Prediction of Protein-Ligand and Protein-Protein Interactions based on Local Surface Similarity

Claudio Garutti

Department of Information Engineering

University of Padova

A thesis submitted for the degree of

PhilosophiæDoctor (PhD)

February 2nd, 2008

Abstract

The three-dimensional structure of a protein determines its function. This thesis describes a suite of methods for the problem of protein binding site recognition, based on a spin-images representation of the molecular surface. A procedure for cavity detection is coupled with a procedure for the recognition of similar regions in two proteins, and applied to the comparison of two protein's cavities, the all-to-all pairwise comparison of a set of cavities, and the recognition of multiple binding sites in one cavity. All the presented methods can be used to screen large collections of proteins.

Abstract

La struttura tridimensionale di una proteina determina la sua funzione. Questa tesi descrive una suite di metodi per il problema del riconoscimento di siti di legame di proteine, basati su una rappresentazione a spin-images della superficie molecolare. Una procedura per l'identificazione di cavita' integrata con una procedura per il riconoscimento di regioni simili in due proteine, e applicata al confronto delle cavita' di due proteine, il confronto *all-to-all pairwise* di un insieme di cavita', e il riconoscimento di siti di legame multipli in una cavita'. I metodi presentato possono essere usati per analizzare vaste collezioni di proteine.

Contents

1 Introduction to the Dissertation **1**

1.1 Proteins . . . 1

1.1.1 Ligand Binding Site . . . 1

1.2 Protein-Ligand and Protein-Protein Interactions . . . 3

1.2.1 Features . . . 3

1.2.2 Effects of Dynamics on Topology . . . 4

1.2.3 Local Surface Similarity . . . 5

1.3 Protein Representation . . . 7

1.3.1 Connolly's Molecular Representation . . . 7

1.3.2 Spin-Images . . . 9

1.3.3 Spin-Image Representation . . . 9

1.3.3.1 Oriented Point Basis . . . 9

1.3.3.2 Quantization of spin-images . . . 11

1.3.3.3 Spin-Image Generation Parameters . . . 12

1.3.4 Protein Surface Spin-Image Representation . . . 14

2 Discovery of Similar Regions on Protein Surfaces **17**

2.1 Introduction . . . 18

2.2 Surface Point Labelling . . . 19

2.3 Spin image profiles . . . 19

2.3.1 Horizontal and vertical profiles of a surface point . . . 19

2.3.2 Amended spin image . . . 20

2.4 Matching shapes . . . 21

2.4.1 Finding regions of similarity on two protein surfaces . . . 21

2.5 Data and results . . . 24

2.5.1 Benchmark Protein-ligand complexes . . . 25
2.5.2 Proteins interacting with other proteins . . . 32
2.5.3 Running times . . . 33
2.6 Conclusions . . . 34

3 Cavity Detection and Matching for Binding Site Recognition 35
3.1 Introduction . . . 36
3.2 Previous work . . . 37
3.2.1 Cavity Detection . . . 37
3.2.2 Binding Site Detection . . . 37
3.3 Characterizing cavities in terms of blocked points . . . 38
3.4 Methods . . . 39
3.4.1 Cavity detection . . . 39
3.4.2 Cavity Matching . . . 41
3.4.3 All-To-All Pairwise Cavity Comparison . . . 43
3.4.4 Background Cavity . . . 44
3.5 Data and results . . . 45
3.5.1 Cavity Detection . . . 45
3.5.2 Cavity Matching . . . 51
3.5.3 All-To-All Pairwise Cavity Comparison . . . 52
3.5.4 Background Cavity . . . 55
3.6 Conclusions . . . 55

4 MolLoc Web Server: a Tool for Local Molecular Surface Alignment 57
4.1 Molecular Surface . . . 57
4.2 Input . . . 58
4.3 Output . . . 60
4.4 FastCav: fast cavity detection . . . 60
4.5 Alignment Methods . . . 62
4.6 Conclusions . . . 64

5 Conclusions 65

Bibliography 67

1

Introduction to the Dissertation

In this dissertation I use computer vision methods to extract information from protein structures on the way they interact with other molecules. The results of this research have been published in three articles: (9), which corresponds to Chapter 2, and (10) and (11), which have been aggregated in chapter 3. Chapter 4 describes a web server that implements these methods, and Chapter 5 has the conclusions.

1.1 Proteins

A protein is an organic compound made of amino acids arranged in one or more linear chains and joined together by peptide bonds between the carboxyl and amino groups of adjacent amino acids. Proteins carry out many vital cellular functions, and the interactions of proteins with other molecules, like ions, small molecules, RNA, DNA, or even other proteins, are critical for these functions.

1.1.1 Ligand Binding Site

When a protein interacts with another molecule, the region of the protein that is in contact with the molecule is called *interface* (or *binding site*) of the protein with the molecule, and is that part of the protein where the molecule is close enough to form stabilizing chemical interactions.

A *ligand* is a chemical substance that is able to bind to a protein, thus forming a complex that plays a role in a biological process (ex. signal transduction). Here a ligand is intended as a small molecule with respect to the protein. The binding

between the ligand and the protein is usually due to "weak" intermolecular forces such as hydrogen bonds and van der Waals forces, while "strong" intramolecular forces like covalent bonds are rare. Therefore, the binding (or *docking*) between the ligand and the protein is usually reversible. The three-dimensional structures of a protein and a molecule before they interact (*unbound conformation*) is different from their structures after binding (*bound conformation*), since all biomolecules are flexible to a certain extent, and they undergo conformational change upon docking, in order to minimize the free energy of the complex.

When a ligand binds to a protein, it is possible to measure the *affinity* of the binding, which is proportional to the strength of the involved chemical bonds. The higher the affinity of the binding, the stronger the forces on the bonds and thus the longer the ligand will be hosted in the binding site. Moreover, if the affinity is high, a relatively low concentration of the ligand is necessary to occupy the binding site.

Ligands can also be *selective*, meaning that they bind to very few types of proteins, or *non-selective*, meaning that they can bind several types of proteins. For example, m2-toxin is a selective molecule that binds to M2 muscarinic receptors (14), while ATP is a non-selective molecule that binds to a wide variety of different proteins (62).

Usually the targets of pharmacological approaches belong to two important classes of proteins, the *enzymes* and the *receptors*. An enzyme is a protein that catalyzes (i.e. increases the rate of) a chemical reaction, while a receptor is a protein that is located in the plasma membrane or cytoplasm of a cell, and that is involved in a signaling that induces a cellular response.

A ligand is called *inhibitor* if it binds to an enzyme and decreases its activity. For example, ritonavir is an inhibitor for HIV protease. An *antagonist* is a term used to describe a ligand that binds and alters the activity of a receptor. For example, IL-1Ra is an antagonist for the interleukin-1 receptor (2), which is involved in the inflammatory response of the body against infection.

Selectivity, inhibitors and agonists are important concepts for pharmacology, that aims at producing selective ligands that specifically inhibit enzymes or are agonists for receptors that are involved in a biological process that causes a disease.

In fact, an inhibitor or agonist that is non-selective binds also to proteins involved in beneficial pathways, thus altering their functions and producing side effects on the health of the hosting organism.

1.2 Protein-Ligand and Protein-Protein Interactions

Currently, experimental techniques such as bimolecular fluorescence complementation (BiFC), yeast two-hybrid, tandem affinity purification (TAP) and many others can be used for high-throughput screening of protein interactions. The high throughput screening returns true interactions and false positives. Co-immunoprecipitation is used to distinguish the true interactions from the false positives, but it's unsuitable for high-throughput screening, since it requires the isolation of the single proteins by means of specific antibodies. Furthermore, these techniques don't define how the proteins interact at the atomic level, which is a crucial information that is required for the understanding of cellular function, of disease-related processes and for the rational design of drugs. The experimental determination of the 3D structure of the protein–protein and protein–ligand complex is performed using NMR, X-ray crystallography and electron microscopy. Again, to date these techniques cannot be used for high-throughput determination of the atomic coordinates of protein–protein or protein–ligand complexes, and there are still many protein families (ex. membrane proteins) that are extremely hard to process. Therefore, computational methods for the determination of putative binding modes in protein–protein and protein–ligand complexes are extensively used. The next paragraph describes the differences between protein–protein and protein–ligand interactions.

1.2.1 Features

Protein-ligand and protein–protein complexes share some structural and physicochemical features and differ substantially in others (22).

On one hand, protein–protein interfaces are often flat, and their size is usually in the range $1600 \pm 400 \mathring{A}^2$, with a few complexes exhibiting very large (2000-4660$\mathring{A}^2$) or very small (less than $1000 \mathring{A}^2$) interfaces (19). Moreover, protein-protein interfaces are enriched in aromatic residues and in water molecules; on average, one water molecule per 100 $\mathring{A}^2$ is found in high-resolution structures (19). It is worthwhile noting that these water molecules have both a structural function, by contributing to the tight docking of the protein-protein interface region, and an energetic role, by mediating intermolecular hydrogen-bond formation. A study on three crystal structures of protein-protein

complexes revealed that water molecules contribute around 25% of the total calculated binding strength (21).

On the other hand, protein–ligand interfaces (or binding sites) usually lie in a protein's cavity, and more than 80% of the protein–ligand binding sites are located in one of the four biggest cavities (25). As the binding partners are smaller than in the case of protein–protein interactions, the interfaces are also smaller. Moreover, a study on 175 enzymes indicates that the residues of the binding site are very often found among the 5% of residues closest to the enzyme centroid (4). Water can play a significant role also in protein–ligand binding. For example, in the human chaperone hsp90, the interaction with radicicol is coordinated by three molecules of water (20).

Both protein-protein and protein–ligand binding energetics are also influenced by the conditions under which binding takes place, like the pH. Finally, the viscosity of the solvent can affect the binding kinetics and thus alter binding affinity (66).

1.2.2 Effects of Dynamics on Topology

One of the major difficulties in predicting the structure of the bound complex from the structures of two unbound molecules is due to the structural changes that take place upon binding. These changes can arise for several reasons. First, the need to establish or improve specific interactions between the two binding partners, to improve their geometric fit at the interface and to avoid any steric clashes between them. Second, the inherent flexibility of the molecules involved. Third, for functional reasons, such as conformational changes that can trigger events in signaling or are important in allosteric effects.

On one hand, local structural changes, such as surface side-chain rotations, occur on time scales of 10^{-11} to 10^{-10} s and length scales of several angstroms. On the other hand, large-scale structural changes occur on time scales of 10^{-11} to 10^{-3} s and can involve motions over several tens of angstroms (46).

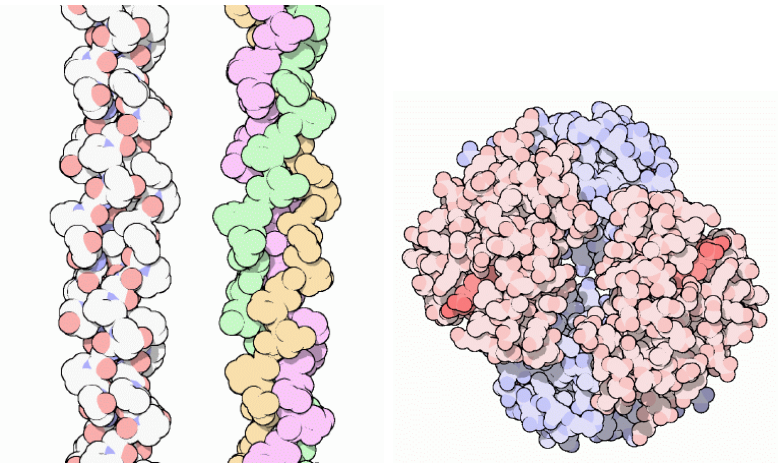
Recent docking and MD simulation studies suggest that the protein–protein docking is at least a two-step process (52). The first step is an initial collision between the two proteins, when recognition takes place through desolvation and burial of key *hot spot* anchor residues at the center of the nascent interface, whose conformations do not significantly change on binding. The second step is a latching phase in which peripheral interface residues adjust their rotameric conformations into complementary

arrangements. There might be also a final induced fit step in which interface side chains adjust their torsion angles to adopt off-rotamer conformations and interfacial waters become frozen into their crystallographically observable positions.

1.2.3 Local Surface Similarity

Protein docking isn't the only approach to discover protein function. Structural similarity between two proteins can give the same information, if the two proteins are found similar and if one of the two proteins has already a functional characterization.

In fact, the function of a protein is mostly determined by its three-dimensional structure. For example, collagen has a super-coiled helical *fold* (i.e. global shape), that is long and resistant to mechanical stress, which makes it a structural protein that provides physical support to the cell. On the other hand, hemoglobin has a globular fold, and its spherical and compact structure makes it a transport protein, that moves molecules around the organism (see fig.1.1).



