Chapter 8

Methods to Characterize the Structure of Enzyme Binding Sites

A. Kahraman* and J. M. Thornton

8.1 Introduction

Enzyme binding sites are regions on the surface of an enzyme specially designed to interact with other molecules. An enzyme can have different sorts of binding sites that differ in their functions and the molecules they bind. Amongst these, the most important is the active site, which consists of two or three parts. The first part is the catalytic site, which contains the catalytic machinery of the enzyme in the form of usually two to six amino acids that perform the catalytic reaction. The second part is the substrate binding site, which has the task of specifically recognizing the molecule upon which the enzyme acts. Besides the specificity, the substrate binding site also provides binding energy to keep the substrate bound on the active site for the time the catalytic reaction progresses. Enzymes can act on a huge variety of substrates, from small molecules, like hormones and sugar, and moderate sized molecules, like polypeptides and oligosaccharides, to macromolecules, like DNA and other proteins. Figure 8.1 shows an exemplary substrate binding site for an asparagine in the structure of the *Escherichia coli* asparagine synthetase [see also Fig. 8.2(a)].

*Corresponding author.
European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB4 1BY, UK. Email: abdullah@ebi.ac.uk.

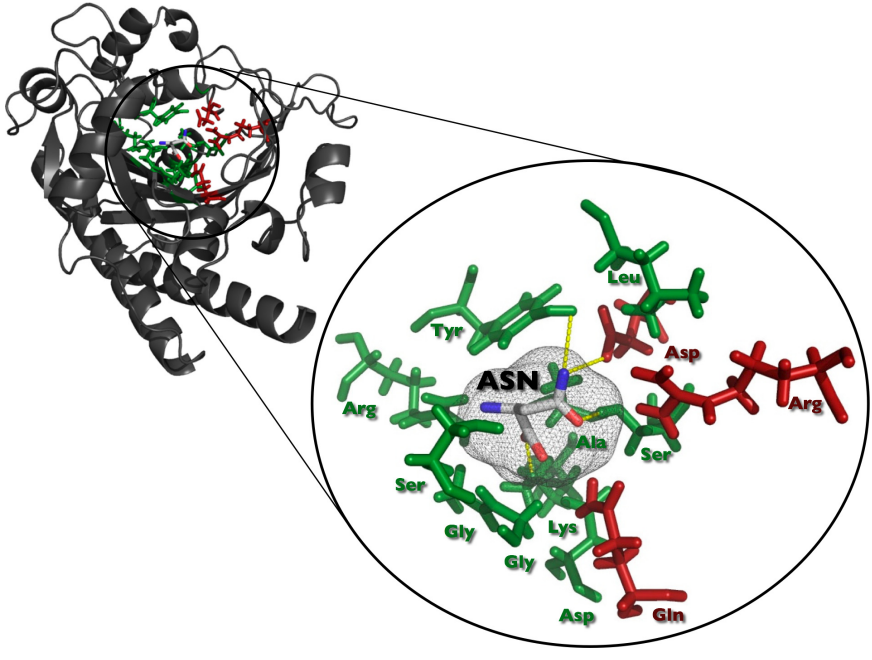

Fig. 8.1 Structure of the *Escherichia coli* Asparagine Synthetase (PDB Id: 12as) with a zoom-in into the binding site of the substrate asparagines. Binding site residues as determined by HBPLUS (see Section 8.2.9) are colored in green; catalytic active residues extracted from CSA (see Section 8.3.1) are coloured in red, and the substrate is varicolored. Hydrogen bonds between binding site residues and substrate are indicated by yellow dashed lines. The binding site shape is shown as a grey mash as approximated with spherical harmonic functions (see Section 8.2.3).

As enzymes are proteins they usually consist of 20 amino acids with either a hydrophobic or polar, charged or uncharged side chain. For some catalytic reactions the chemical properties of these amino acids may be sufficient, but for the majority of reactions such as redox reactions or chemical group transfers, enzymes require the assistance of additional molecules that bind on the third part of an active site. These molecules are defined as either cofactors, which are tightly bound to the enzyme throughout the catalytic reaction or coenzymes, which are released during the reaction. Cofactors distinguish themselves from coenzymes by being not consumed in the catalytic reaction.

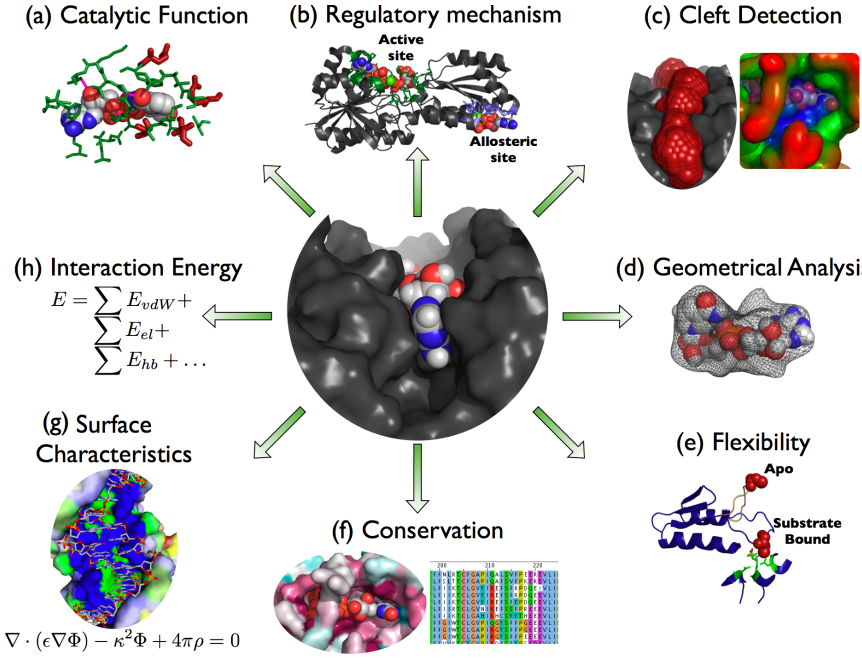

Fig. 8.2 Characteristics of enzyme binding sites. **(a)** The active site is a specific binding site in an enzyme that contains catalytic residues to perform the enzymatic reaction on a substrate. **(b)** The activity of an enzyme can be regulated, for example, by allosteric regulator molecules that bind to a remote binding site. **(c)** In most enzymes the active site is found in the largest or deepest cleft of the enzyme, **(d)** and encloses at least partially the ligand with amino acids, resulting in similar geometrical shapes for the binding site and ligand. **(e)** Binding sites can undergo major conformational changes upon subtrate binding, especially when some parts of the site are located in flexible loops. **(f)** As binding sites are essential for the function of a protein, their residues are often amongst the most highly conserved residues. **(g)** The binding affinity of a ligand is influenced by the physicochemical properties on the binding site surface like complementary electrostatic potentials or perturbed pK_a values **(h)**, which can be exploited to calculate estimated binding energies between ligand and binding sites.

Though they get altered while the catalysis takes place, they are recovered again in the same catalytic process. In contrast, coenzymes support the enzyme reaction by providing chemical groups to the substrate, and subsequently, detaching from the enzyme to start

a recovery process outside the enzyme. Typical cofactors are the inorganic metals and sulphate ions or the organic flavin and heme groups. Examples of coenzymes are vitamins or the cellular energy carrier, ATP.

Some enzymes, especially ones assembled by several domains or several chains, can have allosteric sites in addition to the substrate and cofactor/coenzyme binding sites [see Fig. 8.2(b)]. These allosteric sites play an important role in the regulation of enzymes as they induce, upon binding a regulator molecule, conformational changes on the whole enzyme structure, which can affect also the atomic constellation of the active sites. Depending on whether the regulator molecule is an effector or an inhibitor, the changes on the active site can either enhance or hamper the enzymatic catalysis.

