts. *J. Mol. Biol.* 349(3): 547–557, 2005.

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

Grosdidier A., Zoete V., Michielin O.: EADock: Docking of Small Molecules into Protein Active Sites with a Multiobjective Evolutionary Optimization. *Porteins* 67(4): 1010–1025, 2007.

Grzesiek S., Bax A., Nicholson L.K., Yamazaki T., Wingfield P., Stahl S.J., Eyermann C.J.E., Torchia D.A., Hedge C.N., Lam P.Y.S., Jadhav P.K., Chang C.-H.: NMR Evidence for the Displacement of a Conserved Interior Water Molecule in HIV Protease by a Non-Peptide Cyclic Urea-Based Inhibitor. *J. Am. Chem. Soc.* 116: 1581–1582, 1994.

Gu J., Qiu Z., Wang X.: Identification of Ligand Binding Pockets for Unbound State Proteins Using Local Residue Preference, p. 14–18. International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012), Singapore 2012.

Gunasekaran K. and Nussinov R.: How Different are Structurally Flexible and Rigid Binding Sites? Sequence and Structural Features Discriminating Proteins that Do and Do not Undergo Conformational Change upon Ligand Binding. *J. Mol. Biol.* 365: 257–273, 2007.

Halgren T.: New method for fast and accurate binding-site identification and analysis. *Chem. Biol. Drug Des.* 69: 146–148, 2007.

Halgren T: Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model* 49: 377–389, 2009.

Hamelberg D., Mongan J., McCammon J.A.: Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys.* 120(24): 11919–11929, 2004.

Hannenhalli S.S. and Russell R.B.: Analysis and Prediction of Functional Sub-types from Protein Sequence Alignments. *J. Mol. Biol.* 303(1): 61–76, 2000.

Harris R., Olson A.J., Goodsell D.S.: Automated prediction of ligand-binding sites in proteins. *Proteins* 70: 1506–1517, 2008.

Harrison R.W., Kourinov I.V., Andrews L.C.: The Fourier-Green's function and the rapid evaluation of molecular potentials. *Protein Eng.* 7(3): 359–369, 1994.

Hartshorn M.J., Verdonk M.L., Chessari G., Brewerton S.C., Mooij W.T.M., Mortenson P.N., Murray C.W.: Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* 50: 726–741, 2007.

Hawkins J.C., Zhu H., Teyra J., Pisabarro M.T.: Reduced False Positives in PDZ Binding Prediction Using Sequence and Structural Descriptors. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9(5): 1492–1503, 2012.

Heaslet H., Rosenfeld R., Giffin M., Lin Y.C., Torbett B.E., Elder J.H., McRee D.E., Stout C.D.: Conformational flexibility in the flap domains of ligand-free HIV-1 protease. *Acta Crystallogr.* D63: 866–875, 2007.

Hendlich M., Rippmann F., Barnickel G.: LIGSITE-The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Graph. Model. 15:* 359–363, 389, 1997.

Hendrix D.K. and Kuntz I.D.: Surface solid angle-based site points for molecular docking. *Pac. Symp. Biocomput.*: 317–326, 1998.

Hernandez M., Ghersi D., Sanchez R.: SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res.* 37(Web Server issue): W413–W416, 2009.

Hetényi C. and van der Spoel D.: Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci.* 11(7): 1729–1737, 2002.

Hetényi C. and van der Spoel D.: Blind docking of drug-sized compounds to proteins with up to a thousand residues. *FEBS Lett.* 580(5): 1447–1450, 2006.

Hetényi C. and van der Spoel D.: Toward prediction of functional protein pockets using blind docking and pocket search algorithms. *Protein Sci.* 20(5): 880–893, 2011.

Ho C.M. and Marshall G.R.: Cavity search: an algorithm for the isolation and display of cavity-like binding regions. *J. Comput. Aided Mol. Des.* 4(4): 337–354, 1990.

Honig B., Sharp K.A., Yang A.S.: Macroscopic Models of Aqueous Solutions: Biological and Chemical Applications. *J. Phys. Chem.* 97: 1101–1109, 1993.

Honig B. and Nicholls A.: Classical electrostatics in biology and chemistry. *Science* 268: 1144–1149, 1995.

Hornak V., Okur A., Rizzo R.C., Simmerling C.: HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* 103: 915–920, 2006.

Hornak V. and Simmerling C.: Targeting structural flexibility in HIV-1 protease inhibitor binding. *Drug Discov. Today* 12: 132–138, 2007.

Hu L, Benson M, Smith R, Lerner M, Carlson H: Binding MOAD (Mother of All Databases). *Proteins Bioinformatics* 66: 333–340, 2005.

Huang B. and Schroeder M.: LIGSITE^{csc}- predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol.* 6: 19, 2006.

Huang B.: MetaPocket: A Meta Approach to Improve Protein Ligand Binding Site Prediction. *OMICS* 13(4): 325–330, 2009.

Huang N., Nagarsekar A., Xia G., Hayashi J., MacKerell A.D. Jr.: Identification of non-phosphate-containing small molecular weight inhibitors of the tyrosine kinase p56 Lck SH2 domain via in silico screening against the pY + 3 binding site. *J. Med. Chem.* 47: 3502–3511, 2004.