(a) (b)

Figure 1.1: Structures of the structural protein collangen (a) and the transport protein hemoglobin (b). Source: the Protein Data Bank, *Molecule of the Month* Archive.

A pletora of computational methods for global alignment have been applied to novel proteins with unknown function in order to find proteins with known function and similar fold, thus trying to infer the function of the novel protein from the ones with similar fold (34).

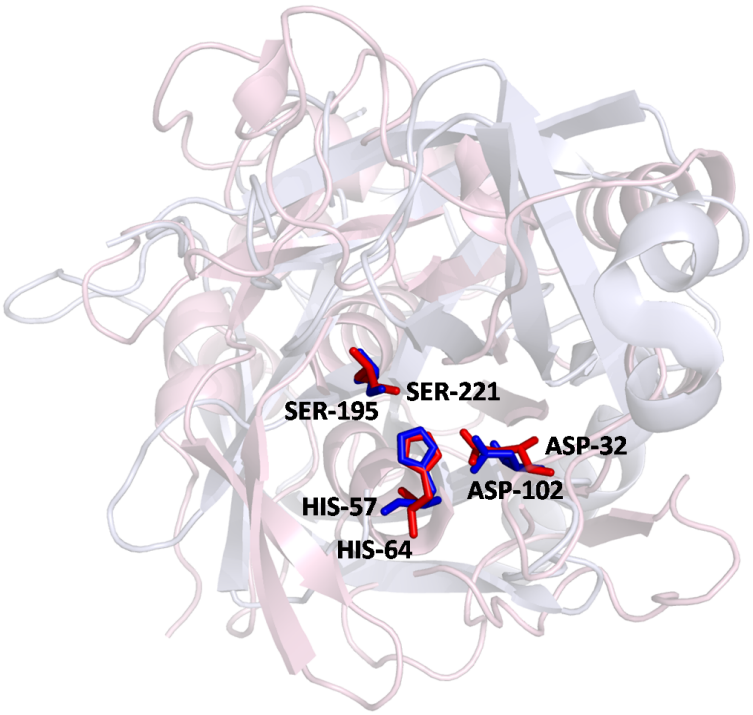


Figure 1.2: Chymotrypsin in lightblue and blue, and subtilisin in darksalmon and red. The two proteins have different folds, but the superimposition shows a local similarity of the catalitic triad Ser-195, His-57, Asp-102 in chymotrypsin and Ser-221, His-64, Asp-32 in subtilisin.

However, the fact that two proteins have the same fold isn't a necessary nor sufficient condition for them to have the same function. A well-known example is that of TIM-barrel fold, that is possessed by different proteins that provide at least 60 different functions. On the other hand, non-homologous proteins can have the same function. For example, chymotrypsin [1] and subtilisin [2] are two non-homologous proteins with different folds but with the same function (proteinase). As shown in fig.1.2, these two proteins have different global structures but are locally similar in the triad Ser-His-Asp, which is also the region that is devoted to the breakdown of other molecules, i.e. the region that is responsible for the function of the two proteins.

Therefore, there is the need for computational methods that can find local structural similarities between two proteins. In fact, finding structural similarities between two proteins in their binding site regions can provide useful information about the molecules that the proteins can bind and about the proteins function as well.

1.3 Protein Representation

Since the three-dimensional structure of a protein determines its function, a convenient representation is crucial to the success of computer-based methods that deal with proteins. Many different representations have been used so far, depending on the level of detail required by the application. Some methods represent the protein as a set of vectors, where a vector is a secondary structure (α-helix or β-sheet) (16; 23); several methods take one point, usually the center of the C_α carbon, for each residue (27; 56; 63); others take one point for each atom center (33; 49).

However, the highest level of detail is reached when using the protein surface. The following section describes in detail the protein surface representation that has been used for the work described in this thesis.

1.3.1 Connolly's Molecular Representation

The simplest space-filling representation of a protein is the model where each atom is represented by a sphere, whose center is the center of the atom, and the radius is its van der Waals radius. Thus, the surface of the molecule is made up of the parts of the

[1] Pdb id: 5cha
[2] Pdb id: 5sic

van der Waals surface of each atom that do not lie inside any other atom. The problem

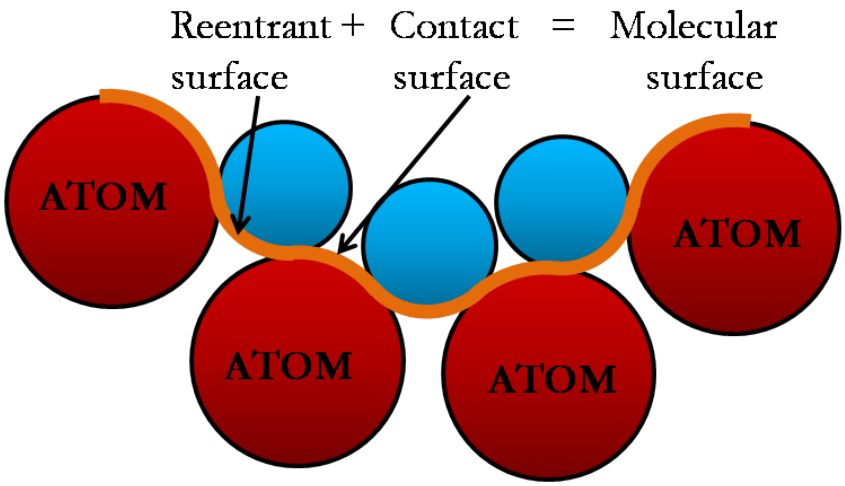


Figure 1.3: Connolly's molecular surface. In red the protein's atoms, in blue the probe.

of this representation is that much of the van der Waals surface of a protein is on the inside of the protein, so this is not suitable for studying the binding of other molecules onto protein surfaces. The outer surface of a protein molecule can be identified (41) by rolling a probe sphere over a molecule. The surface of a molecule is thus defined to be the part of the molecule that is accessible to solvent, where the solvent molecule (water) is represented by a sphere of radius in the range 1.4-1.7 Å. In this way, the volume of space that the probe is excluded from by collisions with the atoms of the molecule doesn't contribute to the molecular surface. The *accessible surface* is the trace of the probe sphere center as it rolls over the protein. The points of the accessible surface are generated in correspondence of the probe ball center. The *contact surface* is that part of the van der Waals surface of the atom that can be touched by a probe sphere. The *reentrant surface* is the inward facing part of the probe sphere as it is touching more than one atom. Together, the contact surface and the reentrant surface form a continuous sheet, the *molecular surface* or *Connolly's surface*(fig.1.3). The contact surface is made up of convex spherical regions, while the reentrant surface is made up of saddle-shaped rectangles and concave spherical triangles. Because the saddle-shaped rectangles are part of the surface of a torus, which is defined by a fourth-degree polynomial, the surface is classified as a piece-wise quartic.

1.3.2 Spin-Images

This section describes the concept of *spin-image* (30), which is a popular representation from computer vision that can be applied to every three-dimensional object that can be described as a cloud of points. It is a data level representation of surfaces, used for *surface matching*, the process that compares surfaces and decides whether they are similar or not.

The main difficulty of surface matching is that the coordinate system where to compare the two surfaces is undefined. Thus, a common approach to this problem is to transform the surfaces being compared into representations where the comparison is easier. Spin-images representation is object-centered, and, since spin-images are constructed with respect to specific surface points, it can also be used to find correspondences between points belonging to similar regions on two surfaces, and then to align the surfaces on the similar regions.

1.3.3 Spin-Image Representation

1.3.3.1 Oriented Point Basis

An *oriented point* is a three-dimensional point with an associated normal direction. It is used to create spin-images, once the surface of the object has been mapped into points. An oriented point O is defined in correspondence of a point p of the object's surface using the 3-D position of that point and its surface normal n. As shown in fig.1.4, an oriented point defines a local basis (p, n), that is to say a cylindrical coordinate system with five degrees of freedom (because the polar angle coordinate cannot be determined using just surface position and normal).
The basis is obtained by the tangent plane P through p oriented perpendicularly to n, and by the line L through p parallel to n. The two coordinates of the basis are α, which is the perpendicular distance to the line L, and β, which is the perpendicular distance to the plane P. A *spin-map* $S_{\mathcal{O}}$ is defined as the function that projects a 3-D point x to the 2-D coordinates of a particular basis (p, n) corresponding to the oriented point O as follows:

$$\begin{gathered} S_{\mathcal{O}} : \mathbb{R}^3 \to \mathbb{R}^2 \\ S_{\mathcal{O}}(x) \to (a, b) = (\sqrt{\|x - p\|^2 - (n \cdot (x - p))^2}, n \cdot (x - p)). \end{gathered} \tag{1.1}$$

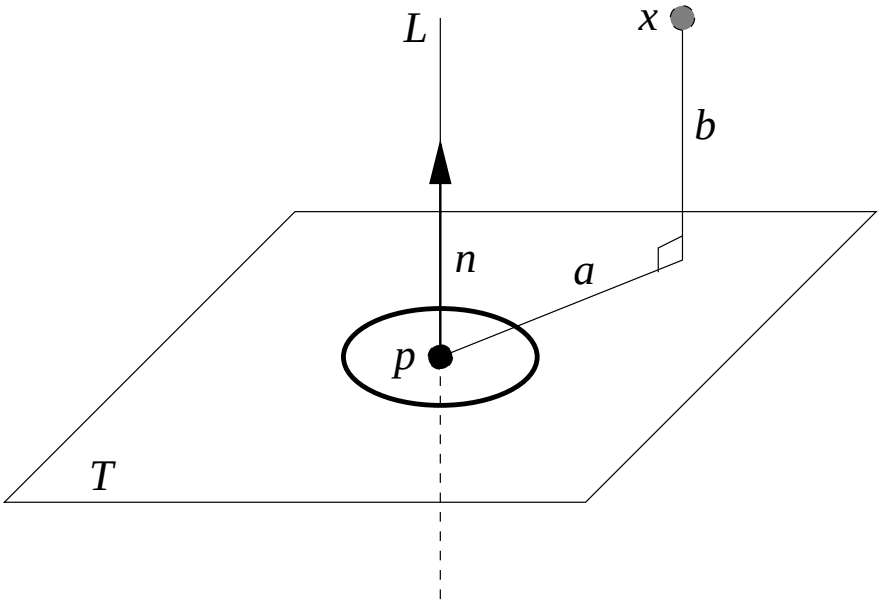


Figure 1.4: An oriented point basis. Source: (29).

Obviously, α cannot be negative, while β can be both positive and negative. The term spin-map comes from the cylindrical symmetry of the oriented point basis. The basis can spin around its axis with no effect on the coordinates of points with respect to the basis. A consequence of the cylindrical symmetry is that points that lie on a circle parallel to P and centered on L will have the same coordinates (α, β).

Oriented points are object-centered coordinate systems, that is to say, they describe the shape of an object with respect to a coordinate system attached to the object. The main advantage of these system is that the description is independent from the position that the object assumes into the space.

Each oriented point O on the surface of an object determines, by means of p and n, a unique spin-map $S_{\mathcal{O}}$ associated with it. When $S_{\mathcal{O}}$ is applied to all the points of the surface, a set of 2-D points is created. The image of a spin-map is a description of the shape of the object. This is because it's the projection of the relative position of 3-D points that lie on the surface of an object to a 2-D space where some of the 3-D metric information is preserved. The image of a spin-map describes the relative position of points on a rigid object with respect to a particular point on the same object, and so the image is independent on rigid transformations applied to the object, and is an object-centered shape description.

It is possible to compare spin-maps images of points belonging to the objects that have to be compared, in order to discover if they are similar or not. That is to say, given two objects, let P be an oriented point of the first object, and Q an oriented point of the second object. If the image of the spin-map of P is similar to that of Q, these two points are put in correspondence. Then this procedure can be repeated for all the points of the first object and all the points of the second. With three or more point correspondences, a rigid transformation between the two objects can be calculated and verified, and therefore it can be established if the objects are similar or not.

The main drawback of a spin-map's image is that it describes the shape of an object by means of the exact position of the points on the surface of the object. Thus, even with two very similar images, the position of the points may be slightly different between an image and the other one, and this leads to the problem of finding an appropriate way to compare the images. That is to say, when comparing images, the goal is to compare global shape conveyed by the two images while not being distracted by local variations.

1.3.3.2 Quantization of spin-images

It is necessary to find a new representation for the images of spin-maps, in order to keep the information on the spatial distribution of the points on the surface, but without the exact position of the points.