The underlying principles of allosteric regulation, as well as the atomic interactions of any binding process between an enzyme and a molecule, have only been elucidated since high-resolution data of the three-dimensional (3D) coordinates of enzyme-molecule complexes were determined. Two main approaches are used for the determination of such high-resolution data for biomolecules, namely "X-ray crystallography" and "Nuclear Magnetic Resonance (NMR) spectroscopy" (see Chapters 22 and 24 for an in-depth description of these methods). The first enzyme structure discovered in 1965 was the X-ray structure of lysozyme, an enzyme found in tears or egg white that digest bacterial cell walls. Since then, many enzyme structures have been determined and their functions analyzed, and the resulting information has been stored in databases. See Table 8.1 for the number of enzymes in some structure-based databases.

The most important among them is the Protein Data Bank (PDB)[1] (http://www.pdb.org) and the Enzyme Commission (EC) number for enzyme reaction.[2] The first is important as it is the worldwide depository for 3D coordinates of enzymes and any other macromolecules like other proteins, nucleic acids, or carbohydrates (see Chapter 26 for further information on the PDB). Structures in the PDB are assigned a unique four alphanumeric PDB Identifier (Id). The importance of the EC number is that it provides a classification scheme for all enzyme reactions and allows their comparison.

Table 8.1 The Extent of Enzyme Data in Some Structural Databases as on 21 July 2007

Number of	Quantity
Known enzyme reactions (unique EC numbers)	~4040
Enzymes in UniProt/Swiss-Prot (56)	~107 400
Enzymes in PDB	~19 600
EC Reactions in PDB	~1390
Enzymes with catalytic residues in CSA	880
Enzymes with catalytic mechanisms in MACiE (57)	202
Enzymes as specified by EC number in PDB with the largest number of structures	
1. Lysozyme, EC 3.2.1.17	~930
2. Non-specific serine/threonine protein kinases, EC 2.7.1.37	~580
3. Trypsin, EC 3.4.21.4	~430
Most enzymes in PDB originate from	
1. Human	~10 700
2. *Escherichia coli*	~4200
3. House mouse	~2100
4. Cow	~1550
5. Baker's yeast	~1300
No of organisms that have one or more enzyme structures in PDB	~1128

The EC number consists of four digits separated by full stops. The first number (class) indicates the reaction type, the second number (sub-class) together with the third number (sub-subclass) represents the occurring chemistry, and the last number gives the substrate specificity.

From the three-dimensional structures of enzymes, it became evident that substrates and secondary molecules like cofactors and coenzymes do not bind randomly on the enzyme surface. The same molecule always binds at the same site within the same enzyme structure. This has led to the assumption that binding sites must have unique features that distinguish them from other areas on the enzyme surface, and in addition, allow the binding site to recognize its associated molecule from the thousands that exist in a living cell. Two

models were suggested to explain in particular the specificity of active sites. First, the Lock and Key model by Fischer,[3] and second, the Induced Fit model by Koshland.[4] The Lock and Key model assumed that a ligand is geometrically complementary to its active site and that both shapes fit exactly into one another. The more recent model of Induced Fit was a modification to the Lock and Key model and incorporated the flexibility of enzymes and substrates. The model suggests an "open" state for an enzyme when the substrate binds, followed by a "closed" state where the enzyme encloses the bound substrate and performs its catalysis. In the process of converting from the open state to the closed state, the active site adjusts its shape to the transition state that is the conformation of the ligand at the highest reaction energy (see Chapter 10), and allows the catalytic reaction to take place.

This chapter addresses different aspects or features of protein binding sites (see Fig. 8.2). It will give some background information to each feature and describe one exemplary methodology to calculate it. A more comprehensive list of computational methods can be found at the end of the next section. All tools and programs introduced in this chapter are not just important to visualize the features in an enzyme but also to try to predict the function for an enzyme. The latter becomes more and more important as more and more enzyme structures are deposited in the PDB without any functional annotation. Many of these structures were targets of global structural genomics initiatives, which aim to develop high-throughput methods for the rapid determination of protein structures. One goal of these initiatives is to determine the structures of all existing protein folds in nature.[5] The high-throughput principle is advantageous for determining many structures in a short time but does not address the functional annotation of proteins, which usually involves many different wet lab experiments and thus is a time-consuming procedure. In order to obtain hints about the function of these unannotated structures, one can extract the features described in this chapter and search for similar features in annotated enzymes. For this purpose, the third part of this chapter will be devoted to algorithms for the comparison of binding site features.

Before we start with the binding site characteristics we would like to note that in this chapter we will refer to any small molecule that is bound by an enzyme as a ligand whether it is a substrate, product, or allosteric effector.

8.2 Enzyme Binding Sites and Their Unique Features

8.2.1 *Active Sites are in Largest Cleft*

Enzyme active sites tend to be within sizeable depressions on the protein's surface, which are known as clefts or pockets. In 70–85% of enzymes, the largest of these clefts is where the substrate and relevant cofactors or coenzymes bind.[6,7] The average volume of a binding site depends on the ligand it binds, and ranges mostly from 400 to 2000 Å^3.[8]

SURFNET[9] is an elegant approach to identify and visualize clefts in proteins. It detects gap regions within the protein by fitting spheres of certain range of sizes between protein atoms. The spheres are not allowed to clash with any neighboring protein atoms. Overlapping SURFNET spheres are clustered and regarded as protein clefts [see Fig. 8.3 and Fig. 8.2(c)]. Placing a grid on the cleft and determining the number of grid cells occupied by a sphere enables the calculation of the volume for each cleft.

8.2.2 *Active Sites are in Deepest Cleft*

The enclosure of a ligand within large and deep clefts helps the enzyme to maximize the number of interactions with its ligand.[10] In particular, active sites are often found in the deepest cleft of an enzyme. The average depth of a cleft that contains a binding site depends on the protein size and can be up to 30 Å.[11]

The algorithm of travel depth[11] is an elegant way to visualize and measure the depth of clefts relative to the convex hull of the enzyme's molecular surface [see Fig. 8.2(c)]. The convex hull is defined for a simplified two-dimensional molecule as the region that is enclosed by

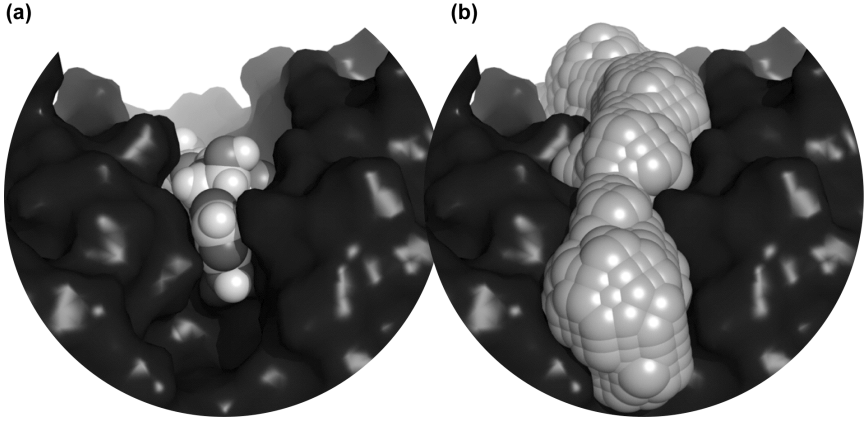

Fig. 8.3 **(a)** Spherical section of the protein structure of ribosyl-transferase (PDB Id: 1og3) colored in black, with bound coenzyme NAD in the active site. **(b)** Largest cleft, as determined by SURFNET, contains the active site. SURFNET spheres are represented by light grey spheres.

a rubber band that is stretched around the whole molecule. The travel depth algorithm finds for a probe sphere on the protein surface the minimum distance to reach the convex hull. It works by placing the protein into a grid and assigning to all grid cells outside the convex hull a depth of zero. For grid cells inside the convex hull, the algorithm scans recursively through the grid and adds to the size of each grid cell the minimum depth of its neighboring cells.

8.2.3 *Binding Site Shapes are Complementary to Ligand Shapes*

It is a common assumption that the shapes of protein binding pockets are complementary to the shapes of the ligands they bind. This assumption became manifest in the Lock and Key model and Induced Fit model for molecular binding (see Section 8.1). A recent study however showed that exact shape complementarity between a binding site and its bound ligand is rarely achieved, and that more often, some free space can be found between the binding site and its ligand[8] (see Figs. 8.2(d) and 8.4).