Hubbard T.J. and Blundell T.L.: Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng.* 1(3): 159–171, 1987.

Innis C.A., Anand A.P., Sowdhamini R.: Prediction of Functional Sites in Proteins Using Conserved Functional Group Analysis. *J. Mol. Biol.* 337: 1053–1068, 2004.

Innis C.A.: siteFiNDER|3D: a web-based tool for predicting the location of functional sites in proteins. *Nucleic Acids Res.* 35(Web server issue): W489–W494, 2007.

Ishima R., Freedberg D., Wang Y.X., Louis J.M., and Torchia D.A.: Flap opening and dimmer-interface flexibility in the free and inhibitor bound HIV protease, and their implications for function. *Structure* 7: 1047–1055, 1999.

Ivanisenko V.A., Pintus S.S., Grigorovich D.A., Kolchanov N.A.: PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res.* 33: D183–D187, 2005.

Jackson R.M.: Q-fit: a probabilistic method for docking molecular fragments by sampling low energy conformational space. *J. Comput. Aided Mol. Des.* 16: 43–57, 2002.

Jain A.N.: Scoring non-covalent ligand-protein interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput. Aided Mol. Des.* 10: 427–440, 1996.

Jorgensen W.L. and Tirado-Rives J.: The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* 110(6): 1657–1666, 1988.

Kalidas Y. and Chandra N.: PocketDepth: a new depth based algorithm for identification of ligand binding sites in proteins. *J. Struct. Biol.* 161(1): 31–42, 2008.

Katoh E., Louis D.A., Yamazaki T., Gronenborn A.M., Torchia D.A., Ishima R.: A solution NMR study of the binding kinetics and the internal dynamics of an HIV-1 protease-substrate complex. *Protein Sci.* 12: 1376–1385, 2003.

Kawabata T. and Go N.: Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. *Proteins* 68: 516–529, 2007.

Kawabata T.: Detection of multi-scale pockets on protein surfaces using mathematical morphology. *Proteins* 78: 1195–1121, 2010.

Keil M., Exner T.E., Brickmann J.: Pattern Recognition Strategies for Molecular Surfaces: III. Binding Site Prediction with a Neural Network. *J. Comput. Chem.* 25(6): 779–789, 2004.

Keller T, Pichota A, Yin Z: A practical view of ‘druggability’. Current opinion in chemical biology 10(4): 357–361, 2006.

Kellogg G.E., Fornabaio M., Cheng D.L., Abraham D.J.: New application design for a 3D hydropathic map-based search for potential water molecules bridging between protein and ligand. *Internet Electron. J. Mol. Des.* 4(3): 194–209, 2005.

Kim D., Cho C.H., Cho Y., Ryu J., Bhak J., Kim D.S.: Pocket extraction on proteins via the Voronoi diagram of spheres. *J. Mol. Graph Model.* 26(7): 1104–1112, 2008.

Kleywegt G.J.: VOIDOO manual, Updated 2008. Checked from internet 14.01.2013: http://xray.bmc.uu.se/usf/voidoo_man.html

Kleywegt G.J. and Jones T.A.: Detection, Delineation, Measurement and Display of Cavities in Macromolecular Structures. *Acta Cryst.* D50: 178–185, 1994.

Knox C., Law V., Jewison T., Liu P., Ly S., Frolkis A., Pon A., Banco K., Mak C., Neveu V., Djoumbou Y., Eisner R., Guo A.C., Wishart D.S.: DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* 39(Database issue): D1035–D1041, 2011.

Kortvelyesi T., Silberstein M., Dennis S., Vajda S.: Improved mapping of protein binding sites. *J. Comput. Aided Mol. Des.* 17(2–4): 173–186, 2003.

Kožíšek M., Bray J., Řezáčová P., Šašková K., Brynda J., Pokorná J., Mammano F., Rulíšek L., Konvalinka J.: Molecular analysis of the HIV-1 resistance development: enzymatic activities, crystal structures, and thermodynamics of nelfinavir-resistant HIV protease mutants. *J. Mol. Biol. 374:* 1005–1016, 2007.

Krasowski A., Muthas D., Sarkar A., Schmitt S., Brenk R.: DrugPred: A Structure-Based Approach To Predict Protein Druggability Developed Using an Extensive Nonredundant Data Set. *J. Chem. Inf. Model.* 51(11): 2829–2842, 2011.

Kullback S. and Leibler A.B.: On Information and Sufficiency. *Ann. Math. Statist.* 22(1): 79–86, 1951.

Kuntz I.D., Blaney J.M., Oatley S.J., Langridge R., Ferrin T.E.: A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* 161: 269–288, 1982.

Kuntz I.D., Chen K., Sharp K.A., Kollman P.A.: The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U.S.A.* 96(18): 9997–10002, 1999.

Lam P. Y., Jadhav P. K., Eyermann C. J., Hodge C. N., Ru Y., Bacheler L. T., Meek J. L., Otto M. J., Rayner M. M., Wong Y. N., Chang C.-H., Weber P. C., Jackson D. A., Sharpe T. R., Erickson-Viitanen S.: Rational Design of Potent, Bioavailable, Nonpeptide Cyclic Ureas as HIV Protease Iinhibitors. *Science* 263: 380–384, 1994.