The solution adopted is the following. The image, which is a set of points in a plane, is divided into a grid (fig.1.5), with cells of appropriate size. Then an hash table is built, with the same dimension of the grid, and a correspondence is established between the cells of the grid and the bins of the hash table. Finally, focusing on a cell, the number of points that fall is counted and stored in the corresponding bin. This phase of the procedure is repeated for every cell of the grid.

By placing the indices in discrete bins of a hash table, the effect of the exact position of individual points on matching is reduced through averaging. Every bin stores information about the density of points in a certain region of space, thus giving a more flexible instrument to perform compare oriented points, and then objects. This new representation of the image of a spin-map is called *spin-image*.

The spin-image generation process can be visualized as a sheet spinning around the oriented point basis, accumulating points as it sweeps through space.

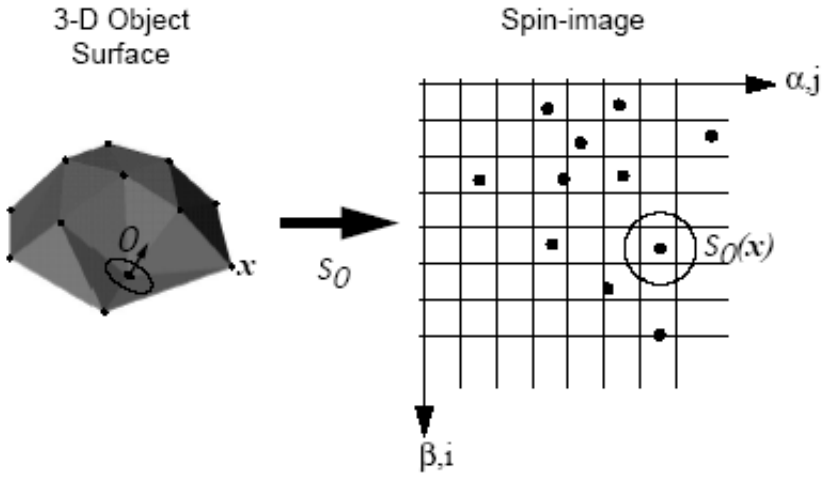
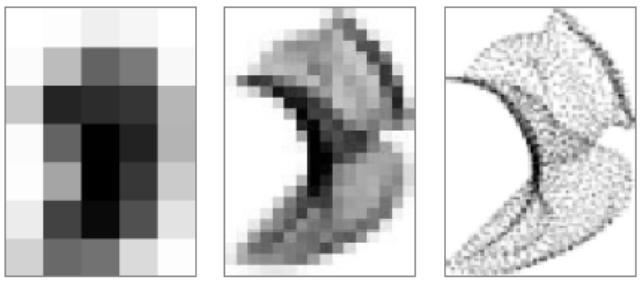


Figure 1.5: Grid on the image of a spin-map. Source: (29).

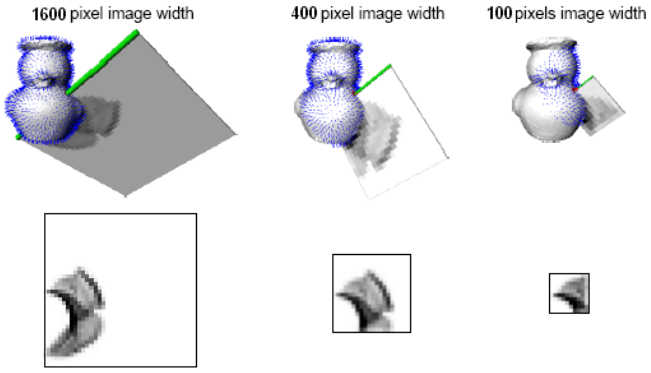
1.3.3.3 Spin-Image Generation Parameters

The first parameter that control spin-image generation is the *bin size*, that determines the storage size of the spin-image, and consequently the averaging in spin-images that reduces the effect of individual point positions. It affects also the descriptiveness of the spin-images. Fig.1.6(a) shows spin-images generated with different size for an oriented point lying on the surface of a duckie model. The spin-images are generated with increasing resolution from left to right. This means that the first from the left is generated with the biggest bin size, that is to say the size of the cell of the grid is the biggest, then the cell's size of the second decreases and finally the third is generated with the biggest bin size, that is to say the size of the cell of the grid is the smallest. The first is not very descriptive of the global shape of the model, the third does not have enough averaging to eliminate the effect of surface sampling, while the second has the proper balance between encoding global shape and averaging of point positions. For protein surfaces, that are generated from atoms with a maximum radius of 2Å, a bin size of 1Å is a good compromise between precision and averaging.

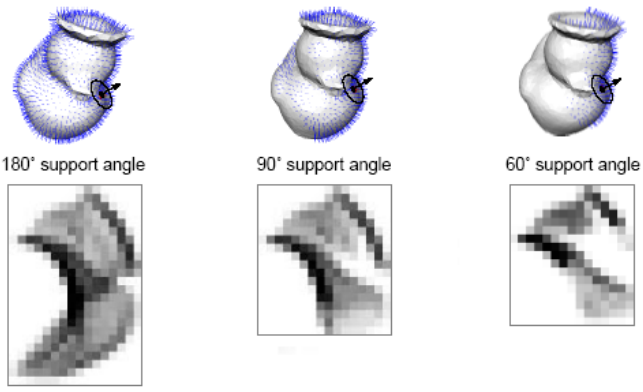
A second parameter is *image width*, which is the number of rows and columns of the spin-image. The smaller the image width, the more local the representation (see fig.1.6(b)). When dealing with global alignment, the image width is big enough to fit all the points of the object. On the other hand, for local alignment, the spin-image of



(a)



(b)



(c)

Figure 1.6: a) different bin sizes for the spin-image of a duckie model; b) different image widths for the spin-image of a duckie model; c)different support angles for the spin-image of a duckie model. Source: (29).

an oriented point P should retain just a set of points around P. Therefore, the image width is taken big enough to fit the smallest region that is searched as similar, but small enough to lose the information about the global shape of the object. Considerations of the same quality apply also if the spin-image is not square, i.e. if the number of rows is different from the number of columns. A related parameter is the *support distance*, which is defined as image width per bin size, and determines the amount of space swept out by a spin-image.

A last parameter is the *support angle*, which is the angle between the normal of O, the point that determines $S_{\mathcal{O}}$, and the normal of x, where x is the generic point on the surface of the object (see fig.1.6(c)). If the support angle is greater than an established maximal value, the point x is rejected during the creation of the spin-image of O. The support angle is used to limit the effect of some drawbacks that occur in analyzing real scenes, such as self occlusion and clutter. Self occlusion may occur because real scenes can give just a partial view of an object. Clutter can occur in scenes that contain multiple objects. However, proteins are ideal models where clutter and self occlusions do not occur, and thus the support angle is always set to 360 degrees.

1.3.4 Protein Surface Spin-Image Representation

Protein surface is represented first using a cloud of points that belong Connolly's surface. Then, for each point, the correspondent spin-map is built. Once the size of the cells of the grid is fixed, the spin-image is built on the *smallest* grid that contains all the points of the spin-map. Thus, the first and last row, and the first and last column of the grid (and consequently of the spin-image), contain at least one non-empty cell (bin). The spin image dimensions depend on the point P and its corresponding tangent plane and corresponding normal n of P to its tangent plane. The number of columns depends on the maximum distance from n of other points on the surface of the protein that are above the tangent plane of P. The number of rows depends on the maximum height of other points on the surface which are above the tangent plane of P. In this work, generated Connolly's surfaces are generated with density $D = 1$ point per Å^2. Since the X-ray resolution of most of the protein stuctures currently available is above 1.5 Å, this value of D guarantees enough precision. An example of protein surface spin-image representation is shown in fig.1.7.

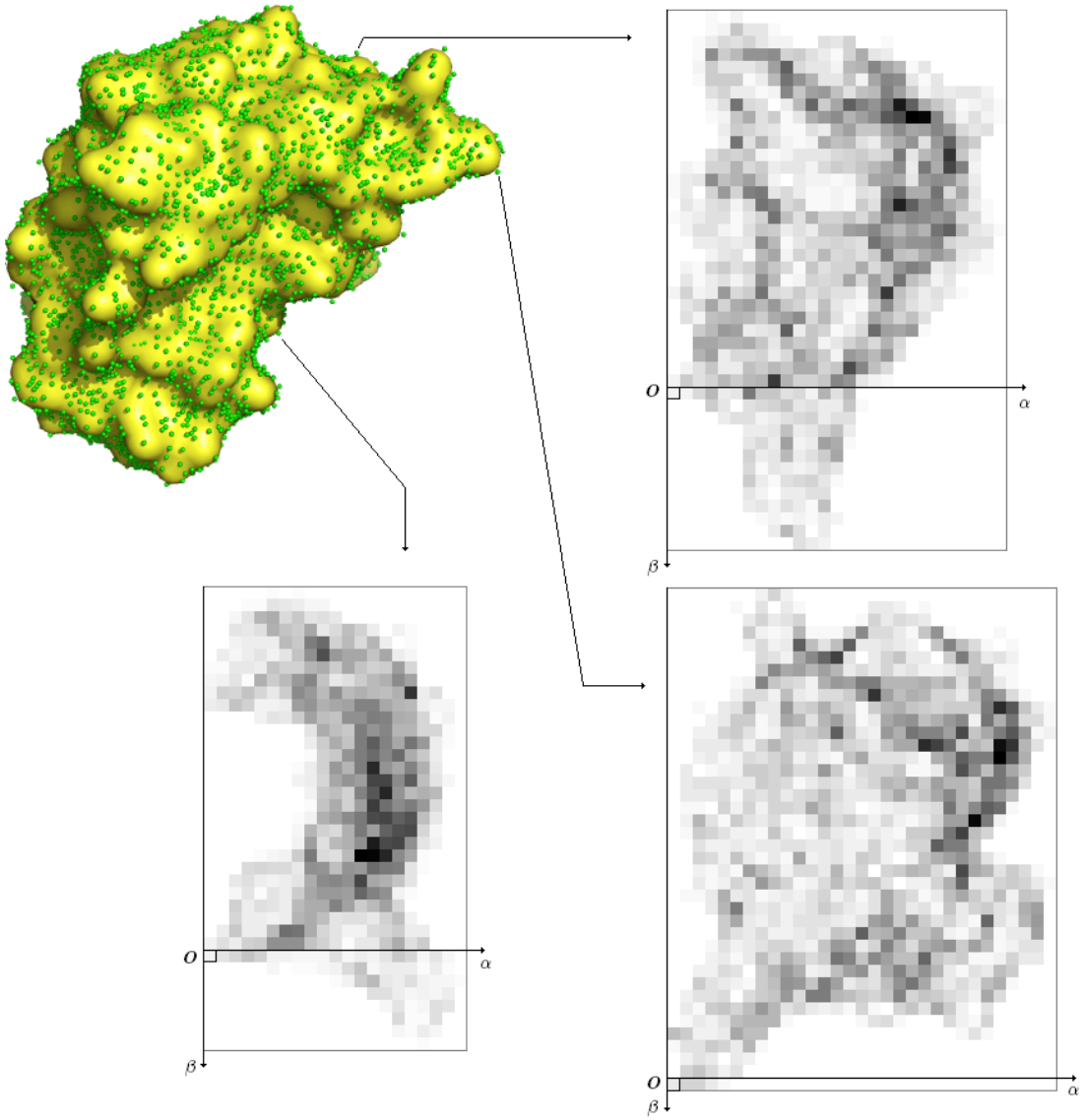


Figure 1.7: Examples of spin-images of a protein surface.

2

Discovery of Similar Regions on Protein Surfaces

Discovery of a similar region on two protein surfaces can lead to important inference about the functional role or molecular interaction of this region for one of the proteins if such information is available for the other. This chapter proposes a characterization of protein surfaces based on a spin image representation of the surfaces that facilitates the simultaneous search of the entire surface of each of two proteins for a matching region. For a surface point, the new concept of *spin image profile* is introduced. Spin image profile is related to the degree of exposure of the point to identify structurally equivalent surface regions in two proteins. Unlike some related methods, here is not assumed that a known fixed region of one of the proteins surfaces is to be matched on the other protein surface. Rather, the new method discussed in this chapter and called *MolLoc* searches for the largest similar regions on each of the two molecular surfaces. In spite of the fact that this approach is entirely geometric and no use is made of physicochemical properties of the protein surfaces or fold information, it is effective in identifying similar regions on both surfaces even when the region corresponds to a binding site on one of the proteins. Experimental results from datasets of more than 50 protein surfaces are presented.