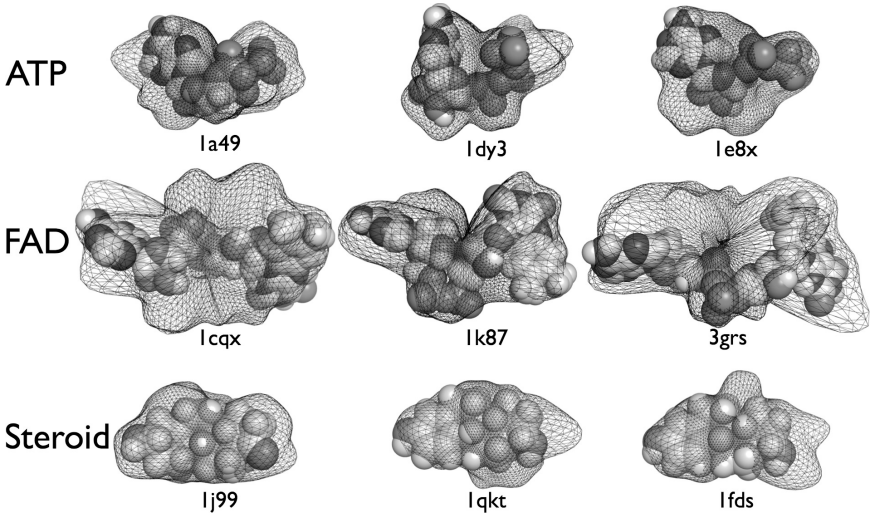

Fig. 8.4 Binding site shapes are not truly complementary to the ligand shape and often show some empty space between the ligand and the binding site like a "buffer zone." The PDB identifier of each associated protein structure is given below each binding site.

For the analysis and visualization of binding sites and ligand shapes, one can apply an elegant approach, which utilizes mathematical functions called spherical harmonics. The computational description of shapes can be simplified by a radius function, which returns for any point on the shape surface its distance to the center of the shape. The common way of obtaining the function is by selecting a number of points on the shape surface and exploiting their radii to approximate the radius function. The approximation can be done by summing up spherical harmonic functions in an equivalent way to the Fourier series, where sine or cosine functions are summed up to obtain a periodic one-dimensional function. While the summation progresses, each spherical harmonic function contributes, with a different weight, to the radius function. The contribution weights are usually referred to as coefficients. Once the approximation finishes, a vector of all coefficients is retrieved and used to reconstruct the shape of the binding site (see Fig. 8.5).

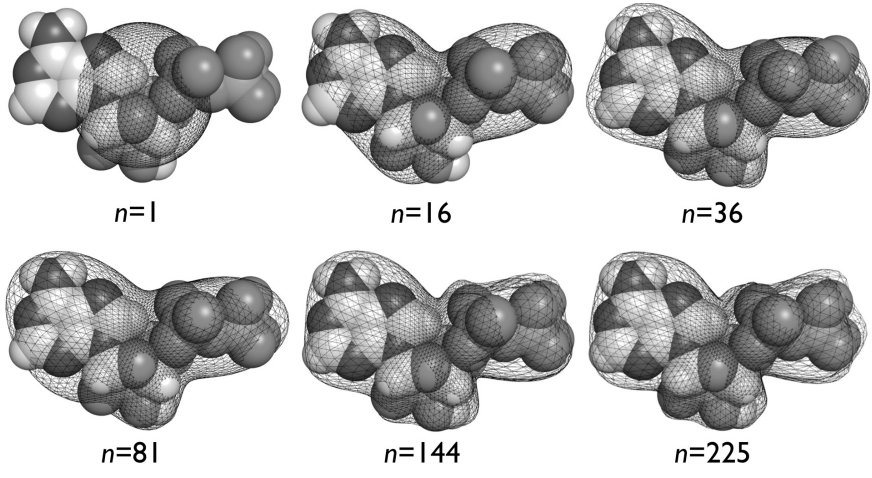

Fig. 8.5 Various approximations of the molecular surface of an ATP (PDB ID: 1e8x) with increasing number n of spherical harmonic functions.

8.2.4 *Binding of the Ligand Induces Conformational Changes in the Binding Site*

The Induced Fit model for molecular binding states that enzymes undergo conformational changes upon substrate binding. For a small fraction of enzymes these changes are large, in particular, if they include a flexible loop region that closes/opens the entrance to the active site, preventing/allowing the binding of a ligand [see Fig. 8.2(e)]. However, for the majority of enzymes, the changes are small. The average RMSD (see below) upon ligand binding between C_a atoms of binding sites and catalytic residues is less than 1 Å.[12] Similar values are observed for the side chain atoms. It is interesting to note that residues in active sites are on average more flexible than other residues in the protein structure. This can be traced back to the geometrical adjustments of the active site to generate the transition state of the ligand (see Section 8.1). But there are also enzymes, like prothymosin-α, that are intrinsically disordered in their native state.[13] Neither the Lock and Key nor the Induced Fit model can describe their functionality. A third model, the "New View" model, has recently been introduced, and it states that a protein exists in an ensemble of pre-existing

conformations with discrete and similar free energies. Among them is also the structure of the bound conformation. The actual binding of the ligand induces a shift in the equilibrium of existing conformations towards the bound conformation and causes the protein to appear well structured in an X-ray crystal.[14]

The standard method for measuring the flexibility of enzymes binding sites is to calculate the Root Mean Square Deviation (RMSD) between different conformations of the binding site. The RMSD is calculated between the Cartesian coordinates of all atom pairs between both proteins using the following formula:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=0}^{N}\left[(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2\right]}{N}} \qquad (8.1)$$

where *x, y, z* are the Cartesian coordinates of the protein atoms and N is the number of compared atoms. Depending on the scientific question or on the available data, one can calculate the RMSD of all atoms, of all residue side chain atoms, or of only the backbone/C_a atoms between two structures. For the qualitative analysis of flexibility, one can use the web server of STRuster.[15] STRuster analyzes an ensemble of different conformations of a protein by first calculating the Euclidean distances between all residues within a conformation, and next, comparing the distances to the distances in the second conformation. The compared distances are summed up and plotted in an "all-conformation versus all-conformation" distance matrix. The distance matrix is utilized to cluster each conformation according to its level of flexibility and group similar conformations.

8.2.5 *Binding Site Residues are Highly Conserved*

Another characteristic of enzyme binding sites is that the residues forming the sites tend to be strongly conserved within the protein family. That is, all members of a protein family tend to have the same residues in the same position in both their sequences and their 3D

structures. The reason for this is that they all have evolved from a common ancestor and have the same function but are found in different organisms. Each family member is however subject to natural variation and selection with mutation and duplication events throughout their protein sequences. However, mutations are not tolerated at all positions in the protein sequence. While those residues that had no functional role in the protein could mutate freely, substitutions of functionally important residues (i.e. residues that are involved in ligand binding or in keeping the structural integrity) are restricted, as these mutations could have led to the loss of protein function. Residues found in binding sites and especially catalytic residues in active sites are amongst the most important residues in an enzyme structure, and consequently, particularly highly conserved. Most often, these residues are either polar or charged (up to 70% of residues are Arg, Asp, Cys, Glu, His, and Lys).[16]

ConSurf[17] calculates the conservation of each amino acid in a protein sequence using the evolutionary trace method.[18] This method first runs a multiple sequence alignment on a set of homologous sequences, i.e. sequences that have a common ancestor. In the second step, the method uses the alignment to compute a phylogenetic tree, which represents the evolutionary relationship of the homologous sequences. In the third step, the homologous sequences are divided into groups and subgroups based on the branches of the tree. In the fourth step, the residue positions in all sequences in each group and subgroups are analyzed for the frequency of residue changes. If at a particular subgroup a residue is invariant throughout all sequences in the subgroup, it becomes assigned a rank, which states how many times the tree was required to be divided to yield the ranked residue. The same procedure is applied to all residues until every residue gets assigned an evolutionary rank. According to the ranks, ConSurf groups the residues of the query sequence into nine classes, with "1" being the least conserved and "9" being the most conserved residues, and the conservation scores are mapped onto the protein structure [see Fig. 8.6 and Fig. 8.2(f)]. A visual inspection of the protein structure can identify clusters of highly conserved residues on the protein surface.