Landau M., Mayrose I., Rosenberg Y., Glaser F., Martz E., Pupko T., Ben-Tal N.: ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* 33(Web Server Issue): W299–W302, 2005.

Landgraf R., Fischer D., Eisenberg D.: Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng.* 12(11): 943–951, 1999.

Landgraf R., Xenarios I., Eisenberg D.: Three-dimensional Cluster Analysis Identifies Interfaces and Functional Residue Clusters in Proteins. *J. Mol. Biol.* 307(5): 1487–1502, 2001.

Lapatto R., Blundell T., Hemmings A., Overington J., Wilderspin A., Wood S., Merson J.R., Whittle P.J., Danley D.E., Geoghegan K.F., Hawrylik S.J., Lee S.E., Scheld K.G., Hobart P.M.: X-Ray analysis of HIV-1 Proteinase at 2.7Å resolution confirms structural homology among retroviral enzymes. *Nature* 342(6247): 299–302, 1989.

Laskowski R.A.: SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph*. 13: 323–330, 1995.

Laskowski R.A., Luscombe N.M., Swindells M.B., Thornton J.M.: Protein clefts in molecular recognition and function. *Protein Sci.* 5: 2438–2452, 1996.

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

Laurie A.T.R. and Jackson R.M.: Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Curr. Protein Pept. Sci.* 7(5): 395–406, 2006.

Lee B. and Richards F.M.: The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55: 379–400, 1971.

Le Guilloux V., Schmidtke P., Tuffery P.: Fpocket: an open source platform for ligand pocket detection. *BMC Bioinf.* 10: 168–179, 2009.

Levitt D.G. and Banaszak L.J.: POCKET: A computer graphies method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.* 10(4): 229–234, 1992.

Li B., Turuvekere S., Agrawal M., La D., Ramani K., Kihara D.: Characterization of local geometry of protein surfaces with the visibility criterion. *Proteins* 71: 670–683, 2008.

Liang J., Edelsbrunner H., Fu P., Sudhakar P.V., Subramaniam S.: Analytical Shape Computation of Macromolecules: I. Molecular Area and Volume Through Alpha Shape. *PROTEINS: Structure, Function, and Genetics* 33: 1–17, 1998a.

Liang J., Edelsbrunner H., Woodward C.: Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Science* 7: 1884–1897, 1998b.

Lichtarge O., Bourne H.R., Cohen F.E.: An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families. *J. Mol. Biol.* 257(2): 342–358, 1996.

Lichtarge O. and Sowa M.E.: Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* 12(1): 21–27, 2002.

Lindow N., Baum D., Hege H.C.: Voronoi-Based Extraction and Visualization of Molecular Paths. *IEEE Transactions on Visualization and Computer Graphics* 17(12): 2025–2034, 2011.

Lipinski C., Lombardo F., Dominy B., Feeney P.: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews 23: 3–25, 1997.

Liu F., Kovalevsky A.Y., Tie Y., Ghosh A.K., Harrison R.W., Weber I.T.: Effect of flap mutations on structure of HIV-1 protease and inhibition by saquinavir and darunavir. J.Mol.Biol. 381: 102–115, 2008.

Lo Y.T., Wang H.W., Pai T.W., Tsou W.S., Hsu H.H., Chang H.T.: Protein-ligand binding region prediction (PLB-SAVE) based on geometric features and CUDA acceleration. BMC Bioinformatics 14(Suppl. 4): S4, 2013.

Lu H., Schulten K.: The key event in force-induced unfolding of Titin's immunoglobulin domains. Biophysical journal 79(1): 51–65, 2000.

Magrane M. and the UniProt consortium: UniProt Knowledgebase: a hub of integrated protein data. *Database* 2011: bar009, 2011.

Marr D. and Hildreth E.: Theory of edge detection. *Proc. R. Soc. London B Biol. Sci.* 207: 187–217, 1980.

Masuya M. and Doi T.: Detection and geometric modeling of molecular surfaces and cavities using digital mathematical morphological operations. *J. Mol. Graph.* 13: 331–336, 1995.

Mayrose I., Graur D., Ben-Tal N., Pupko T.: Comparison of Site-Specific Rate-Inference Methods for Protein Sequences: Empirical Bayesian Methods Are Superior. *Mol. Biol. Evol.* 21(9): 1781–1791, 2004.

Miller M., Schneider J., Sathyanarayana B.K., Toth M.V., Marshall G., Clawson L., Selk L., Kent S.B.H., Wlodaver A.: Structure of a complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3Å resolution. *Science* 246(4934): 1149–1152, 1989.

Ming D. and Wall M.E.: Quantifying allosteric effects in proteins. *Proteins* 59(4): 697–707, 2005.

Ming D. and Wall M.E.: Interactions in native binding sites cause a large change in protein dynamics. *J. Mol. Biol*. 358(1): 213–23, 2006.

Ming D., Cohn J.D., Wall M.E.: Fast dynamics perturbation analysis for prediction of protein functional sites. *BMC Struct. Biol.* 8: 5, 2008.

Minke W.E., Diller D.J., Hol W.G.J., Verlinde C.L.M.J.: The Role of Waters in Docking Strategies with Incremental Flexibility for Carbohydrate Derivatives: Heat-Labile Enterotoxin, a Multivalent Test Case. *J. Med. Chem.* 42(10): 1778–1788, 1999.