2.1 Introduction

The detection of structural similarities between regions on the surfaces of proteins is of interest in molecular biology, as discussed in chapter 1. If the surface region of one protein is similar to that of the ligand binding site of another protein with known function, the function of the one protein can be inferred and its molecular interaction with the ligand predicted. Much work has been done on the analysis of the binding sites of proteins and their identification (24; 31; 32; 33; 45; 48; 58; 64) using various approaches based on different protein representations and matching strategies. Different instances of the surface shape matching problem have been considered in the literature:

1. given two protein surfaces find similar patches on the two surfaces (8);

2. for a given binding site on a first protein, find the surface region of a second protein most similar to the given binding site (3; 15; 39; 53; 58; 67);

3. given binding sites for numerous proteins, the sites are compared and classified (48; 58).

This chapter considers the first formulation of the problem and proposes an approach to identify regions of similarity based on a spin-image representation of molecular surfaces. A protein is described by a collection of spin images, each associated to a Connolly point.

MolLoc approach consists of finding point correspondences on the two surfaces based on the correlation of their associated spin image profiles as well as on the correlation of the 2D spin images. Regions of similarity on the protein surfaces are obtained by grouping the obtained correspondences into sets of geometrically consistent correspondences. The results on different datasets of proteins show that this approach performs well also when the regions of similarity correspond to almost flat surfaces.

The chapter is organized as follows. Section 2.2 presents a labeling of the surface points based on spin images. Section 2.3 introduces the spin image profile. Section 2.4 considers the general problem of matching surfaces based on the spin image similarity. Experimental results are presented in section 2.5, and conclusions in section 2.6.

2.2 Surface Point Labelling

Here is described a labelling scheme for protein surface points based on the spin images. This labelling allows one to speed up the matching procedure by restricting correspondences to points with the same label.

A protein surface point P is labelled as blocked or unblocked depending on whether or not the normal n at P oriented outwards intersects the protein surface at another point above the tangent plane T at P and perpendicular to n. The label of point P is computed from the information stored in the spin image. Since spin images use a discrete representation of the space, the above definition is modified as follows. A point P is labelled as *blocked* if there is at least another surface point lying above the tangent plane T that is within ϵ distance from normal n, where ϵ is the spin image pixel size (1Å). In other words, there is at least one surface point that in the reference frame of a blocked point has coordinates (a, b), with $a = 0$ and $b > 0$. This implies that only the first column (corresponding to $a = 0$) of the spin image needs to be examined for labelling: if it contains a non-zero pixel with positive b, then the point is blocked, otherwise it is *unblocked*. Examples of blocked and unblocked points and their corresponding spin images are shown in Figure 2.1, where (α, β) are the column index and row index, respectively, and the origin O is the cell $(0, 0)$. The images are displayed with darker pixels corresponding to higher accumulator values.

2.3 Spin image profiles

2.3.1 Horizontal and vertical profiles of a surface point

The *horizontal profile* of a spin-image of a point is the one-dimensional array whose element i is given by the number of contiguous zero-elements in row i, ($i \geq 0$, corresponding to $b = i$) of the spin image array starting at column 0 (corresponding to $a = 0$) and ending at the first non-zero cell along row i. See Figure 2.2(a) for an example of horizontal profile. The *vertical profile* of a spin image of a point is the one-dimensional array whose element i is given by the number of contiguous zero-elements in column i (corresponding to $a = i$) of the spin image array starting at the last row (corresponding to $\lceil \beta_{max}/\epsilon \rceil$) and ending up at the first non-zero cell along column i from the bottom. See Figure 2.2(b) for an example of vertical profile.

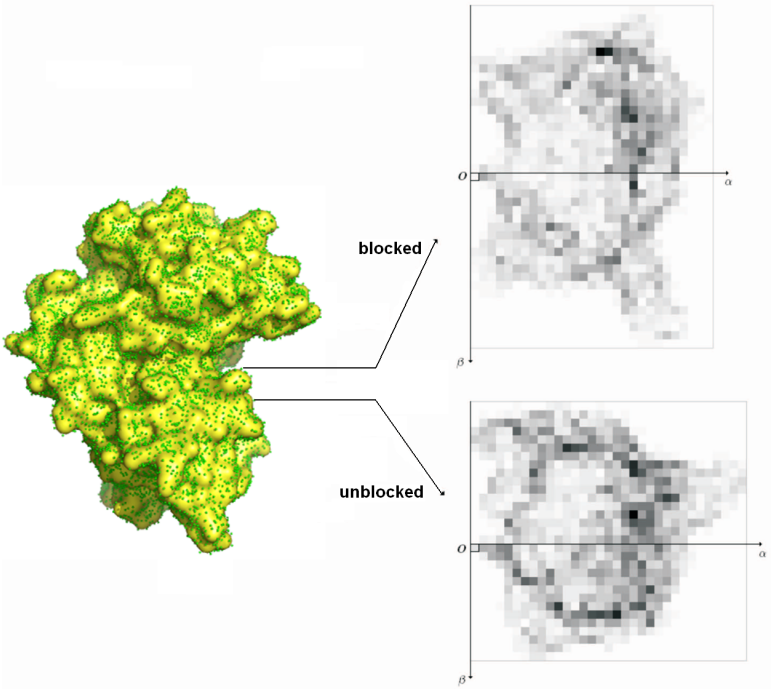


Figure 2.1: Blocked and unblocked surface points and corresponding spin-images.

2.3.2 Amended spin image

In global matching, a spin-image describes the whole protein surface from the oriented point that generates it. However, the problem of finding similar patches on two protein surfaces is a problem of local matching, and as discussed in chapter 1, in this case the spin-image of an oriented point P should retain just a set of points around P.

Therefore, the *amended spin image* of a point imposes limitations on the size of the spin image. Let Z be the count of the number of successive zero elements along the $a = 0$ column of the spin-image for $b \geq 0$ starting $b = 0$.

For the amended spin image of a blocked point, only surface points with b value between $-\xi$ and $Z + \xi$, where ξ is a threshold set to 8 Å, are included. The amended spin image of a blocked point includes only surface points with value for a less than ξ plus an integer equal to the largest value of the horizontal profile among the first Z elements.

For the amended spin image of an unblocked point, only surface points with b value greater than $-\xi$ and a value smaller than $\min\{20, a^* + \xi\}$, where a^* is the last value in the horizontal profile, are included.

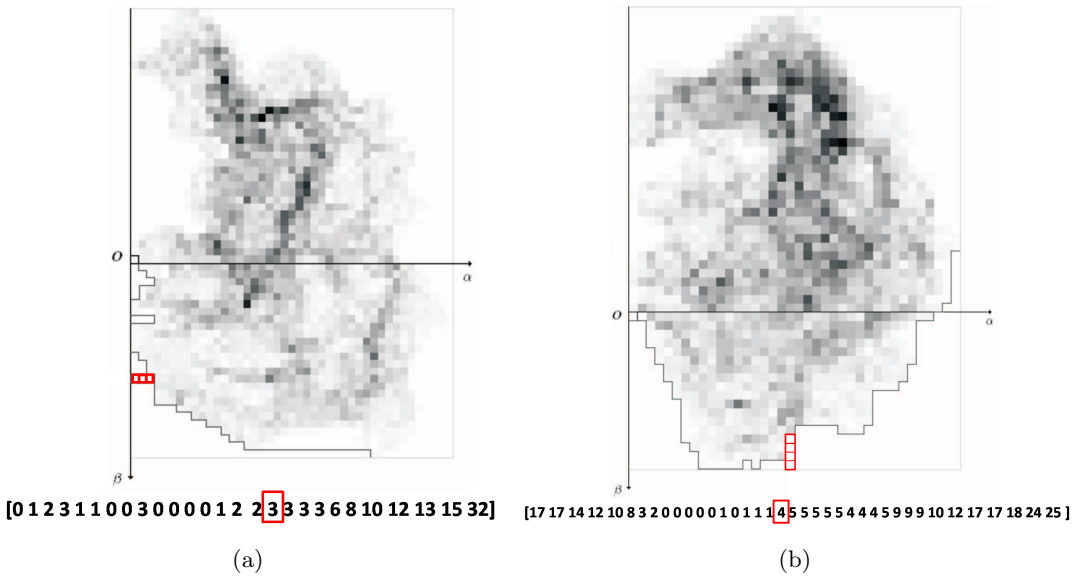
2.4 Matching shapes

2.4.1 Finding regions of similarity on two protein surfaces

This section presents a matching routine based on amended spin images to identify regions of structural similarity on two surfaces and to align them. The routine is based on the observation that surfaces with similar shape tend to have similar spin images, thus reducing a complex 3D matching problem into a 2D problem for which a simpler and more efficient solution exists.

Given two spin images each with $N = n \times m$ pixels, the similarity between them can be measured by the linear correlation coefficient R for the two sets of pixels. The non independence of the pixels on the same spin image does not appear to cause a serious problem for the matching when the filtering methods given below are used. A high value of R indicates similarity of the two spin images.

Since the amended spin images may have different sizes, the correlation value is computed on the two sub-images that overlap. More precisely, if the two amended spin images have size $n_1 \times m_1$ and $n_2 \times m_2$, then the correlation is determined on the two sub-images with dimensions $n = \min\{n_1, n_2\}$ and $m = \min\{m_1, m_2\}$ centered in the origins. Given two sets S and T consisting of s and t surface points on two proteins,



(a) (b)

Figure 2.2: a) horizontal profile of a blocked point; b) vertical profile of an unblocked point.

the first step is to establish correspondences among pairs of points on the two surfaces, based on the similarity of their spin images. The computational requirements of all correlation values can be quite high due to the typically large number of point pairs. Consequently, various filters are introduced to discard sets of point pairs which tend to generate low correlation.

First, only pairs of points with the same label (either blocked or unblocked) are considered. If h and k points are blocked on the two protein surfaces, respectively, the number of pairs is reduced from $O(s \times t)$ to $O\left(h \times k + (s-h) \times (t-k)\right)$, a significant improvement.

Second, only pairs of points with similar profiles are retained, where the similarity between spin image profiles is measured by the linear correlation coefficient R_p of two one-dimensional profile arrays. This computation is obviously less costly than that of the correlation of two spin images. Pairs of points with value R_p below a given minimum value are eliminated. Only for the remaining pairs the correlation coefficient R of the two spin images is computed.

This filtering operation reduces the number of pairs of points for which the correlation of the spin images is computed to about one third of the original pairs. The final set of point correspondences consists of all pairs of points with a correlation value R of their spin images above a given threshold (0.5). Such correspondences, ranked according to their R correlation values, are inserted into a linked list L.

Once individual point correspondences are established as described above, they are grouped in regions of consistent point correspondences on the two proteins. The grouping criterion (30) is the geometric consistency of distances of corresponding points and of angles formed by their normals. More specifically, a correspondence $C = (P, P')$ between two points P and P' on the two protein surfaces is defined to be geometrically consistent with a group of already established correspondences $C_1 = (Q_1, Q_1'), \ldots, C_i = (Q_i, Q_i'), \ldots, C_n = (Q_n, Q_n')$ if the following criteria are satisfied:

1. the spin images of P and P' are highly correlated;
2. for every $i = 1, \ldots, n$, the distances between P and Q_i and between P' and Q_i' are within some user-defined tolerance;
3. for every $i = 1, \ldots, n$, the angle between the normals at P and Q_i is the same as the angle between the normals at P' and Q_i' within some user-defined tolerance.

A greedy algorithm finds groups of geometrically consistent correspondences as follows. The top element of the list L, i.e. the correspondence with the highest correlation value, forms the seed of a group of correspondences. Then, after removing the top element, the algorithm scans the list L in decreasing order with respect to the correlation values; if a correspondence is found that is geometrically consistent with those already in the group, then it is added to group and removed from L. When no more correspondences can be added, but the list L is not empty, the process starts over again with the reduced correspondence list to create a new group. In this way, several groups of consistent corresponding points are generated, each identifying two similar surface regions, one on each protein.

The score of a solution is given by the number of corresponding pairs of points. Groups with less than a threshold number of elements (5 in this case) are discarded. The rigid transformation that best overlaps the two sets of corresponding points on the two regions is determined and the RMSD of corresponding points computed.

Additional information on two proteins can help reduce the amount of computation by selecting on the protein surfaces particular areas of interest and restricting the match to those areas. For instance, one can select only cavities on both proteins if the goal is to determine similar binding sites. This approach is explored in chapter 4.

In the general case, the following procedure is used to speed up the matching process. The surface points of the two proteins are mapped onto two 3D grids, where the cell dimension is equal to 6Å. The grids allow finding easily points that are close in 3D space. The matching procedure described above is applied to pairs of grid cells. If a good match is found for a given pair of cells then it is extended to points in adjacent cells.