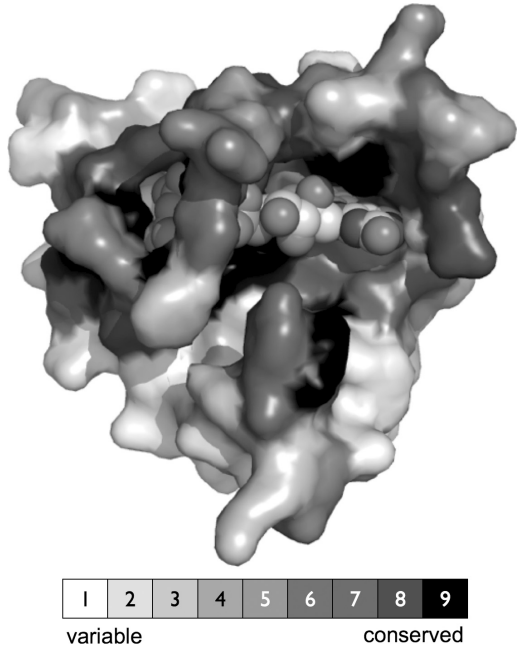

Fig. 8.6 ConSurf conservation scores mapped on PDB structure 1p4m. Note the higher conservation in and around the binding sites.

8.2.6 *Complementary Electrostatic Potentials Between Binding Sites and Ligands*

Electrostatic potentials are long-range potential energies and one of nature's strongest forces at the atomic scale. All energies between atoms and molecules are electrostatic in origin, whether they are transient dipole-dipole interactions as in the case of van der Waals interactions, charge-charge interactions, or hydrogen bond interactions. They differ in the rate of decreasing interaction energy with increasing atomic distance.[19]

One theory about electrostatic complementarity between binding sites and ligands suggests that electrostatic potentials are strong enough to attract the ligand from the solvent into the active site. This assumption has been derived from enzymes that have catalysis rates approaching the diffusion limit, like the copper-zinc-superoxide-dismutase

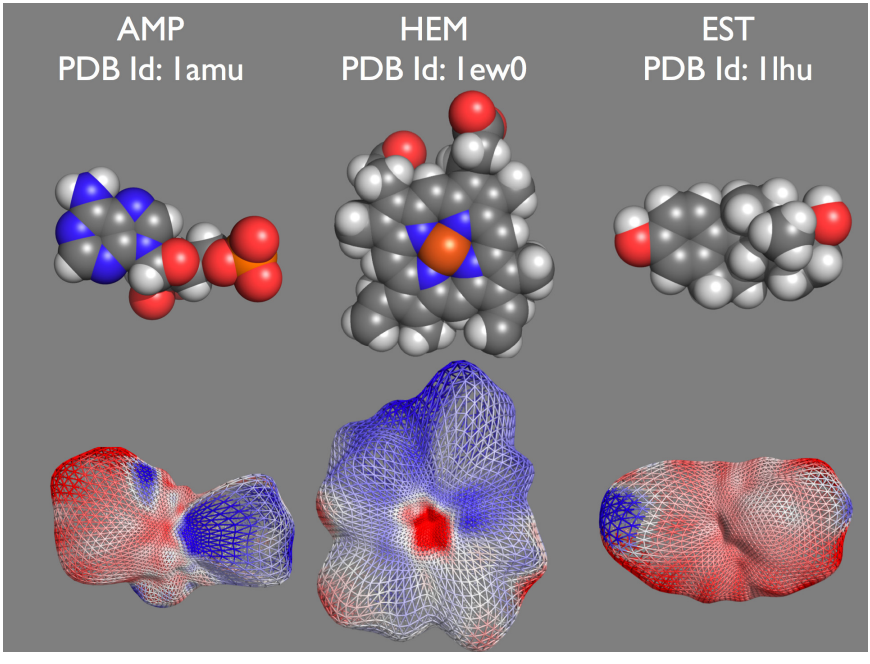

Fig. 8.7 Electrostatic potential of three proteins mapped on the molecular surface of their ligands as represented by spherical harmonics (see Section 8.2.3): AMP, heme, and Estradiol. Negative potentials are colored red, neutral potentials are colored white, and positive potentials are colored blue.

protein family. This protein family exerts a positive electric field over the active site, which attracts negatively charged oxygen radicals towards the active site copper ion.[20] The visualization of the electrostatic potentials mapped on the structure surface, also referred to as potential surfaces, is particular useful for identifying DNA binding sites. Many DNA binding proteins possess a large patch of positively charged amino acids on their surface to electrostatically attract their negatively charged binding partner[21] [see Fig. 8.2(g)]. Figure 8.7 visualizes the electrostatic potentials by showing the potentials of three proteins on the molecular surface of their ligands.

The eF-site[22] database contains pre-calculated potential surfaces for all PDB structures. Auxiliary servers to the eF-site database allow the calculation of the electrostatic potential for any user-provided

structure and the search for similar surface potentials in the eF-site database. The electrostatic potentials in eF-site are calculated by a standard procedure applied by many electrostatic methodologies, among them APBS and Delphi (see Table 8.2).

The methodology simplifies the representation of the protein and the solvent by ignoring the molecular details of the solvent molecules and treating all solvent molecules as a single continuum. The simplification is necessary as the explicit calculation of all interactions between water molecules to each other and to the protein is computational demanding and most often not feasible. In combination with the simplification, the electrostatic potential of a protein is calculated by solving the Linear Poisson-Boltzmann differential Equation (LPBE).[23] As every protein has an arbitrary shape, the LPBE is solved numerically by discretizing the space occupied by the protein with a grid and calculating iteratively the electrostatic potential for each grid cell using the finite difference technique.[24]

8.2.7 *Catalytic Residues Destabilize the Enzyme Structure and Have Perturbed pK_a-values*

The ability to calculate the electrostatic potential for a protein structure facilitated the computational analysis of two further phenomena in active sites. Both phenomena are unique properties of ionizable catalytic residues (all Lys, Arg, Asp, Glu, His, Tyr, Cys, N-terminus, C-terminus) and distinguish them from the remaining residues in the enzyme structure. One of these properties is their capacity to destabilize the integrity of enzyme structures, especially when they are found in active sites that exert repulsive electrostatic forces towards the ionizable catalytic residues. Experiments have shown that the replacement of the affected residues with neutral or oppositely charged residues tended to stabilize the protein structure.[25]

Another of these properties is a perturbed pK_a-value for ionizable catalytic residues. The pK_a is defined as the pH for which the average protonation state of an ionizable molecular group is 0.5. It can be measured by titration curves that plot the solvent's pH against the net charge of the ionizable group. For non-catalytic residues, in general, these

Table 8.2 Programs and Web Servers to Analyze Different Aspects of Enzyme Binding Sites

Method	Program/Server	URL	Notes
Size	SURFNET	http://www.biochem.ucl.ac.uk/~roman/surfnet/surfnet.html	Active sites are most likely in the largest protein cleft.
	CASTp	http://sts.bioengr.uic.edu/castp/	
	VOIDOO	http://xray.bmc.uu.se/usf/voidoo.html	
Depth	TravelDepth	http://crystal.med.upenn.edu/travel_depth.tar.gz	Binding sites are often found in deep protein clefts.
	PocketPicker	http://gecco.org.chemie.uni-frankfurt.de/pocketpicker/index.html	
Flexibility	STRuster	http://struster.bioinf.mpi-inf.mpg.de/	Protein structures can undergo conformational changes upon ligand binding.
	MolMovDB	http://www.molmovdb.org/	
Conservation	ConSurf	http://consurf.tau.ac.il/index.html	Binding sites are among the most conserved regions on the protein.
	Evolutionary Trace	http://www-cryst.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html	
	JevTrace	http://www.cmpharm.ucsf.edu/~marcinj/JEvTrace/	
3D templates	CSA	http://www.ebi.ac.uk/thornton-srv/databases/CSA/	Catalytic residues are often found to be highly conserved in their spatial disposition.
	PINTS	http://www.russell.embl-heidelberg.de/pints/	
	Rigor	http://xray.bmc.uu.se/usf/rigor_man.html	