Mohamedali K.A., Kurz L.C., Rudolph F.B.: Site-Directed Mutagenesis of Active Site Glutamate-217 in Mouse Adenosine Deaminase. *Biochemistry* 35: 1672–1680, 1996.

Molecular Operating Environment (MOE), 2012.10; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2012.

Morgan D.H., Kristensen D.M., Mittelman D., Lichtarge O.: ET viewer: an application for predicting and visualizing functional sites in protein structures. *Bioinformatics* 22(16): 2049–2050, 2006.

Morita M., Nakamura S., Shimizu K.: Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures. *Proteins* 73(2): 468–479, 2008.

Morris G.M., Goodsell D.S., Halliday R.S., Huey R., Hart W.E., Belew R.K., Olson A.J.: Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comp. Chem.* 19: 1639–1662, 1998.

Muchmore S.W. and Hajduk P.J.: Crystallography, NMR and virtual screening: integrated tools for drug discovery. *Curr. Opin. Drug Discov. Devel.* 6(4): 544–549, 2003.

Murzin A.G., Brenner S.E., Hubbard T., Chothia C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247(4): 536–540, 1995.

Navia M.A., Fitzgerald P.M.D., McKeever B.M., Leu C.-T., Heimbach J.C., Herber W.K., Sigal I.S., Darke P.L., Springer J.P.: Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature* 337(61): 5–20, 1989.

Nayal M. and Honig B.: On the nature of cavities on protein surfaces Application to the identification of drug-binding sites. *Proteins* 63(4): 892–906, 2006.

Nicholls A., Sharp K.A., Honig B.: Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* 11(4): 281–296, 1991.

Nicholsan L.K., Yamazaki T., Torchia D.A., Grzesiek S., Bax A., Stahl S.J., Kaufman J.D., Wingfield P.T., Lam P.V.S., Jadhav P.K., Hodge C.N., Domaille P.J., Chang C.H.: Flexibility and function in HIV-1 protease. *Nat. Struct. Biol.* 2: 274–280, 1995.

Nilsson L, Thomas W, Sokurenko E, Vogel V: Beyond Induced-Fit Receptor-Ligand Interactions: Structural Changes that Can Significantly Extend Bond Lifetimes. *Structure* 16: 1047–1058, 2008.

Nimrod G., Glaser F., Steinberg D., Ben-Tal N., Pupko T.: In silico identification of functional regions in proteins. *Bioinformatics* 21(Suppl 1.): i328-i337, 2005.

Nimrod G., Schushan M., Steinberg D.M., Ben-Tal N.: Detection of Functionally Important Regions in "Hypothetical Proteins" of Known Structure. *Structure* 16(12):1755-1763, 2008.

Nissink J.W., Murray C., Hartshorn M., Verdonk M.L., Cole J.C., Taylor R.: A new test set for validating predictions of protein-ligand interaction. *Proteins* 49(4): 457–471, 2002.

Pan Y., Huang N., Cho S., MacKerell A.D. Jr.: Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.* 43: 267–272, 2003.

Pearlman D.A., Case D.A., Caldwell J.W., Ross W.R., Cheatham T.E., DeBolt S., Ferguson D., Seibel G., Koll-man P.: AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Comp. Phys. Commun.* 91: 1–41, 1995.

Pern Tan K., Varadarajan R., Madhusudhan M.S.: DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. *Nucl. Acids Res.* 39(Suppl. 2): W242-W248, 2011.

Perola E., Walters W.P., Charifson P.S.: A Detailed Comparison of Current Docking and Scoring Methods on Systems of Pharmaceutical Relevance. *Proteins* 56(2): 235–249, 2004.

Perola E., Herman L., Weiss J.: Development of a Rule-Based Method for the Assessment of Protein Druggability. *J. Chem. Inf. Model.* 52: 1027−1038, 2012.

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

Perryman A.L., Lin J.H., McCammon J.A.: Restrained molecular dynamics simulations of HIV-1 protease: The first step in validating a new target for drug design. *Biopolymers* 82: 272–284, 2006.

Peters K.P., Fauck J., Frömmel C.: The Automatic Search for Ligand Binding Sites in Proteins of Known Three-dimensional Structure Using only Geometric Criteria. *J. Mol. Biol.* 256(1): 201–213, 1996.

Pettit F.K., Bare E., Tsai A., Bowie J.U.: HotPatch: A Statistical A pproach to Finding Biologically Relevant Features on Protein Surfaces. *J. Mol. Biol.* 369(3): 863–879, 2007.

Pierce L.C., Salomon-Ferrer R., Augusto F. de Oliveira C., McCammon J.A., Walker R.C.: Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. *J. Chem. Theory Comput.* 8(9): 2997–3002, 2012.

Pillai B., Kannan K.K., Hosur M.V.: 1.9Å X-ray study shows closed flap conformation in crystals of tethered HIV-1 PR. *Proteins* 43(1): 57–64, 2001.

Prlić A., Yates A., Bliven S.E., Rose P.W., Jacobsen J., Troshin P.V., Chapman M., Gao J., Hock Koh C., Foisy S., Holland R., Rimsa G., Heuer M.L., Brandstatter-Muller H., Bourne P.E., Willis S.: BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics* 20: 2693–2695, 2012.