A selection of only a subset of grid cells helps to reduce the computation of the procedure. In this case, only the cells which contain at least five points are retained. For any pair of selected cells, one on each protein, corresponding points in the two cells are identified.

Then, the point correspondences are grouped into sets of geometrically consistent correspondences, as described above. If the number of correspondences within a group is more than three, the minimum number to obtain a rototranslation, the group is extended by adding correspondences of points in adjacent cells using the same geometric

consistency criterion. The overall approach is sketched as follows:

MatchingProcedure

1. Map the surface points of the two proteins onto 3D grids, G and G'
2. Select subsets of cells $G_S \subseteq G$ and $G'_S \subseteq G'$
3. Generate the amended spin images for all points in the selected cells G_S and G'_S
4. For all pairs of selected cells $(g, g') \in G_S \times G'_S$ do
   (a) $L \leftarrow$ empty list
   (b) For all pairs of points Q, Q' in g and g' with the same label (either blocked or unblocked) compute the correlation R_p of their spin image profiles. If $R_p > 0.5$, then compute the correlation R of their spin images. If also $R > 0.5$, then add the pair Q, Q' to the list of correspondences L
   (c) Group the correspondences of L into sets of geometrically consistent correspondences
   (d) For each obtained group with more than a threshold number of correspondences, extend the group by adding consistent correspondences among points belonging to adjacent grid cells
   (e) Score each group by the number of pairs of corresponding points.

The procedure outputs the 30 top-ranked solutions. Some of the solutions may share several residues and consist mostly of correspondences that are geometrically consistent. The last step of the processing tries to merge such solutions using the same criteria of geometric consistency described above.

2.5 Data and results

This section reports on experiments conducted for the identification of regions of similarity on protein surfaces. A correspondence between points in the corresponding region on each of the proteins is obtained.

In the first experiment the method is benchmarked by considering the problem of comparing a pair of proteins or chains that bind the same or different ligands to check that a similar region on each protein can be found containing the binding site. Note

that MolLoc does not use any information about the existence or the location of the binding site on either protein or chain. A different problem which is computationally easier is to start with the known binding site on one protein or chain and search the other protein or chain for a similar site (58). Typically the binding regions lie in large cavities but no use is made of this information or fold information in the method.

In the second experiment, MolLoc was also checked by comparing proteins binding ligand NAD to see if it is able to detect a region on each chain that corresponded to the interface of the chain with its ligand.

Further examples in this paper of the application of the method involve proteins interacting with other proteins in an interface area that is relatively flat and much larger than the typical binding site of a ligand.

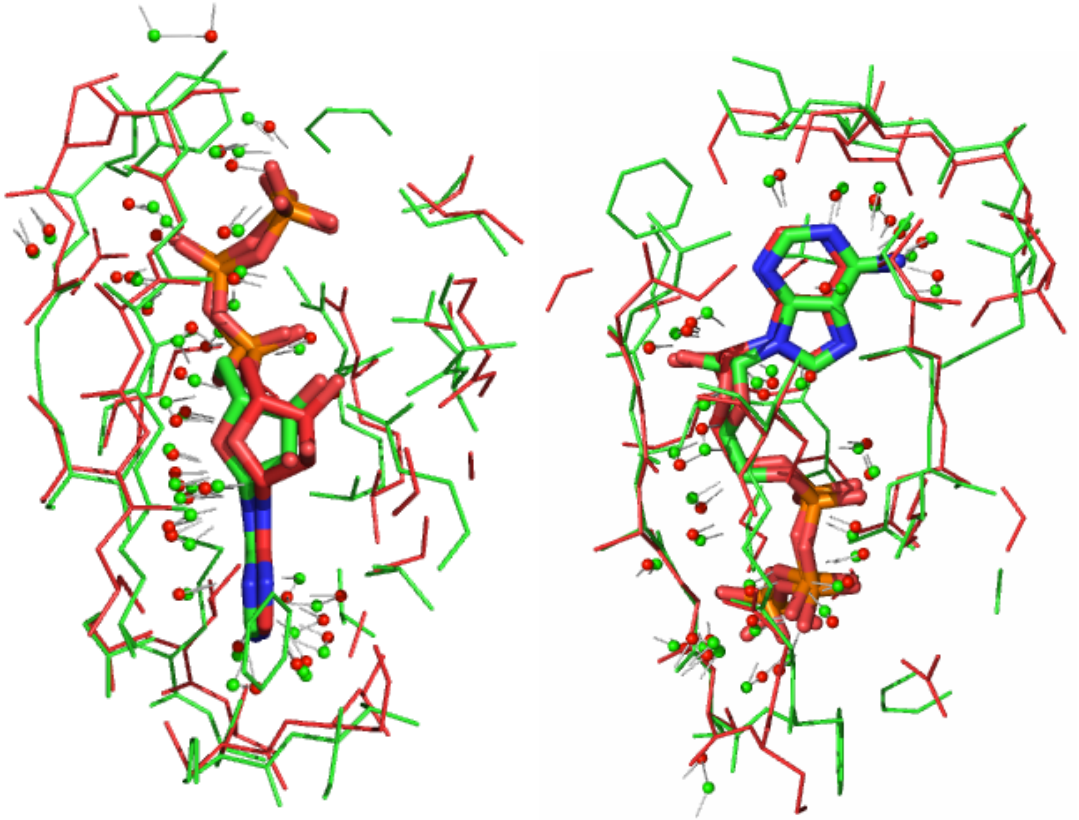
The dataset of the experiments includes proteins of the Cyclophilin-like fold from different species all interacting with cyclosporin.

The protein structures considered in this study are taken from the PDB (5). In some cases, only a chain from the protein is considered. For each chain or protein, the surface points and their normals are generated using Connolly's `msroll` program (18), after the removal of any ligand. Surface normals contribute to define the reference frames for the spin image construction. Amended spin images are created for surface points, as described in section 2.3.2.

2.5.1 Benchmark Protein-ligand complexes

MolLoc is benchmarked on different sets of proteins or chains that potentially bind to a ligand. The first is a subset of the representative set chosen in (58) which in turn included proteins used in the study by (36). This set includes 46 proteins, 12 proteins with a chain binding to ATP and 10 with a chain binding to other adenine-containing ligands. Other proteins are from diverse functional families that can bind estradiol, equilin and retinoic acid. Other different protein families from the set are: HIV-1, anhydrase, antibiotics, fatty acid-binding proteins, chorismate mutases and serine proteases. Table 2.1 lists all proteins chosen for this experiment.

Comparisons of a query protein or chain surface with the entire set of 46 proteins or chains were performed to retrieve those with high score when matched with the query. The score of a comparison is defined as the number of correspondences between points on the pair of matching regions identified on the two surfaces. the pair of regions for



(a) (b)

Figure 2.3: The superimpositions of the two most similar regions on proteins 1csn (in red) and 1atp (in green) aligns also their ligands. a) and b) are two views that show the alignment from different perspectives. The spheres are the pairs of correspondent Connolly points with their normals that belong to the top solution.

Protein family	Pdb ID
Adenine-binding	1ads 1byq 1b4v 1bx4 1byq 1kpf 1mmg 2src 1zin 9ldt
ATP binding proteins	1a82 1atp 1csn 1e2q 1f9a 1hck 1j7k 1jjv 1mjh 1nhk 1nsf 1phk
Serine proteases	1abi 4sgb 4tgl
Fatty acid binding proteins	1b56 1kqw 1lib 2cbr
Estradiol	1a27 1e6w 1fds 1lhu 1qkt 3ert
Anhydrase	1jd0
Retinoic acid-binding	1g5y 1gx9
Antibiotics	1alq 1bt5 1dcs
HIV-1	1mu2
Viral proteinase	1cqq 1mbm 1q2w
Chorismate mutase	1fnj

Table 2.1: The dataset for the first experiment

the two surfaces. For example, for the proteins 1atp and 1csn, which both bind to the ligand ATP, the two similar regions on each protein are part of the binding site. Figure 2.3 shows the two proteins aligned by the rigid tranformation derived from the correspondences of the solution.

The first set of results is obtained using the Catalytic Subunit of cAMP-dependent Protein-Kinase (pdb code 1atp, chain E) as query protein. This chain binds ATP. The ATP binding pockets in different proteins show great structural variability. Therefore one cannot expect the algorithm always to identify the common regions that correspond to the active sites on a pair of proteins. For instance, in the DNA ligase from bacteriophage T7 complex with ATP (1a0i) the ligand sticks out into the solvent, while in casein kinase-1 (1csn), a phosphate-directed protein kinase, the ligand ATP lies entirely at the bottom of a large cavity. Note that protein 1a0i is not included in the set of (58). The surface regions of the active sites in 1a0i and 1csn vary both in size and shape and the matching algorithm fails to identify the sites when they are compared to each other. This may be a case where the use of physico-chemical properties might have helped.

Table 2.2 lists the proteins or chains with the top 10 highest scores when compared with chain E of 1atp. For each protein, the table shows: its rank in the match; the PDB code and the chain id in case the ligand has contact; the protein name and fold; the number of corresponding pairs of surface points in the obtained solution region (based on which the proteins are ranked); the name of the ligand in the actual binding site; and finally, the rmsd of the rigid transformation RT that best aligns the two sets

of corresponding points. Table 2.3 shows the list of residues of 1atp in each of the 10 solutions when compared with the top highest scored proteins.

As can be seen from Table 2.2, most proteins that have a region on their surface resembling a region on 1atp (typically the binding site) have an adenine ring binding site. The results are compared with those of (58) where they search a database of complete protein structures by comparing them with the adenine binding site extracted from 1atp. Taking the top 10 proteins not including 1atp that were compared with 1atp from the set of proteins of Shulman-Peleg, eight of the proteins appear on both lists. The discrepancies are that the protein 1mu2, ranked number 5 in their list, is not even in the top 25. However, 1jd0, ranked number 9 in their list, has rank 20 in the procedure. These two proteins in their list, 1mu2 and 1jd0, do not bind ATP and do not have an adenine ring binding site. By contrast, the two proteins that appear in the top 10 and not in theirs are 1bx4 and 1f9a. The protein 1f9a has an ATP binding site and 1bx4 has a binding site for an adenine ring.

Rank	PDB:chain	Protein	Fold	# Corr.	Ligand	Rmsd
1	1phk	g-Subunit of glycogen phosphorylase kinase	Protein-kinase	190	ATP	1.1
2	1csn	Casein kinase-1, CK1	Protein-kinase	92	ATP	1.9
3	1mjh:B	"Hypothetical" protein MJ0577	Adenine nucleotide a hydrolase-like	56	ATP	0.7
4	1g5y:B	Retinoid-X receptor alpha	Nuclear receptor ligand-binding domain	55	REA	1.0
5	1bx4:A	Human Adenosine Kinase	Ribokinase-like	46	ADN	1.8
6	1b4v:A	Cholesterol Oxidase	FAD/NAD(P)-binding domain	46	FAD	1.8
7	2src	Tyrosine-protein Kinase SRC	Protein kinase-like (PK-like)	44	ANP	1.3
8	1hck	Cyclin-dependent PK	Protein-kinase	43	ATP	2.6
9	1nsf	Hexamerization domain of N-ethylmalemide-sensitive fusion protein	P-loop containing nucleoside triphosphate hydrolases	43	ATP	1.4
10	1f9a:A	"Hypothetical" Protein MJ0541	Adenine nucleotide alpha hydrolase-like	43	ATP	0.9

Table 2.2: High scoring pair-wise comparisons with 1atp:E.