(*Continued*)

Table 8.2 (*Continued*)

Method	Program/Server	URL	Notes
Electrostatic potential	APBS	http://apbs.sourceforge.net/	DNA binding proteins have often large, positively charged binding sites.
	DELPHI	http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:DelPhi	
	PCE-Pot	http://bioserv.rpbs.jussieu.fr/cgi-bin/PCE-Pot	
	eF-site	http://ef-site.hgc.jp/eF-site/	
pK_a-values	PROPKA	http://propka.ki.ku.dk/	Catalytic residues have often perturbed titration curves.
	WHAT IF pKa	http://enzyme.ucd.ie/Science/pKa/Software	
	PCE-pKa	http://bioserv.rpbs.jussieu.fr/cgi-bin/PCE-pKa	
Hydrophobicity	GRASP	http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:GRASP	Hydrophobic binding sites often bind hydrophobic ligands.
Hydrogen bond	HBPLUS	http://www.biochem.ucl.ac.uk/bsm/hbplus/home.html	Hydrogen bonds provide specificity for ligand binding.
	LIGPLOT	http://www.biochem.ucl.ac.uk/bsm/ligplot/ligplot.html	

(*Continued*)

Table 8.2 (*Continued*)

Method	Program/Server	URL	Notes
Potential function	Q-SiteFinder	http://www.bioinformatics.leeds.ac.uk/qsitefinder/	Binding sites can often develop high interaction energies that can be assessed by potential functions.
	Grid	http://www.moldiscovery.com/soft_grid.php	
	MCSS	http://www.accelrys.com/insight/mcss.html	
Biological Unit	PQS	http://pqs.ebi.ac.uk/pqs-quick.html	PDB structures often represent not the biological active conformation of the protein.
	Pita	http://www.ebi.ac.uk/thornton-srv/databases/pita/	
	PISA	http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html	
Cognate Ligand	PROCOGNATE	http://www.ebi.ac.uk/thornton-srv/databases/procognate/	Not all bound ligands to a protein structure are functionally related.
Enzyme mechanism	MACiE	http://www.ebi.ac.uk/thornton-srv/databases/MACiE/	Enzyme reactions consist of various catalytic steps.

curves adopt a specific shape in which the net charge decreases with increasing pH and with a sharp decline around the pK_a-value. For catalytic residues, however, these curves can be perturbed, generating regions of constant protonation state or shifts in the pK_a-value.[26]

Theoretical microscopic titration curves (THEMATICS)[26] can be computed for every ionizable residue in a protein using electrostatic potential calculations. The superposition of titration curves obtained for all residues of the same type within the protein identifies perturbed curves and may indicate ionizable catalytic residues.

8.2.8 *Hydrophobic Interactions are Essential for Binding*

In a study where organic solvent molecules were computationally mapped on the protein surface to predict potential binding sites of ligands, it was found that hydrophobic patches are also important within binding sites, inducing organic solvents to cluster therein.[27] The results are in agreement with earlier experiments that showed that binding affinities of ligands can increase by promoting hydrophobic interactions between binding sites and ligands.[28] Our own calculations confirmed that hydrophobic ligands like heme and steroids are often bound by binding sites that expose mainly hydrophobic residues.

Computationally, the hydrophobicity of amino acids can be calculated by exploiting the fact that hydrophobic amino acids are usually surrounded by other amino acids in the protein's core and not accessible to solvent molecules. Calculating the mean fractional area loss upon protein folding of a residue provides an estimate on the residue's hydrophobicity. The area loss is obtained by relating the solvent accessible surface area (SASA) of an amino acid in a fully extended conformation to the mean SASA of the amino acid in the protein structure. The SASA can be calculated by rolling a probe sphere over the atomic van der Walls surfaces and placing a fixed number of dots per unit area on the roll track of the probe sphere. The number of dots multiplied by the area that a dot occupies gives the accessible surface area. The ASA of the extended conformation is usually given as the surface area of the residue within the extended tripeptide Gly-X-Gly.[29]

8.2.9 *Hydrogen Bonds Provide Binding Specificity*

Unlike other chemical interactions, hydrogen bonds require directionality between the hydrogen-bond acceptor and donor. This directionality provides the enzyme's specificity for its ligand. Only ligand atoms that have a specific orientation towards a particular binding site can form hydrogen bonds. Ligands that do not have the right atoms at the right place cannot form hydrogen bonds and must rely on other forms of interaction to achieve binding.[30] Most hydrogen bonds in binding sites are formed among the atoms of the binding site in order to stabilize the positions of the catalytic residues. Only a small portion (10–20%) are formed with ligand atoms.[16] In protein-ligand complexes there are on average 10 bonds, of which two-thirds are hydrogen bond acceptors and a third hydrogen bond donors.[31]

The program LIGPLOT[32] uses the application HBPLUS[33] to extract and plot hydrogen bonds between the binding site and ligand atoms. The algorithm of HBPLUS begins with placing hydrogen atoms in the protein structure. This is necessary, as most X-ray crystal structures do not include hydrogen atoms except for NMR or very high-resolution X-ray structures. Once the hydrogen atoms are generated, the hydrogen bonds are determined by applying purely geometrical criteria[32] to the protein-ligand complex. In addition to hydrogen bonds, HBPLUS also calculates non-covalent bond interactions by applying a simple cut-off of 3.9 Å to atomic distances between the binding site and ligand. Finally, LIGPLOT draws a schematic two-dimensional diagram of the binding site ligand complex and highlights the calculated hydrogen bonds and non-covalent bond interactions (see Fig. 8.8).

8.2.10 *Potential Functions for Estimating Binding Energy*

The process of molecular binding requires in the first instance shape complementarity to allow ligand atoms to approach the binding site atoms. The proximity between both binding partners is important as

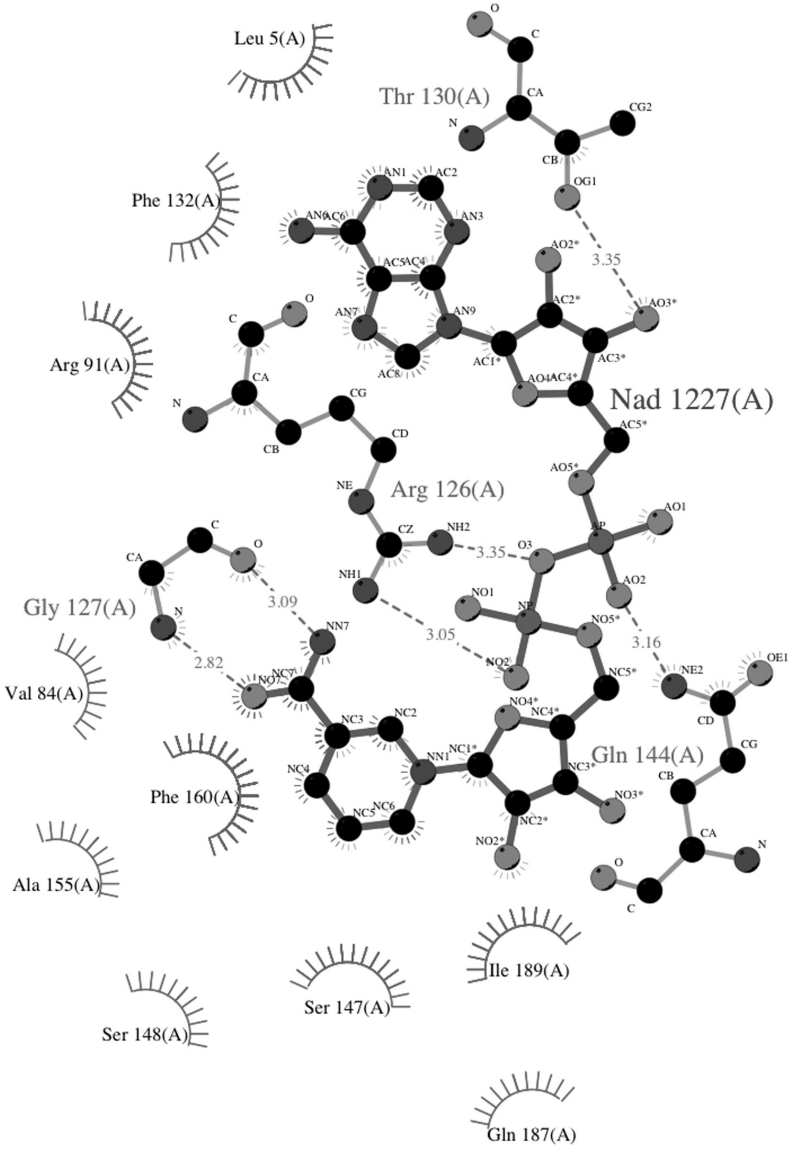

Fig. 8.8 Schematic diagram of the non-covalent interactions between NAD and its binding site in PDB structure 1p4m. Thick lines belong to the ligand and thin lines to the hydrogen-bonded residues in the binding site. Hydrogen bonds are indicated with dashed lines. Non-covalent bond interactions are shown as spoked arcs pointing towards the ligand.