Punta M., Coggill P.C., Eberhardt R.Y., Mistry J., Tate J., Boursnell C., Pang N., Forslund K., Ceric G., Clements J., Heger A., Holm L., Sonnhammer E.L.L., Eddy S.R., Bateman A., Finn R.D.: The Pfam protein families database. *Nucleic Acids Research* Database Issue 40: D290–D301, 2012.

Pupko T., Bell R.E., Mayrose I., Glaser F., Ben-Tal N.: Rate4Site: an Algotirthmic tool for the indentification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioninformatics* 18 Suppl. 1: S71–S77, 2002.

Qiu Z. and Wang X.: Identification of ligand-binding pockets in proteins using residue preference methods. *Protein Pept. Lett.* 16(8): 984–990, 2009.

Requicha A.A.G. and Voelcker H.B.: Solid Modeling: A Historical Summary and Contemporary Assessment. *IEEE Computer Graphics and Applications* 2: 9, 1982.

de Rinaldis M., Ausiello G., Cesareni G., Helmer-Citterich M.: Three-dimensional Profiles: A New Tool to Identify Protein Surface Similarities. *J. Mol. Biol.* 284(4): 1211–1221, 1998.

Rogers D.F.: Procedural Elements for Computer Graphics, McGraw-Hill, New York, 1985.

Rossi A., Marti-Renom M.A., Sali A.: Localization of binding sites in protein structures by optimization of a composite scoring function. *Protein Sci.* 15(10): 2366–2380, 2006.

Ruppert J., Welch W., Jain A.N.: Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci.* 6(3): 524–533, 1997.

Saitou N. and Nei M.: The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol.* 4(4): 406–425, 1987.

Sander C. and Schneider R.: Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9(1): 56–68, 1991.

Sanner M.F. and Olson A.J.: REDUCED SURFACE: an Efficient Way to Compute Molecular Surfaces. *Biopolymers* 38(3): 305–320, 1996.

Schmidtke P., Le Guilloux V., Maupetit J., Tuffery P.: fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.* 38(Web server issue): W582–W589, 2010.

Schmidtke P. and Xavier B.: Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites. *J. Med. Chem.* 53: 5858–5867, 2010.

Schmidtke P., Luque J., Murray J., Xavier B.: Shielded Hydrogen Bonds as Structural Determinants of Binding Kinetics: Application in Drug Design. *J. Am. Chem. Soc.* 133(46): 18903–18910, 2011.

Schrödinger Suite 2011 Schrödinger Suite; Epik, version 2.2, Schrödinger, LLC, New York, NY, 2011; Impact, version 5.7, Schrödinger, LLC, New York, NY, 2011; Prime, version 2.3, Schrödinger, LLC, New York, NY, 2011. SiteMap, version 2.5, Schrödinger, LLC, New York, NY, 2011.

Schwartz R.M. and Dayhoff M.O.: Matrices for detecting distant relationships. In book: Atlas of Protein Sequence and Structure, pp. 353–358, Edit. Dayhoff M.O., National Biomedical Research Foundation, Washington DC 1979.

Scott W.R.P. and Schiffer C.A.: Curling of flaps tips in HIV-1 protease as a mechanism for substrate entry and tolerance of drug resistance. *Structure* 8: 1259–1265, 2000.

Shan Y., Kim E.T., Eastwood M.P., Dror R.O., Seeliger M.A., Shaw D.E.: How Does a Drug Molecule Find Its Target Binding Site. *J. Am. Chem. Soc.* 133(24): 9181–9183, 2011.

Shao W., Everitt L., Manchester M., Loeb D.D., Hutchison C.A., Swanstrom R.: Sequence requirements of the HIV-1 protease flap region determined by saturation mutagenesis and kinetic analysis of flap mutants. *Proc. Natl. Acad. Sci. U.S.A.* 94: 2243–2248, 1997.

Sharp K.A. and Honig B.: Calculating Total Electrostatic Energies with the Nonlinear Poisson-Boltzman Equation. *J. Phys. Chem.* 94: 7684–7692, 1990.

Sheridan R.P., Maiorov V.N., Holloway M.K., Cornell W.D., Gao Y.D.: Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *J. Chem. Inf. Model.* 50(11): 2029–2040, 2010.

Shrake A. and Rupley J.A.: Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* 79(2): 351–371, 1973.

Sideraki V., Mohamedali K.A., Wilson D.K., Chang Z., Kellems R.E., Quiocho F.A, Rudolph F.B: Probing the Functional Role of Two Conserved Active Site Aspartates in Mouse Adenosine Deaminas. *Biochemistry* 35: 7862–7872, 1996a.

Sideraki V., Wilson D.K., Chang Z., Kurz L.C., Quiocho F.A, Rudolph F.B: Site-Directed Mutagenesis of Histidine 238 in Mouse Adenosine Deaminase: Substitution of Histidine 238 Does Not Impede Hydroxylate Formation. *Biochemistry* 35: 15019–15028, 1996b.