A result of this experiment is to show that in many cases for the procedure the largest paired regions discovered with high similarity on two protein surfaces actually correspond to the area around the binding site. Tables 2.4-2.8, present more details on the results of pair-wise surface comparisons for proteins binding ATP and with high rank in Table 2.2. For each comparison the coverage of the solution is determined with respect to the actual binding sites on the two proteins. The binding sites of all proteins were derived with the CSU software that analyzes the interatomic contacts in protein complexes (59). In the Tables 2.4-2.8, for each protein of the matched pair, "#residues in solution" and "# residues in binding site" gives the number of residues in the solution region and in the binding site, respectively. Note that an atom belongs

Protein	List of residues of 1atp in the solution
1phk	49 50 51 52 57 70 71 72 104 120 121 123 127 165 168 170 171 173 183 184 187 200 201 204 205 209 219 223
1csn	50 51 52 55 57 70 72 88 91 104 118 120 121 166 167 168 170 171 173 183 184 185 186
1mjh:B	49 57 121 122 173 327
1g5y:B	49 50 51 57 70 104 120 123 173 183
1bx4:A	49 50 51 53 55 57 168 170 171 173 176 183
1b4v:A	49 50 51 55 57 120 127 168 170 173 183 184
2src	49 50 51 57 72 168 183 184 185 326
1hck	55 57 70 120 121 123 170 171 173 183
1nsf	49 57 70 104 120 123 170 171 173 183
1f9a	49 50 57 70 104 173 183

Table 2.3: List of residues of protein 1atp:E in the solutions of the pair-wise comparisons of Table 2.2. The underlined residues are in contact with a ligand in the PDB 1atp:E according to CSU software.

to the solution if at least one of the Connolly's surface points close to it belongs to the solution. A residue with at least one atom in the solution is considered to be in the solution. *Cov* is defined as the percentage of residues in the binding site of the protein that is found in the solution. Column 4 of Tables 2.4-2.8 shows the coverage *Cov* of the binding site. Also the coverage of the part of the binding site in contact with the adenine ring of ATP is considered, called *Cov Adenine Ring*, which is defined as the percentage of residues in contact with the adenine ring (according to CSU) that is found in the solution. These coverage values, shown in Tables 2.4-2.8 column 5, appear to be higher than the coverage values of the entire binding site for almost all comparisons. As shown in Tables 2.4-2.8, the matching procedure identifies the most similar regions in all 4 protein pairs to correspond to the ATP binding sites with a relatively good coverage of the binding site.

Note that in spite of the fact that proteins 1atp and 1csn have high structural similarity overall it is still the region about the binding site that is found to be most similar on the surface. A structural alignment algorithm such as PROuST (17) or CE (57) aligns the two overall structures (not just the surfaces) fairly well. Out of 336 residues of 1atp and 293 residues of 1csn, the method finds 248 residues superimposed with rmsd less than 2.5. On the other hand, the two proteins appear by visual inspection not identical in almost all areas on the surface except in the binding sites. These binding sites are clearly recognized as the most similar areas by this strategy, as can be seen in Table 2.9. The table lists a subset of pairs of corresponding atoms of 1atp and 1csn in the solution and shows that such atoms are in contact with the same atoms of ligand

ATP in the two complexes. For the remaining pairs of corresponding atoms of the solution (not listed in the table), the contact is with nearby atoms of the ligand.

For a few pairs of proteins binding ATP, the solutions do not correspond to the binding sites. For instance, for the pair 1atp and 1e2q the solution consists of 36 corresponding points that are outside the binding site. Protein 1e2q is a thymidylate kinase complexed with ATP, TMP, and a magnesium ion. The ligand ATP is located at the bottom of a cavity of 1atp, but appears more exposed in 1e2q flat on a surface rather than in a cavity. As a consequence, while in protein 1atp almost all surface points in contact with ATP are labeled blocked by this procedure, in 1e2q they are labeled unblocked and therefore no correspondence between such points is found by this matching procedure.

Pdb ID	# residues in solution	# residues in binding site	Cov	Cov Adenine Ring
1atp	28	23	78%	82%
1phk	27	26	69%	100%

Table 2.4: Comparison of 1atp (cAMP-dependent Protein-Kinase) with 1phk (Subunit of glycogen phosphorylase kinase). *Cov* is the percentage of residues in the binding site of the protein that is found in thie solution, while *Cov Adenine Ring* is the percentage of residues in contact with the adenine ring that is found in the solution.

Pdb ID	# residues in solution	# residues in binding site	Cov	Cov Adenine Ring
1atp	23	23	70%	64%
1csn	22	26	62%	50%

Table 2.5: Comparison of 1atp (cAMP-dependent Protein-Kinase) with 1csn (Casein kinase-1). *Cov* is the percentage of residues in the binding site of the protein that is found in the solution, while *Cov Adenine Ring* is the percentage of residues in contact with the adenine ring that is found in the solution.

Pdb ID	# residues in solution	# residues in binding site	Cov	Cov Adenine Ring
1atp	6	23	26%	55%
1mjh	6	25	24%	25%

Table 2.6: Comparison of 1atp (cAMP-dependent Protein-Kinase) with 1mjh:B ("Hypothetical" protein MJ0577). *Cov* is the percentage of residues in the binding site of the protein that is found in the solution, while *Cov Adenine Ring* is the percentage of residues in contact with the adenine ring that is found in the solution.

Pdb ID	# residues in solution	# residues in binding site	Cov	Cov Adenine Ring
1atp	10	23	39%	64%
1hck	10	24	42%	58%

Table 2.7: Comparison of 1atp (cAMP-dependent Protein-Kinase) with 1hck (Cyclin dependent PK). *Cov* is the percentage of residues in the binding site of the protein that is found in the solution, while *Cov Adenine Ring* is the percentage of residues in contact with the adenine ring that is found in the solution.

Pdb ID	# residues in solution	# residues in binding site	Cov	Cov Adenine Ring
1atp	10	23	43%	73%
1nsf	10	23	35%	75%

Table 2.8: Comparison of 1atp (cAMP-dependent Protein-Kinase) with 1nsf (Examerization domain of N-ethilmalemide-sensitive fusion protein). *Cov* is the percentage of residues in the binding site of the protein that is found in the solution, while *Cov Adenine Ring* is the percentage of residues in contact with the adenine ring that is found in the solution.

	1atp			1csn		
Residue	Properties	Atom	Residue	Properties	Atom	Corresponding atom on ATP
THR 51	neutral , polar	N	GLY 19	neutral , non-polar	CA	C4*
THR 51	neutral , polar	O	GLU 20	neutral , polar	O	C5*
GLY 52	neutral , non-polar	CA	GLY 21	neutral , non-polar	CA	O3B/O3A
GLY 55	neutral , non-polar	O	GLU 20	neutral , polar	O	C5*
VAL 57	neutral , non-polar	CB	ILE 26	neutral , non-polar	CB	04*/C1*
VAL 57	neutral , non-polar	CG1	ILE 26	neutral , non-polar	CG2	N9/C5
VAL 57	neutral , non-polar	CG2	ILE 26	neutral , non-polar	CG1	O4*/C8
ALA 70	neutral , non-polar	CB	ALA 39	neutral , non-polar	CB	C6/N6/N1
LYS 72	basic , polar	CE	LYS 41	basic , polar	CE	O3A
VAL 104	neutral , non-polar	CG1	LEU 88	neutral , non-polar	CD1	N6
MET 120	neutral , non-polar	SD	ALA 39	neutral , non-polar	CB	N6
MET 120	neutral , non-polar	SD	ASP 86	acidic , polar	O	N6
GLU 121	neutral , polar	O	LEU 88	neutral , non-polar	N	N6
GLU 170	neutral , polar	O	ASP 135	acidic , polar	O	C3*/O3*/C2*
GLU 170	neutral , polar	CB	ASP 135	acidic , polar	O	O3*
GLU 170	neutral , polar	CB	ASP 135	acidic , polar	CB	O3*
ASN 171	neutral , polar	OD1	ASP 135	acidic , polar	O	O3*
LEU 173	neutral , non-polar	CD1	LEU 138	neutral , non-polar	CD1	C2/C4/C6
LEU 173	neutral , non-polar	CD2	LEU 138	neutral , non-polar	CD1	C2*

Table 2.9: A subset of correspondences in the solution between protein 1atp and 1csn. Each row represents a correspondence of an atom of 1atp and an atom of 1csn. For each correspondence, the atoms of the ligand ATP listed in the last column are in contact with both atoms of 1atp and of 1csn in the same row.

The second dataset includes the following 11 proteins that bind NAD (Nicotinamide-adenine-dinucleotide): 1a27, 1ads, 1c1d, 1dqs, 1ee2, 1ew6, 1gzf, 1ici, 1k4m, 2bkj, 9ldt. In a recent study by (48) for a given set of proteins some of which bind ATP, NAD, heme, etc., it was observed that the sites in contact with NAD tend to cluster well and certainly better than those binding ATP. This fact is also found in this study. All-to-all pair-wise comparisons of the proteins of the above set of 11 was performed. For almost all comparisons, the common area on the two protein surfaces included residues of the binding sites.

The comparison of a single protein chain 1dqs:A with the remaining 10 proteins of the set reveals another interesting fact. 1dqs:A is a multi-domain protein chain with Dehydroquinate synthase-like fold binding two ligands, NAD and CRB, and two metal ions, Zn and Cl. Table 2.10 shows the list of residues of 1dqs in each of the 10 solutions. In all pair-wise comparisons the area on the surface of 1dqs most similar to an area on the second protein contains residues of the binding sites. Although no residue is present in all such lists, i.e. the intersection of such lists is empty, some residues appear very frequently. HIS 271, HIS 287, GLU 194 are present in 9, 7 and 6 solutions, respectively, out of 10 possible solutions. It is interesting to note that these three residues are in contact with more than one ligand in the complex 1dqs. More precisely, HIS 271 is in contact with ZN and CRB, residue 194 is in contact with ZN and NAD and residue 287 is in contact with all three ligands, CRB, NAD, ZN.

This fact was also observed for other proteins. For instance, in all pair-wise comparisons of 1k4m with the remaining 10 proteins of the set, the residues 134, 19, 16 that appear most frequently in the 10 solutions are in contact with both ligands NAD and CIT of 1k4m.

2.5.2 Proteins interacting with other proteins

Another test included 6 proteins of the Cyclophilin-like fold from different species all interacting with cyclosporin. The set includes:1cyn, 1bck, 1m63, 1mf8, 1qng, and 2rmc. All pair-wise comparisons returned large regions of similarity on the surfaces corresponding to the actual interface areas with cyclosporin. For instance, for the pair 1cyn and 1bck the solution consists of approximately 600 corresponding points with rmsd=0.5 and with good coverage of the interface area (see table 2.11). Here *Cov* is the percentage of residues in the interface site of the proteins that is found in the

solution. The proteins 1cyn and 1bck have a good structural superposition according to PROuST (17) and CE (57) (164 alignment length, 63% sequence identity, rmsd 0.9).

Protein	List of residues of 1dqs:A in the solution
1a27	79 84 115 116 140 142 146 183 187 190 194 271 286 287
1ads	163 268 272 275 276 351 352 354 355 356 357
1ici	84 119 143 153 154 166 194 267 268 271 287
1k4m	119 194 267 268 271 287 355 356 357
2bkj	152 161 162 264 267 268 271 356 357
1gzf	119 146 147 162 194 268 271 287 357
9ldt	146 152 154 162 264 267 268 271 357
1ee2	84 116 117 119 146 194 271 287
1c1d	84 115 116 119 194 267 271 287
1e6w	84 115 119 142 161 197 271 287

Table 2.10: List of residues of protein 1dqs:A in the solutions of all pair-wise comparisons. The underlined residues are in contact with a ligand in the PDB 1dqs:A according to CSU software. Note that only 1ads has a tim-barrel fold.

Pdb ID	# residues in solution	# residues in binding site	Cov
1cyn	26	20	75%
1bck	25	17	59%

Table 2.11: Comparison of 1cyn with 1bck. *Cov* is the percentage of residues in the interface site of the proteins that is found in the solution.

2.5.3 Running times

The program is written in C++ and uses the LEDA library (47) for the handling of the data structures and standard matrix operation. The execution time for the matching of two complete surfaces ranges from 20 minutes for small molecules up to 2 hours for the largest proteins. This execution time includes also the generation of the spin images. Most of the execution time is spent in the determination of the correlation of spin images to identify the points on the two surfaces with most similar spin images.

2.6 Conclusions

MolLoc is a method to find regions of similarity on two protein surfaces that produces good results when applied to known families of proteins. The method is based on a new geometric protein surface descriptor, the spin image profile, that is crucial for obtaining reasonable execution times for the matching procedure. These facts qualify spin images as a powerful tool in a variety of applications, from the analysis of protein structure, to protein structural alignment.

3

Cavity Detection and Matching for Binding Site Recognition

This chapter describes a suite of methods for the problem of protein binding site recognition, based on a representation of the protein structures by a collection of spin-images. A procedure for cavity detection is coupled with the method previously described(MolLoc) for the recognition of similar regions in two proteins, and applied to the comparison of two protein's cavities, the all-to-all pairwise comparison of a set of cavities, and the recognition of multiple binding sites in one cavity. All the presented methods can be used to screen large collections of proteins.

The detection of the cavities in a given protein is often the preliminary step in protein binding site recognition, since binding sites usually lie in cavities. The comparison of two cavities identifies two similar regions in the two cavities, and hints at a common functional structure when one or both regions include a binding site. The all-to-all pairwise comparison of a set of cavities is clustered according to the measure of similarity of the cavities, obtaining a clustering that groups together cavities with the same binding sites, when their structures are similar enough. The recognition of multiple binding sites in one cavity is performed by the comparison of a cavity, called *background cavity*, with a dataset of cavities, and clustering its residues that match the residues of other cavities in the data set. The four methods are benchmarked on different databases, and their effectiveness is discussed.