their binding energy depends very much on the distances between their atoms. Since ligand molecules do not bind at random sites on a protein structure, their binding sites should feature particular high binding energies towards the ligand.

Q-SiteFinder[34] calculates the potential binding energies on a protein surface and detects energetically favorable surface patches that may present ligand binding sites. The favorable patches are found by placing the protein in a grid and rolling a probe sphere along the grid points over the molecular surface. At each grid point an energy function, which incorporates van der Waals potential, electrostatic potential, and hydrogen bond potential, is applied to the probe sphere [see Equation in Fig. 8.2(h)]. Grid points that exceed a predefined energy threshold are clustered if they are below a certain separation. For each cluster, the single interaction energies of the grid points are summed up and ranked according to their total interaction energy. The cluster with the most favorable interaction energy is identified and is considered as a potential binding site.

8.2.11 *Unusual Amino Acids*

There are 20 standard amino acids used by nature to build up proteins, however, under certain circumstances some amino acids can be catalytically altered, giving rise to a 21st amino acid. One such change occurs in active sides of copper amine oxidases (PDB Id: 1pu4), which increases the catalytic activity of the enzyme. The change occurs at the catalytic active tyrosine, which becomes autocatalytically oxidized to tri-hydoxy-phenyalalanine (Topa) in the presence of a copper ion.[35] Another example is the phosphomannose isomerase, which when expressed in *E. coli* has a di-hydoxy-phenylalanine (Dopa) substituting for a tyrosine.[36]

8.2.12 *Precautions with PDB Structures*

Structures deposited as single chains in the PDB are often actually dimers or tetramers or sometimes vice versa. When analyzing binding

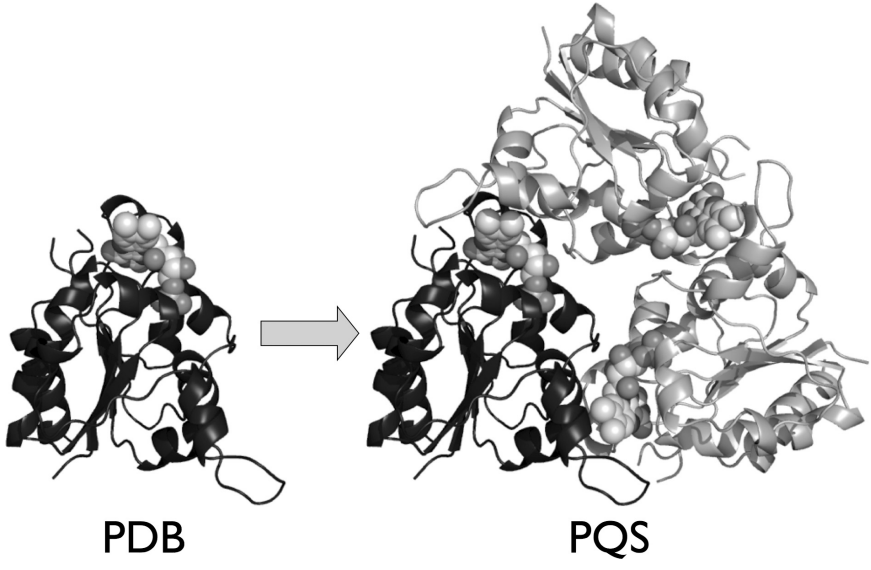

Fig. 8.9 The PDB structure of the decarboxylase 1mvl shows only a monomer with the FMN being exposed to the solvent. However, the biological relevant conformation is a trimer as calculated by PQS, with a FMN binding site shared between two subunits.

sites, one has to bear in mind this obstacle, especially as many binding sites in dimers are found at the interface of the two monomers (see Fig. 8.9). The PQS (Protein Quaternary Structure)[37] file server is a depository of estimated quaternary structures of all PDB structures. We would encourage the reader to use in any of their protein studies these assembles for their proteins from the PQS database, since although not perfect, they are much more reliable than using the single chain.

Ligands that are found attached to an enzyme in a crystal structure may not always be the native substrate or cofactor, etc., of an enzyme. Many such ligands found in the active site are substrate analogues or enzyme inhibitors that compete with the substrate for binding into the active site. In addition, some ligands can be artifacts of the crystallization buffer, which is a mixture of different solvents to promote the crystallization process of a protein. In general, all ligands

that are not required for the enzyme function are called non-cognate, whereas ligands that are functionally related to an enzyme are designated as cognate. The PROCOGNATE[38] database has been established to address this problem and contains information about cognate ligands in enzymes and provides similarity scores for non-cognate ligands that allow their structural comparison to the cognate ones.

8.3 Methods and Tools for Comparing Enyzme Binding Sites

Tools to assess the similarity between binding sites compare either atomic coordinates or surface properties. In the field of computer vision, many methods exists for comparing three-dimensional coordinates, features, or surfaces. See Ref. 39 for a review on the existing methods. However, only a few of them have been realized in structural biology. Among these, the most important ones are the kd-tree search, graph matching, geometrical hashing, and coefficient comparison of spherical harmonic functions. A detailed description of each follows.

8.3.1 *Comparing Catalytic Templates*

As mentioned in the introduction, two to six catalytic residues within an active site perform the catalytic reaction of an enzyme (see red colored residues in Fig. 8.1). Usually, the spatial conformation of these residues is highly conserved for the same enzymatic reaction and can be recovered in evolutionary unrelated enzymes, as in the case of serine proteases and their Ser-His-Asp catalytic triad. The Catalytic Site Atlas (CSA)[40] stores a catalogue of catalytic residues as templates and provides the motif finder JESS[41] to search for the existence of the templates in a query protein structure. The JESS algorithm works by extracting constraint conditions from the template, which include the type of residues that are allowed to participate in a catalytic site and the allowed separations between these residues. The aim of JESS is to find residues in the protein structure that fulfill these constraints.

8.3.2 *Comparing Atomic Coordinates*

According to graph theory, a binding site can be regarded as a graph, with atoms being the nodes and the distance vectors between the atoms being the edges. A new association graph can be inferred using all atoms in both binding sites that are similar with respect to their physicochemical property and spatial location. Given such an association graph, the task is to find the maximum clique, i.e. the largest subset of nodes that are all connected with each other in a pair-wise manner. This problem is computationally demanding since every additional node increases the computation time with N^2. To reduce the complexity, the program IsoCleft[42] uses exclusively C-alpha atoms as a pre-filtering step, and only in a second stage, runs a more demanding all atom comparison. Another method, CavBase[43] (http://www.ccdc.cam.ac.uk/products/life_sciences/relibase/) uses a small set of pseudospheres to represent the location and physicochemical property of a binding site residue. In any case, once the maximum clique is found, the similarity between two binding site is assessed by relating the size of the maximum clique to the smaller binding sites.