Singh G. and Senapati S.: Molecular Dynamics Simulations of Ligand-Induced Flap Closing in HIV-1 Protease Approach X-ray Resolution: Establishing the Role of Bound Water in the Flap Closing Mechanism. *Biochemistry* 47: 10657–10664, 2008.

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

Soga S, Shirai H, Kobori M, Hirayama N.: Use of amino acid composition to predict ligand-binding sites. *J. Chem. Inf. Model.* 47(2): 400–406, 2007a.

Soga S., Shirai H., Kobori M., Hirayama N.: Identification of the druggable concavity in homology models using the PLB index. *J. Chem. Inf. Model.* 47(6): 2287–2292, 2007b.

Sonka M., Hlavac V., Boyle R.: Image Processing, Analysis, and Machine Vision. Chapter 5, Border tracing, 2nd ed. PWS Pub Co., pp. 142, 1998.

Spinelli S., Lin Q.Z., Alzari P.M., Hirel P.H., and Polzak R.J.: The three dimensional structure of aspartyl protease from HIV-1 isolate BRU. *Biochimie* 73: 1391–1396, 1991.

Stahl M., Bur D., Schneider G.: Mapping of proteinase active sites by projection of surface-derived correlation vectors. *J. Comput. Chem.* 20(3): 336–347, 1999.

Stahl M., Taroni C., Schneider G.: Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Engineering* 13(2): 83–88, 2000.

Surleraux D.L., Tahri A., Verschueren W.G., Pille G.M., de Kock H.A., Jonckers T.H., Peeters A., De Meyer S., Azijn H., Pauwels R., de Bethune M.P., King N.M., Prabu-Jeyabalan M., Schiffer C.A., Wigerinck P.B.: Discovery and selection of TMC114, a next generation HIV-1 protease inhibitor. *J. Med. Chem.* 48: 1813–1822, 2005.

Suzek B.E., Huang H., McGarvey P., Mazumder R., Wu C.H.: UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23(10): 1282–1288, 2007.

Taylor W.R.: The Classification of amino acid conservation. *J. Theor. Biol.* 119(2): 205–218, 1986.

Thangudu R.R., Bryant S.H., Panchenko A.R., Madej T.: Modulating Protein–Protein Interactions with Small Molecules: The Importance of Binding Hotspots. *J. Mol. Biol.* 415(2): 443–453, 2012.

Thompson J.D., Higgins D.G., Gibson T.J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids. Res.* 22(22): 4673–4680, 1994.

Till M.S. and Ullmann G.M.: McVol -A program for calculating protein volumes and identifying cavities by a Monte Carlo algorithm. *J. Mol. Model.* 16(3): 419–429, 2010.

Tóth G. and Borics A.: Flap opening mechanism of HIV-1 protease. *J. Mol. Graph. Model.* 24: 465–474, 2006a.

Tóth G. and Borics A.: Closing of the flaps of HIV-1 protease induced by substrate binding: A model of flap closing mechanism in retroviral aspartic proteases. *Biochemistry* 45: 6606–6614, 2006b.

Tozzini V. and McCammon J.A.: A coarse grained model for the dynamics of flap opening in HIV-1 protease. Chemical Physics Letters 413: 123–128, 2005.

Tripathi A. and Kellogg G.E.: A Novel and Efficient Tool for Locating and Characterizing protein cavities and binding sites. *Proteins* 78(4): 825–842, 2010.

Tseng Y.Y., Dupree C., Chen Z.J., Li W.H.: SplitPocket: identification of protein functional surfaces and characterization of their spatial patterns. *Nucleic Acids Res.* 37: W384–W389, 2009.

Tseng Y.Y., Chen Z.J. and Li W.H.: FPOP: footprinting functional pockets of proteins by comparative spatial patterns. *Nucleic Acids Res.* 38: D288–D295, 2010.

Tsodikov O.V., Record M.T. Jr., Sergeev Y.V.: Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *J. Compt. Chem.* 23: 600–609, 2002.

UniProt Consortium.: The universal protein resource (UniProt). *Nucleic Acids Res.* 35(Database issue): D193–D197, 2007.

Varghese J.N.: Development of neuraminidase inhibitors as anti-influenza virus drugs. *Drug. Dev. Res.* 46: 176–196, 1999.

Venkatachalam C.M., Jiang X., Oldfield T., Waldman M.: LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph. Model.* 21: 289–307, 2003.

Vidler L.R., Brown N., Knapp S., Hoelder S.: Druggability Analysis and Structural Classification of Bromodomain Acetyl-lysine Binding Sites. *J. Med. Chem.* 55(17): 7346–7359, 2012.

Volkamer A., Griewel A., Grombacher T., Rarey M.: Analyzing the Topology of Active Sites: On the Prediction of Pockets and Subpockets. *J. Chem. Inf. Model.* 50: 2041–2052, 2010.

Volkamer A., Kuhn D., Grombacher T., Rippmann F., Rarey M.: Combining Global and Local Measures for Structure-Based Druggability Predictions. *J. Chem. Inf. Model.* 52: 360−372, 2012.

von Itzstein M., Wu W.Y., Kok G.B., Pegg M.S., Dyason J.C., Jin B., Van Phan T., Smythe M.L., White H.F., Oliver S.W., Colman P.M., Varghese J.N., Ryan D.M., Woods J.M., Bethell R.C., Hotham V.J., Cameron J.M., Penn C.M.: Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* 363(6428): 418–423, 1993.