3.1 Introduction

Cavity detection is often the first step for functional analysis, since binding sites in proteins usually lie in cavities. Typically the surface region that constitutes the binding site of a ligand in a cavity is only a small part of the total surface area of the cavity and the volume of the cavity is much larger than needed to accommodate the ligand. Moreover, the binding site by definition surrounds that part of the ligand that interacts with the protein, and thus similar conformations and orientations of a ligand in two cavities correspond to two geometrically similar binding sites in those cavities. Therefore, MolLoc is adapted to the problem of finding similar regions in two cavities. Then, given a dataset of proteins, all-to-all-pairwise comparison of their cavities is used to cluster the proteins based on the structurally similar regions in their cavities. This often corresponds to clusters of proteins cavities that interact with the same ligand. Finally, one cavity called *background cavity* is compared to a dataset of protein cavities, and it's show that is possible to identify multiple binding sites that lie in the background cavity, by using a novel distance measure that clusters its residues that match those in the cavities of the dataset.

The cavity detection procedure is tested with the nonredundant set of 244 protein structures used in (24), and the results obtained on the dataset with this procedure that uses only geometric criteria are comparable to the SURFNET-ConSurf method, which adds information on the conserved residues from the ConSurf-HSSP database (25) to the surface pocket predictor SURFNET by (37). Then, the combined use of this cavity detection and cavity comparison procedures for the comparison of two cavities was benchmarked on five pairs of proteins, containing distant homologues, used in (9). The combined approach achieves better results in identifying the binding site, while it improves on the execution times reported in the protein surface comparison method alone, from 1-2 hours down to few minutes or even seconds. The dataset for the all-to-all pairwise cavity comparison has been introduced in (48); 40 proteins divided in four groups of ten proteins, where each group contains proteins that bind the same ligand (ATP, NAD, heme and steroid). The proteins are then clustered according to the results obtained in the all-to-all pairwise comparisons, and the more conserved the conformation of the ligand, the more precisely the proteins are clustered for the ligand that they bind. In one case (pdb:1jtv) a protein hosting two binding sites in the same

cavity belonged to the cluster corresponding to its biggest binding site. Thus, using a background cavity comparison, also the second binding site was correctly identified.

The chapter is organized as follows. Section 3.2 presents a short survey of the existing methods for cavity detection and binding site recognition. Section 3.3 discusses how the labeling of the protein surface points is used in the identification of protein cavities. Section 3.4 describes the methods for cavity detection and matching. Experimental results are provided in section 3.5 and conclusions in section **??**.

3.2 Previous work

3.2.1 Cavity Detection

Several methods and procedures exist to detect protein cavities, either internal to a molecule or external on a protein surface (12; 24; 28; 35; 37; 42; 43; 44; 65). The cavity detection algorithms are often based on fitting probe spheres into the spaces between the atoms.

In DOCK(35) algorithm, for each pair i, j of surface points, a sphere is generated tangent to the surface at i and j and with center on the surface normal at i. Then the cluster program of the DOCK suite performs a clustering of the obtained spheres. Finally, geometric values of the resulting clusters, such as volume and depth, are determined. In many cases, the largest cluster is the ligand binding site of the molecule.

The program SURFNET by (37) for visualizing molecular surfaces builds a sphere for each pair of nearby atoms with the center halfway between the two atoms and then adjusts the radius if it clashes with any neighboring atom. The predicted cleft volume is in many cases much larger than the ligand that occupies it. A trimming procedure called SURFNET-ConSurf reduces the size of the clefts generated by SURFNET by cutting away regions distant from highly conserved residues (24). In the POCKET program (42) trial spheres are placed on a regular three-dimensional grid and their radii are reduced in size until no neighboring atom penetrates the sphere. For a review of cavity detection methods, refer to (40).

3.2.2 Binding Site Detection

Much work has been done on the recognition of the binding sites of proteins (3; 6; 7; 13; 31; 32; 33; 36; 45; 48; 51; 55; 58; 60; 64; 67) using various approaches based on

different protein representations and matching strategies.

Three recognition problems are generally addressed: 1) the comparison of known binding sites to determine their degree of similarity, 2) the search for a given binding site in a set of complete protein structures, and 3) the search for putative binding sites of a given protein in a set of known binding sites. In SiteEngine (58) all three problems are considered and extensive experimentation is conducted for each. Recognition is obtained by hashing triangles of points and their associated physico-chemical properties and by application of a clever scoring mechanism. A method for binding pocket comparison and clustering has been proposed (48) based on a protein shape representation in terms of spherical harmonic coefficients. This method is interesting and fast; however, as pointed out by the authors, it requires a registration phase, to align the two shapes, that it is not always very reliable. A geometric hashing approach have been used(13) to compare and cluster phosphate binding sites in proteinnucleotide complexes, leading to the identification of 10 clusters. These are the structural P-loop, di-nucleotide binding motif [1] and FAD binding motif. A cavity-aware match technique(15) which uses C-spheres to represent active clefts which must remain vacant for ligand binding. The technique reduces the number of false positives while maintaining most of the true positive matches found with identical motifs lacking C-spheres.

3.3 Characterizing cavities in terms of blocked points

Surface points are labeled as *blocked* or *unblocked* depending on the shape of their spin-images. As seen in chapter 2, a surface point P with normal n is labeled blocked if n intersects the surface at any other point lying above the tangent plane T at P perpendicular to n; otherwise it is labeled unblocked. To label a point, only the first column of its spin-image needs to be examined: if it contains a non-zero pixel with positive β, then the point is blocked, otherwise it is unblocked.

Crucial to the cavity detection procedure is the identification of blocked points on the protein surface. Typically, the number of blocked points on a protein surface is smaller than that of unblocked points, i.e. of points whose normal does not intersect the surface at any other point. Not surprisingly, the opposite is true for points of the binding sites.

[1] FAD/ NAD(P)-binding and Rossman-like fold

Fig.3.2(a) shows the statistics of blocked points of proteins and binding sites (the proteins are taken from the nonrendundant dataset of (24), that will be discussed in more detail later). For most proteins, less than half of the surface points are blocked, while for the majority of the binding sites, more than 70% of points are blocked.

For example, out of 5039 Connolly's points of the D2 Hexamerization domain of N-Ethylmaleimide sensitive factor[1], just the 35% are blocked. For the binding site of 1nsf with ligand ATP, the percentage of blocked points goes up to 74%. As another example, protein 1mjh, an hypothetical protein binding ATP, has an even higher percentage of blocked points on the binding site, i.e. above 80%.

Furthermore, blocked points are strongly present in cavities, especially in internal cavities. In fact, if a cavity is internal, then the normals at all points of the cavity intersect the protein at some other points of the cavity. If a cavity is external, there might be few unblocked points at the bottom of the cavity.

The identification of blocked points can be done very easily once the spin-images of surface points have been constructed. If the first column (corresponding to $0 \leq \alpha < \varepsilon$) of a spin-image contains a non-zero pixel with positive β, then the point is blocked, otherwise is unblocked. This works under the assumption that the normal n intersects the surface at some other point Q if n is within ε distance from Q, where ε is the spin-image pixel size (1Å).

3.4 Methods

3.4.1 Cavity detection

Surface cavities are detected using blocked points. More precisely, for each blocked point, the cavity detection method builds the largest sphere that can fit at that point; then it determines the cavities as clusters of overlapping spheres. Given a blocked point p with normal n and spin-image s, the associated sphere $\mathbb{S}(s(p))$ is obtained from the biggest semi-circle in s, tangent to the cell in the origin, with center on the normal n and radius s.t. the sphere contains only empty pixels (see Fig.3.1). Due to the cylindrical symmetry of spin-images, the semi-circle of s corresponds to the sphere in 3-D. Defining the sphere starting from the spin-image allows fast construction of the spheres.

[1] Pdb id: 1nsf

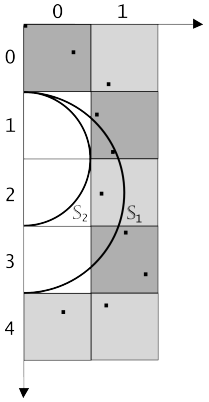


Figure 3.1: The cavity detection procedure starts determining sphere $\mathbb{S}_1$ with radius R; then, it scans row 1 and determines a stricter constraint on the sphere radius, obtaining $\mathbb{S}_2$. Rows 2 and 3 don't impose new constraints on the sphere radius, and thus $\mathbb{S}_2$ is the final sphere.

Algorithm 1 contains the pseudocode for the procedure to build a sphere $\mathbb{S}(s(p))$ with radius $R = R(\mathbb{S}(s(p)))$ and center $C = C(\mathbb{S}(s(p)))$ for a blocked point p with spin s and horizontal profile $\underline{h}(s(p))$, where $\underline{h}(s(p))(i)$ indicates the i^{th} element of the profile. Fig.3.1 gives a visual explaination of the procedure. The algorithm is linear in Z, and the time required to generate all spheres is $\mathrm{O}(b \times d)$, where b is the number of considered blocked points, typically much smaller than the number m of all surface points, and d is the maximum Z value of the spin-images.

Blocked points with large Z values are not typical of cavities, since they can also be found at the top of a region if their normal intersects the surface at a far away region. Thus, for a molecule with a set B of blocked points, spheres are generated only for the subset B' of points of B with a Z value below a given threshold Z_{max}, and the time complexity becomes $\mathrm{O}(b \times Z_{max})$.

Once all spheres of blocked points are obtained, those with R below a certain threshold R_{min} are removed so that small gaps between atoms are not considered. From the remaining spheres, a clustering procedure determines collections of interpenetrating spheres corresponding to the points of the surface cavities. The clusters are identified as the connected components of the undirected graph $G = (V, E)$, in which the vertices are the blocked points, and an edge connects two vertices if their spheres overlap. The cavity detection procedure is described in algorithm 2.

Algorithm 1 BUILD SPHERE($s(p)$)

Input: spin-image s of a point p

Output: center C and radius R of the sphere $\mathbb{S}(s(p))$

1: $R \leftarrow |\underline{h}(s(p))|/2$
2: **for** $j = 1, \ldots, |\underline{h}(s(p))|$ **do**
3: $\quad i \leftarrow \underline{h}(s(p))(j)$
4: $\quad$ **if** $i \geq j$ **then**
5: $\quad\quad R \leftarrow \min\{R, (i^2 + j^2)/2j\}$
6: $\quad$ **else**
7: $\quad\quad R \leftarrow \min\{R, (i^2 + (j-1)^2)/2(j-1)\}$
8: $\quad$ **end if**
9: $\quad C \leftarrow (0, (R+1)\varepsilon)$
10: **end for**
11: **return** C, R

Taking into account the pre-processing phase needed to create m spin-images, the overall time complexity of the cavity detection procedure becomes $\mathrm{O}(m \times \max\{m, D\} + b \times d)$, where D is the size of the spin-image. This represents a computational advantage with respect to methods for cavity detection that generate m^2 trial spheres, one for each pair of surface points, and check the non penetration of other surface points into each sphere, obtaining an overall time complexity of $\mathrm{O}(m^3)$.

3.4.2 Cavity Matching

MolLoc takes as input a pair of proteins and finds the regions on the two surfaces that most resemble each other. The method is adapted for comparing pairs of cavities, and the resulting procedure is described in Algorithm 3.

On line 3, the formula of the statistical correlation is

$$R(P,Q) = \frac{N \sum p_{ij} q_{ij} - \sum p_{ij} \sum q_{ij}}{\sqrt{(N \sum p_{ij}^2 - (\sum p_{ij})^2)(N \sum q_{ij}^2 - (\sum q_{ij})^2)}}, \tag{3.1}$$

where p_{ij} e q_{ij} are the common cells i, j of the spin-images P and Q, and N is the number of elements evaluated.