The second approach to structural atomic comparisons is geometric hashing, which consists firstly of a preliminary preprocessing stage that runs offline only once and is followed by a recognition stage. In the preliminary stage, a database is created with a hash table for each binding site following four steps:

1. Three atoms being non-collinear to each other are picked out from a binding site. The triplet represents a plane in space from which an orthonormal reference frame can be built. The reference frame will help to describe the geometrical positions of the remaining binding site atoms independent from their original Cartesian coordinates.
2. Each remaining atom in the binding site is located within the triplet reference frame.
3. Representative information on the triangle together with location information of the fourth atom (quadruplet) is stored in a hash. If required, any other properties of the atoms can be added.
4. Repeat steps 1–3 for all other triplet combinations in the binding site.

Once the database of hash tables is built, the recognition stage can begin by applying the same approach as above to a query binding site. However, instead of storing the quadruplets in a hash table, they are checked for their existence in the database. Hash tables that exceed a user-defined minimum match value are considered as similar and are further analyzed for atom clusters.

8.3.3 *Comparing Binding Surfaces*

The molecular surface is crucial in intermolecular interactions as it is the interface through which the molecules interact. Different surface models exist for molecules with the two most important ones being the molecular surface and the solvent accessible surface. Both surfaces can be obtained by rolling a probe sphere over the van der Waals atom shells of a molecule. Whilst the inward-facing surface of the probe sphere produces the molecular surface, the solvent accessible surface is built by tracing the centre of the probe sphere. The radius of the probe sphere influences the appearance of both surfaces. A smaller probe sphere will show the surfaces in greater detail, whereas a larger probe sphere will reveal only major surface characteristics. Usually, the radius of a water molecule with 1.4 Å is used as the probe sphere radius.

Different representations exist for molecular surface models. The piecewise-quartic representation splits up the molecular surface into concave spherical triangles, saddle shape rectangles, and convex spherical regions. The Connolly dot representation spreads over the surface dots that allow a transparent view of the molecular surface. Another transparent representation is gained by tessellating the surface into linked empty triangles.

Although the visualization of molecular surfaces is well established, their comparison is just the opposite. Only a few attempts have been made to compare molecular surfaces. Their methodology is based mainly on the comparison techniques mentioned above, in which points on the molecular surface are compared using geometric hashing[44] or, as in the case of the publicly accessible eF-seek webserver (http://ef-site.hgc.jp/eF-seek/), using graph matching.

An elegant approach for surface comparison is to compare the coefficient vectors of binding site shapes that are approximated with spherical harmonics functions (see Section 8.2.3). The comparison between two shapes reduces to a Euclidean distance calculation between two coefficient vectors, with smaller distances for similar shapes.[8] Note that this approach is not fully comparable to the graph matching and geometric hashing methods mentioned above, as it compares the shape and not the molecular surface of the binding site. The volumetric shape represents not directly the molecular surface but the negative imprint of a binding site that is occupied by the ligand in the binding process.

8.3.4 *Other Comparison Methodologies*

Instead of describing the binding site properties specifically, FFF[45] (Fuzzy Functional Forms) explore to what extend properties can be relaxed and still allow a recognition of a binding site in a database scan.

pvSoar[46] (http://pvsoar.bioengr.uic.edu/) compares local sequence and geometric similarities of binding sites. It extracts the residues building up the wall of clefts from the CASTp[47] database and runs a sequence alignment to detect any highly conserved sequence patterns. In a second step, the geometric positions of the conserved residues are compared using a simple RMSD (see Equation 8.1) calculation.

8.4 Future Outlook

Even with the wide variety of identified binding site features and the methods described above, it remains difficult to correctly predict potential interactions between proteins and ligands. Drug discovery programs report some promising results on the prediction of interaction energies with *in silico* docking programs. However, in general, the prediction successes of docking and mapping applications remain rather moderate. The reasons are mainly the oversimplification of the physical conditions in the interaction process as well as persisting problems in recognizing the fundamental processes of molecular

recognition[48,49] (see Chapter 17 for an in-depth discussion on *in silico* docking).

The current concept of molecular recognition states that molecular binding occurs primarily due to complementary physicochemical properties between a binding site and its ligand. This hypothesis might require amendments, as more and more examples arise that show binding despite non-complementarity. The most striking examples occur in phosphate receptors (PDB Id: 1pbp), sulphate binding proteins (PDB Id: 1sbp), flavodoxin structures (PDB Id: 2fox), and DNase I structures (PDB Id: 2dnj).[50] All binding sites in these structures exert a negative electrostatic field over their binding sites despite binding a highly negative substrate. The question remains open whether in their evolutionary past these enzymes were binding ligands with complementary electrostatic potentials. Enzyme promiscuity might play a decisive role to answer this question. The current view on proteins, which is mainly governed by their specificity towards their functionality, is likely to change towards functional promiscuity, which states that a protein can exert different functions with the same active site. More and more enzymes are discovered that, despite their specificity, promiscuously catalyze other sometimes very different and unrelated reactions.[51] The increasing amount of data coming from growing 3D structure databases, annotations of catalytic mechanisms and in-depth binding site analyses will provide useful information to reveal the fundamental process in molecular recognition.

Structural information about enzymes have been derived mainly from X-ray crystal structures, which provide atom coordinates of unparallel high resolution. X-ray crystallography has one major drawback, which is that it provides only a static picture of an otherwise flexible protein. Protein dynamics and motions in crystals are usually only visible as a lack of "clarity" caused by the averaging process over many molecules. Molecular dynamics simulations attempted to overcome this obstacle by simulating motions in proteins using the X-ray structure as the starting point for their calculation. The steadily growing computer power, new developments of faster algorithms and better physicochemical parameterizations in recent years have improved

dynamic simulations. Soon, larger molecular dynamic simulations will be possible and hopefully allow a deeper investigation of the importance of protein dynamics in molecular binding.[52]

But most likely the explicit simulation of water molecules in and around proteins will have the biggest impact on our comprehension about molecular binding. Molecules are solvated in water and their interaction occurs in water. For many years, water was necessarily omitted in molecular docking and mapping applications as their *in silico* simulation was computational expensive. It was hoped that, in general, shape and physicochemical complementarity would be sufficient to drive molecular interactions. But many crystal structures of proteins show conserved water molecules at binding interfaces or next to binding sites and suggest an active role of water molecules in the protein-ligand complex.[53] Especially for molecular parts that interact via hydrophobic interactions by decreasing the entropy of the water molecule network, water acts as a "molecular glue" and induces the approach of protein and ligand molecules. The first methodologies that simulated hydration effects on protein structures considered water as a continuum, but had, in general, limited success. A second generation of simulation software treated water molecules explicitly but did not reached the expected accuracy especially due to the immense computational cost that dynamic simulations require. The growing computer power will eventually also help in this field to provide simulations of hydration effects under physical conditions.[53]

Once we achieve a comprehensive understanding of the fundamental processes in molecular binding, the *de novo* design of enzymes, i.e. the alternation of the enzymatic function, will be within reach. Other than inorganic catalysts, enzymes catalyze their reactions under mild conditions with high specificity and rate enhancements. This unique property makes enzymes attractive for many industrial processes although often they do not catalyze the required chemical reactions. Methods like rational-design and directed evolution in protein engineering have shown to be very useful in producing desired functionality in enzymes. As the factors for protein integrity namely, hydrogen bonds and hydrophobic effects, are well understood, many enzymes have been successfully altered to

stabilize the structural integrity against harmful chemicals, or extreme temperature and pH conditions. Comparable results could not be obtained for altering the catalytic machinery of enzymes.[54] Only few enzymes so far have been successfully altered, like the modification of an inert ribose-binding protein into a highly active triose-phosphate-isomerase.[55] As long as the general mechanisms of molecular binding and catalysis are little understood, such successful examples will remain rare. Improved understanding of the mechanisms for molecular binding will have an impact not only to function prediction in structural biology, but will also have effects within the fields of medicine and biotechnology.