Voorinholt R., Kosters M.T., Vegter G., Vriend G., Hol W.G.J.: A very fast program for visualizing protein surfaces, channels and cavities. *J. Mol. Graph.* 7(4): 243–245, 1989.

Voter A.F.: Hyperdynamics: Accelerated Molecular Dynamics of Infrequent Events. *Phy. Rev. Lett.* 78: 3908−3911, 1997.

Wallach I. and Lilien R.: The protein-small-molecule database, a non-redundant structural resource for the analysis of protein-ligand binding. *Bioinformatics* 25(5): 615–620, 2009.

Wang R., Fang X., Lu Y., Wang S.: The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *J. Med. Chem.* 47(12): 2977–2980, 2004.

Wang Y.F., Tie Y., Boross P.I., Tozser J., Ghosh A.K., Harrison R.W., Weber I.T.: Potent new antiviral compound shows similar inhibition and structural interactions with drug resistant mutants and wild type HIV-1 protease. J. Med. Chem. 50: 4509–4515, 2007.

Wang Z. and Quiocho F.A.: Complexes of Adenosine Deaminase with Two Potent Inhibitors: X-ray Structures in Four Independent Molecules at pH of Maximum Activity. Biochemistry 37: 8314–8324, 1998.

Warring M: Defining optimum lipophilicity and molecular weight ranges for drug candidates—Molecular weight dependent lower logD limits based on permeability. Bioorganic & Medicinal Chemistry Letters 19: 2844–2851, 2009.

Wei Y., Ko J., Murga L.F., Ondrechen M.J.: Selective prediction of interaction sites in protein structures with THEMATICS. *BMC Bioinformatics* 8: 119, 2007.

Weisel M., Proschak E., Schneider G.: PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.*1: 7, 2007.

Wildman, S.A. and Crippen, G.M.: Prediction of Physiochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* 39(5): 868–873, 1999.

Wilson D.K., Rudolph F.B., Quiocho F.A.: Atomic structure of adenosine deaminase complexed with a transition-state analog: understanding catalysis and immunodeficiency mutations. Science 252(5010): 1278–1284, 1991.

Wilson D.K. and Quiocho F.A.: A pre-transition-state mimic of an enzyme: X-ray structure of adenosine deaminase with bound 1-deazaadenosine and zinc-activated water. Biochemistry 32(7):1689–1694, 1993.

Wishart D.S., Knox C., Guo A.C., Shrivastava S., Hassanali M., Stothard P., Chang Z., Woolsey J.: DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 34(Database issue): D668–D672, 2006.

Wishart D.S., Knox C., Guo A.C., Cheng D., Shrivastava S., Tzur D., Gautam B., Hassanali M.: DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36(Database issue): D901–D906, 2008.

Wlodawer A., Miller M., Jaskolski M., Sathyanarayana B.K., Baldwin E., Weber I.T., Selk L.M., Clawson L., Schneider J., Kent S.B.H.: Conserved folding in retroviral proteases: Crystal structure of a synthetic HIV-1 protease. Science 245: 616–621, 1989.

Wlodawer A. and Erickson J.W.: Structure-based inhibitors of HIV-1 protease. *Annu. Rev. Biochem.* 62: 543–585, 1993.

Xie L. and Bourne P.E.: A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics* 8(Suppl. 4): S9, 2007.

Yang Z.R., Thomson R., McNeil P., Esnouf R.M.: RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 21(16): 3369–3376, 2005a.

Yang Z.R., Wang L., Young N., Trudgian D., Chou K.C.: Pattern recognition methods for protein functional site prediction. Curr. Protein Pept. Sci. 6(5): 479–491, 2005b.

Yao H., Kristensen D.M., Mihalek I., Sowa M.E., Shaw C., Kimmel M., Kavraki L., Lichtarge O.: An Accurate, Sensitive, and Scalable Method to Identify Functional Sites in Protein Structures. *J. Mol. Biol.* 326(1): 255–261, 2003.

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

Zhang M-Q, Wilkinson B: Drug discovery beyond the ‘rule-of-five’. *Current opinion in biotechnology* 18(6): 478–488, 2007.

Zhang Y. and Skolnick J.: TM-align: A protein structure alignment algorithm based on TM-score. *Nucleic Acids Res.* 33(7): 2302–2309, 2005.

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

Zhong S. and MacKerell A.D. Jr.: Binding Response: A Descriptor for Selecting Ligand Binding Site on Protein Surfaces. *J. Chem. Inf. Model.* 47(6): 2303–2315, 2007.

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

Appendixes

Appendix I. VBA-script for the automated processing of Binding MOAD

ATTENTION: Before the start of the script running it is crucial to add the value “xxx” in each column of spreadsheet (at least columns A–M). Otherwise, the running of script never stops.

```
Sub DoItAll()

'This part of code copies the EC numbers until the value ”xxx” is achieved. If cell is not empty, script
'copies it and moves to the next row. If cell is empty, script paste copied data and moves to the next
'row.

   Range("A1").Select  'This informs the cell, which is the starting point. Now A1
   Do Until Selection.Value = "xxx"
      If Selection.Value <> "" Then
         Application.CutCopyMode = False
         ActiveCell.Copy
         Selection.Offset(1, 0).Select
      Else
         ActiveCell.PasteSpecial
         Selection.Offset(1, 0).Select
      End If
   Loop


'This part of code copies the PDB ID numbers until the value ”xxx” is achieved. If cell is not empty,
'script copies it and moves to the next row. If cell is empty, script paste copied data and moves to the
'next row.

   Range("C2").Select  'This informs the cell, which is the starting point. Now C2
   Do Until Selection.Value = "xxx"
      If Selection.Value <> "" Then
         Application.CutCopyMode = False
         ActiveCell.Copy
         Selection.Offset(1, 0).Select
      Else
         ActiveCell.PasteSpecial
         Selection.Offset(1, 0).Select
      End If
   Loop


'This part of code erase empty cells and cells that include “invalid” value until the value ”xxx” is
'achieved.

   Range("E2").Select  'This informs the cell, which is the starting point. Now E2
   Do Until Selection.Value = "xxx"
      If Selection.Value = "" Then
         Selection.EntireRow.Delete
      ElseIf Selection.Value = "invalid" Then
         Selection.EntireRow.Delete
```

```
    Else
      Selection.Offset(1, 0).Select
    End If
  Loop

'This part of code erases other affinity values than Ki

  Range("F1").Select  'This informs the cell, which is the starting point. Now F1
  Do Until Selection.Value = "xxx"
    If Selection.Value = "ki" Then
      Selection.Offset(1, 0).Select
    ElseIf Selection.Value = "Ki" Then
      Selection.Offset(1, 0).Select
    Else
      Selection.EntireRow.Delete
    End If
  Loop

'This part of code changes the values of affinities to comparable magnitude

  Range("I1").Select  'This informs the cell, which is the starting point. Now I1
  Do Until Selection.Value = "xxx"
    If Selection.Value = "M" Then
      Selection.Offset(0, 6).Value = 1
      Selection.Offset(0, 7).Formula = Selection.Offset(0, -1).Value * Selection.Offset(0, 6).Value
      Selection.Offset(1, 0).Select
    ElseIf Selection.Value = "mM" Then
      Selection.Offset(0, 6).Value = 10 ^ -3
      Selection.Offset(0, 7).Formula = Selection.Offset(0, -1).Value * Selection.Offset(0, 6).Value
      Selection.Offset(1, 0).Select
    ElseIf Selection.Value = "uM" Then
      Selection.Offset(0, 6).Value = 10 ^ -6
      Selection.Offset(0, 7).Formula = Selection.Offset(0, -1).Value * Selection.Offset(0, 6).Value
      Selection.Offset(1, 0).Select
    ElseIf Selection.Value = "nM" Then
      Selection.Offset(0, 6).Value = 10 ^ -9
      Selection.Offset(0, 7).Formula = Selection.Offset(0, -1).Value * Selection.Offset(0, 6).Value
      Selection.Offset(1, 0).Select
    ElseIf Selection.Value = "pM" Then
      Selection.Offset(0, 6).Value = 10 ^ -12
      Selection.Offset(0, 7).Formula = Selection.Offset(0, -1).Value * Selection.Offset(0, 6).Value
      Selection.Offset(1, 0).Select
    ElseIf Selection.Value = "fM" Then
      Selection.Offset(0, 6).Value = 10 ^ -15
      Selection.Offset(0, 7).Formula = Selection.Offset(0, -1).Value * Selection.Offset(0, 6).Value
      Selection.Offset(1, 0).Select
    Else
      Selection.Offset(1, 0).Select
    End If
  Loop

  'This piece of code calculates the common logarithmically values of affinities, other words pKi

  Range("F1").Select  'This informs the cell, which is the starting point. Now E2
  Do Until Selection.Value = "xxx"
    If Selection.Value = "Ki" Then
      Selection.Offset(0, 10).Select
```

```
        Selection.Offset(0, 1).Formula = "=-Log10(" & Selection.Address & ")"
        Selection.Offset(1, -10).Select
      ElseIf Selection.Value = "ki" Then
        Selection.Offset(0, 10).Select
        Selection.Offset(0, 1).Formula = "=-Log10(" & Selection.Address & ")"
        Selection.Offset(1, -10).Select
      Else
        Selection.Offset(1, 0).Select
      End If
    Loop

End Sub
```

This script copies the values of M_w, which are calculated from SMILES strings by ChemDraw add-in, to next column. This part was made, because the results of ChewDraw add-in will not show with computers, which not contain ChewDraw software.

```
Sub Mw-paste()

'This part of code copies the values of calculated Mw until the value ”xxx” is achieved. If cell is not
'empty, script copies it and moves to the next row. If cell is empty, script paste copied data and moves
'to the next row.

    Range("M2").Select  'This informs the cell, which is the starting point. Now M2
    Do Until Selection.Value = "xxx"
      If Selection.Value <> "" Then
        ActiveCell.Copy
        Selection.Offset(0, 1).Select
        Selection.PasteSpecial Paste:=xlPasteValues
        Selection.Offset(1, -1).Select

        Application.CutCopyMode = False

      Else
        Selection.Offset(1, 0).Select
      End If
    Loop

End Sub
```