References

1. Berman HM, Westbrook J, Feng Z, *et al.* (2000) The Protein Data Bank. *Nucl Acids Res* **28**: 235–242.
2. International Union of Biochemistry and Molecular Biology. Nomenclature Committee and Webb, E.C. (1992) *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Academic Press, San Diego, London.
3. Fischer E. (1894) Einfluss der configuration auf die wirkung der enzyme. *Ber Dtsch Chem Ges* **27**: 2985–2993.
4. Koshland DE. (1958) Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci USA* **44**: 98–104.
5. Brenner SE. (2001) A tour of structural genomics. *Nature Rev* **2**: 801–809.
6. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. (1996) Protein clefts in molecular recognition and function. *Protein Sci* **5**: 2438–2452.
7. Nayal M, Honig B. (2006) On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins* **63**: 892–906.
8. Kahraman A, Morris RJ, Laskowski RA, Thornton JM. (2007) Shape variation in protein binding pockets and their ligands. *J Mol Biol* **368**: 283–301.
9. Laskowski RA. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* **13**: 323–330.
10. Kraut DA, Sigala PA, Pybus B, *et al.* (2006) Testing electrostatic complementarity in enzyme catalysis: hydrogen bonding in the ketosteroid isomerase oxyanion hole. *Plos Biol* **4**: 501–519.
11. Coleman RG, Sharp KA. (2006) Travel depth, a new shape descriptor for macromolecules: application to ligand binding. *J Mol Biol* **362**: 441–458.

12. Gutteridge A, Thornton J. (2005) Conformational changes observed in enzyme crystal structures upon substrate binding. *J Mol Biol* **346**: 21–28.
13. Uversky VN, Gillespie JR, Millett IS, *et al.* (2000) Zn(2+)-mediated structure formation and compaction of the "natively unfolded" human prothymosin alpha. *Biocheml Biophys Res Commun* **267**: 663–668.
14. James LC, Tawfik DS. (2003) Conformational diversity and protein evolution — a 60-year-old hypothesis revisited. *Trends Biochem Sci* **28**: 361–368.
15. Domingues FS, Rahnenfuhrer J, Lengauer T. (2004) Automated clustering of ensembles of alternative models in protein structure databases. *Protein Eng Des Sel* **17**: 537–543.
16. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. (2002) Analysis of catalytic residues in enzyme active sites. *J Mol Biol* **324**: 105–121.
17. Glaser F, Pupko T, Paz I, *et al.* (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19**: 163–164.
18. Lichtarge O, Bourne HR, Cohen FE. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* **257**: 342–358.
19. Morris GM, Goodsell DS, Halliday RS, *et al.* (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**: 1639–1662.
20. Livesay DR, Jambeck, P, Rojnuckarin, A, Subramaniam, S. (2003) Conservation of electrostatic properties within enzyme families and superfamilies. *Biochemistry* **42**: 3464–3473.
21. Tsuchiya Y, Kinoshita K, Nakamura H. (2004) Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins* **55**: 885–894.
22. Kinoshita K, Furui, J, Nakamura H. (2002) Identification of protein functions from a molecular surface database, eF-site. *J Struct Funct Genomics* **2**: 9–22.
23. Honig B, Nicholls A. (1995) Classical electrostatics in biology and chemistry. *Science* **268**: 1144.
24. Klapper I, Hagstrom R, Fine R, Sharp K, Honig B. (1986) Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: effects of ionic strength and amino-acid modification. *Proteins* **1**: 47–59.
25. Elcock AH. (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. *J Mol Biol* **312**: 885–896.
26. Ondrechen MJ, Clifton JG, Ringe D. (2001) THEMATICS: a simple computational predictor of enzyme function from structure. *Proc Natl Acad Sci USA* **98**: 12473–12478.
27. Silberstein M, Dennis S, Brown L, Kortvelyesi T, Clodfelter K, Vajda S. (2003) Identification of substrate binding sites in enzymes by computational solvent mapping. *J Mol Biol* **332**: 1095–1113.

28. Davis AM, Teague SJ. (1999) Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis. *Angew Chem Int Ed* **38**: 736–749.
29. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. (1985) Hydrophobicity of amino acid residues in globular proteins. *Science* **229**: 834–838.
30. Kortemme T, Morozov AV, Baker D. (2003) An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol* **326**: 1239–1259.
31. Panigrahi SK, Desiraju GR. (2007) Strong and weak hydrogen bonds in the protein-ligand interface. *Proteins* **67**: 128–141.
32. Wallace AC, Laskowski RA, Thornton JM. (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng* **8**: 127–134.
33. McDonald IK, Thornton JM. (1994) Satisfying hydrogen bonding potential in proteins. *J Mol Biol* **238**: 777–793.
34. Laurie AT, Jackson RM. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **21**: 1908–1916.
35. Matsuzaki R, Fukui T, Sato H, Ozaki Y, Tanizawa K. (1994) Generation of the topa quinone cofactor in bacterial monoamine oxidase by cupric ion-dependent autooxidation of a specific tyrosyl residue. *FEBS Lett* **351**: 360–364.
36. Smith JJ, Thomson AJ, Proudfoot AE, Wells TNC. (1997) Identification of an Fe(III)-dihydroxyphenylalanine site in recombinant phosphomannose isomerase from *Candida albicans. Eur J Biochem/FEBS* **244**: 325–333.
37. Henrick K, Thornton JM. (1998) PQS: a protein quaternary structure file server. *Trends Biochem* Sci **23**: 358–361.
38. Bashton M, Nobeli I, Thornton JM. (2006) Cognate ligand domain mapping for enzymes. *J Mol Biol* **364**: 836–852.
39. Iyer N, Jayanti S, Lou K, Kalyanaraman Y, Ramani K. (2005) Three-dimensional shape searching: state-of-the-art review and future trends. *Comput-Aided Des* **37**: 509–530.
40. Porter CT, Bartlett GJ, Thornton JM. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucl Acids Res* **32**: D129–D133.
41. Barker JA, Thornton JM. (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* **19**: 1644–1649.
42. Najmanovich RJ, Allali-Hassani A, Morris RJ, *et al.* (2007) Analysis of binding site similarity, small-molecule similarity and experimental binding profiles in the human cytosolic sulfotransferase family. *Bioinformatics* **23**: e104–e109.
43. Schmitt S, Kuhn D, Klebe G. (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol* **323**: 387–406.

44. Rosen M, Lin SL, Wolfson H, Nussinov R. (1998) Molecular shape comparisons in searches for active sites and functional similarity. *Protein Eng* **11**: 263–277.
45. Fetrow JS, Skolnick J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J Mol Biol* **281**: 949–968.
46. Binkowski TA, Adamian L, Liang J. (2003) Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J Mol Biol* **332**: 505–526.
47. Binkowski TA, Naghibzadeh S, Liang J. (2003) CASTp: computed Atlas of Surface Topography of proteins. *Nucl Acids Res* **31**: 3352–3355.
48. Jain AN. (2006) Scoring functions for protein-ligand docking. *Curr Protein Peptide Sci* **7**: 407–420.
49. Sotriffer C, Klebe G. (2002) Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. *Farmacology* **57**: 243–251.
50. Ledvina PS, Yao N, Choudhary A, Quiocho FA. (1996) Negative electrostatic surface potential of protein sites specific for anionic ligands. *Proc Natl Acad Sci USA* **93**: 6786–6791.
51. Khersonsky O, Roodveldt C, Tawfik DS. (2006) Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr Opin Chem Biol* **10**: 498–508.
52. Karplus M, McCammon JA. (2002) Molecular dynamics simulations of biomolecules. *Nat Struct Biol* **9**: 646–652.
53. Levy Y, Onuchic JN. (2006) Water mediation in protein folding and molecular recognition. *Ann Rev Biophys and Biomol Struct* **35**: 389–415.
54. Bolon DN, Voigt CA, Mayo SL. (2002) *De novo* design of biocatalysts. *Curr Opin Chem Biol* **6**: 125–129.
55. Dwyer MA, Looger LL, Hellinga HW. (2004) Computational design of a biologically active enzyme. *Science* **304**: 1967–1971.
56. Apweiler R, Bairoch A, Wu CH, *et al.* (2004) UniProt: Rhe universal protein knowledgebase. *Nucl Acids Res* **32**: 115.
57. Holliday GL, Bartlett GJ, Almonacid DE, *et al.* (2005) MACiE: a database of enzyme reaction mechanisms. *Bioinformatics* **21**: 4315–4316.