The grouping of the correspondences on line 8 is based on a greedy algorithm that proceeds as follows. The correspondence with the highest correlation value, i.e. the

Algorithm 2 CAVITY DETECTION($\mathcal{S}$)

Input: spin-images surface representation $\mathcal{S}$ of protein P

Output: set of cavities $\mathcal{C}$ of P

1: list of cavities $\mathcal{C} \leftarrow \emptyset$
2: determine the set of blocked points $B \subseteq \mathcal{S}$
3: **for all** the points $b \in B$ **do**
4: compute $\underline{h}(s(b))$
5: **end for**
6: determine $B' = \{b \in B : |\underline{h}(s(b))| \leq Z_{max}\}$
7: **for all** the points $b \in B'$ **do**
8: compute BUILD SPHERE($s(b)$)
9: **end for**
10: determine $B'' = \{b \in B' : R(\mathbb{S}(s(b))) \geq R_{min}\}$
11: build the undirected graph $G = (V, E)$, where where a node $v \in V$ corresponds to a blocked point $b \in B$, and $e = (v_i, v_j) \in E \Leftrightarrow dist(C_i, C_j) < R_i + R_j$, where $C_i \doteq C(\mathbb{S}(s(b_i)))$, $R_i \doteq R(\mathbb{S}(s(b_i)))$
12: find the connected components $G_1, \cdots, G_n$ of G using Breadth First Search
13: **for all** $G_i \in G$ **do**
14: define the cavity c_i as the set of residues of P with at least one point $b \in G_i$
15: $\mathcal{C} \leftarrow \mathcal{C} \cup \{c_i\}$
16: **end for**
17: **return** $\mathcal{C}$

top element of the list L, forms the seed of a group of correspondences. Then, after removing the top element, the algorithm scans the list L in decreasing order with respect to the correlation values; if a correspondence is found that is geometrically consistent with those already in the group, then it is added to the group and removed from L. The consistency criterion states that the angles between normals at two surface points on one protein and the distances between the two points must be preserved, within fixed thresholds (28° and 3 Å), between the corresponding points of the other protein. When no more consistent correspondences are found, but the list L is not empty, the process starts over again with the reduced correspondence list to create a new group.

This procedure considers the thirty top-ranked solutions. Some of the solutions may share several residues and consist mostly of correspondences that are geometrically

Algorithm 3 PAIRWISE COMPARISON($\mathcal{C}_1, \mathcal{C}_2$)

Input: list of cavities $\mathcal{C}_1$ and $\mathcal{C}_2$ of proteins P_1 and P_2

Output: list of points correspondences L that identifies the most extended similar regions on $\mathcal{C}_1$ and $\mathcal{C}_2$

1: list of points correspondences $L_{start} \leftarrow \emptyset$
2: **for all** the pairs of points (p_1, p_2) such that $p_1 \in \mathcal{C}_1, p_2 \in \mathcal{C}_2$ with the same label (either blocked or unblocked) **do**
3: compute the statistical correlation r of their spin-images
4: **if** $r \geq 0.5$ **then**
5: $L_{start} \leftarrow L_{start} \cup \{(p_1, p_2)\}$
6: **end if**
7: **end for**
8: group the correspondences of L_{start} into lists of geometrically consistent correspondences $L_1, \ldots, L_m$
9: score each list by the number of pairs of corresponding points
10: merge the top 30 lists into a list L that mantains only geometrically consistent correspondences
11: **return** L

consistent. The last step of algorithm 3 merges such solutions using the same criteria of geometric consistency described above.

The aim of the procedure is to find similar binding sites in two proteins, when these binding sites share a structurally similar region. In this case, comparing just the cavities of the two proteins is often the best option, because the cavities of a protein usually host the protein binding site, and because it's faster to compare just the cavities of two proteins than to compare the whole protein surfaces. In fact, MolLoc takes between one and two hours to compare two complete proteins, while the pairwise comparison described in algorithm 3 takes few minutes to compare the cavities of two proteins.

3.4.3 All-To-All Pairwise Cavity Comparison

The methods described so far can divide a dataset of proteins into groups of proteins with similar structures using a complete linkage clustering of all the proteins, that depends on the results of the pairwise protein comparisons. The clustering distance is $d = 1/(1 + c)$, where c is the number of correspondences between two proteins.

Algorithm 4 decribes the method.

Algorithm 4 ALL-TO-ALL PAIRWISE CAVITY COMPARISON($\mathcal{P}$)

Input: list of proteins $\mathcal{P} = P_1, \ldots, P_n$

Output: clustering of proteins according to their pairwise similarity

1: build the spin-image surface representations $\mathcal{S}_1, \ldots, \mathcal{S}_n$ for $P_1, \ldots, P_n$
2: **for all** the proteins $P_i \in \mathcal{P}$ **do**
3: $\quad \mathcal{C}_i \leftarrow$ CAVITY DETECTION($\mathcal{S}_i$)
4: **end for**
5: **for all** the sets of cavities $\mathcal{C}_i$ **do**
6: $\quad$ keep only the four biggest cavities
7: **end for**
8: **for all** the pairs of proteins $(P_i, P_j) \in \mathcal{P}, i \neq j$ **do**
9: $\quad L_{ij} \leftarrow$ PAIRWISE COMPARISON($\mathcal{C}_i, \mathcal{C}_j$)
10: **end for**
11: define the distance between two proteins P_i and P_j as $d \doteq 1/(1+c)$, where c is the number of correspondences in L_{ij}
12: **return** complete linkage clustering of $P_1, \ldots, P_n$ with distance d

The comparison between the cavities of two proteins, using algorithm 3, returns the most extended similar regions between the two proteins, considering just their cavities. One can choose which cavities to consider in the comparison, depending on the study he wants to perform. The four biggest cavities for each protein are considered(line 6), since section 3.5.1 shows that the binding site lies in one of these cavities in most of the cases.

3.4.4 Background Cavity

Using the all-to-all pairwise comparison gives us a functional classification of each protein cavity as a whole, but, when more than one binding site is present in the cavity, the bigger binding site shadows the smallest one, and the cavity is assigned to just one cluster. To overcome this, consider the comparisons of a protein cavity, called the *background cavity*, with all the others of the dataset. Each comparison identifies a set of correspondences, i.e. a set of residues in the background cavity and another set of residues in the other cavity. A complete linkage clustering of all the other proteins in the background cavity is produced, this time with distance $d = 1/(1+c)$, where c is the

number of common residues in the background cavity that two comparisons identify. The procedure is outlined in algorithm 5.

Algorithm 5 BACKGROUND CAVITY COMPARISON($c_B, \mathcal{P}$)

Input: background cavity c_B of protein P and the list of proteins $\mathcal{P} = P_1, \ldots, P_n$
Output: clustering of regions $R_1, \ldots, R_n$ in c_B similar to regions on $P_1, \ldots, P_n$

1: build the spin-image surface representations $\mathcal{S}_1, \ldots, \mathcal{S}_n$ for $P_1, \ldots, P_n$
2: **for all** the proteins $P_i \in \mathcal{P}$ **do**
3: $\quad \mathcal{C}_i \leftarrow$ CAVITY DETECTION($\mathcal{S}_i$)
4: **end for**
5: **for all** the sets of cavities $\mathcal{C}_i$ **do**
6: $\quad$ keep only the four biggest cavities
7: **end for**
8: **for all** the proteins $P_i \in \mathcal{P}$ **do**
9: $\quad L_i \leftarrow$ PAIRWISE COMPARISON($c_B, \mathcal{C}_i$)
10: $\quad$ define the region R_i as the set of residues on c_B that partecipate in at least one correspondence in L_i
11: **end for**
12: define the distance between two regions R_i and R_j as $d \doteq 1/(1+c)$, where c is the number of common residues
13: **return** complete linkage clustering of $R_1, \ldots, R_n$ with distance d

3.5 Data and results

3.5.1 Cavity Detection

Experiments for cavity detection have been performed on the data set of 244 nonredundant proteins used in (24). The protein structures are taken from the PDB (5), and for each structure only the chain (or chains) and ligand that represent the functional unit of the protein are retained. Of these proteins, 112 are enzymes (45.9%), 129 nonenzymes (52.9%), and three "hypothetical" (1.2%) proteins, according to PDBsum (38) and Uniprot (1).

These PDB entries contained 464 ligands not covalently bound to the protein and then for each complex protein-ligand there is a binding site. The binding sites of these complexes are determined in the following way. For a ligand binding to a protein, the

binding site consists of the atoms of the protein that are (i) closer than a given threshold (5 Å in these experiments) to at least one atom of the ligand, and (ii) have at least one surface point that is *blocked by the ligand*. A protein surface point P with normal n is said to be blocked by the ligand if there is at least one ligand surface point whose distance from n is less than ε. The surface points and their normals are generated using Connolly's algorithm (18). The obtained binding sites are generally very close to the binding sites derived with the CSU software that analyzes the interatomic contacts in protein complexes (59).

The ligands in the data set form a very heterogeneous set, that includes sugars, co-factors, substrate analogs, peptides, etc. They also show great variability in the size of their binding sites, varying from 3 atoms for NAG-21 in 1o7d, to 141 atoms for CDN in 1nek.

Although there is a correlation between the number of atoms of the binding sites and of the ligands, the binding sites of the same ligand with different proteins may vary significantly in size. For example, the binding sites of ligand MPD in protein complexes 1d3c, 1h6g, 1hty, 1i78, 1lvo, 1nvm, 1oo0, 1srq consist of a number of atoms ranging from 3 to 28. A ligand can have more than one binding site in the same protein, and also these binding sites can vary considerably in size. Thus, the ligand UPL (unknown branched fragment of phospholipid) has 27 binding sites on the same protein (1lsh), of which the smallest has only 4 atoms, while the biggest has 56 atoms. The ligand of the dataset that shows the largest variability is FAD (flavin-adenine dinucleotide), where the biggest of its 11 binding sites has 114 atoms and the smallest has just 10 atoms.

The cavity detection algorithm was run on the whole data set of 244 proteins. For each protein, it returned all cavities with more than a threshold number of atoms, ranked according to the number of atoms they contain. Thus rank one identifies the largest cavity, rank two the second largest cavity, and so on. This number is taken as an approximate measure of extension of the cavity. The number of cavities found on a protein vary considerably, depending on the size of the protein and its shape.

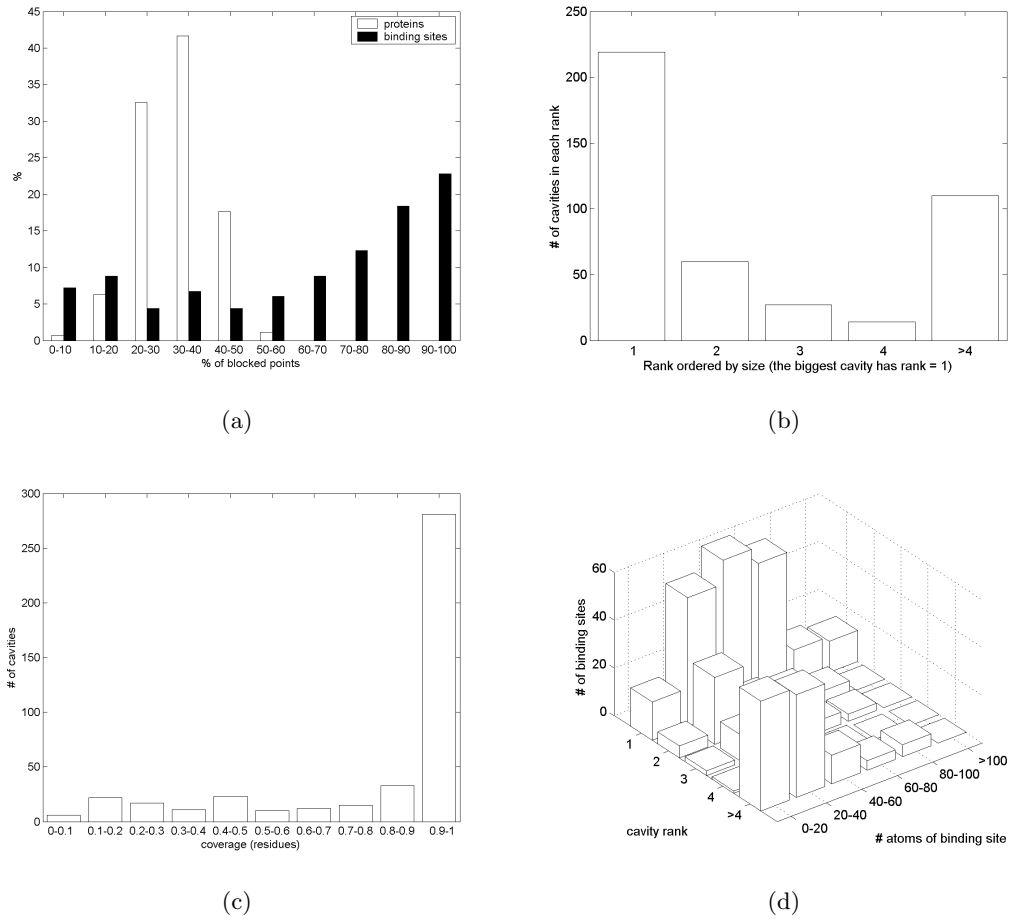
In analyzing the solutions, once again the measure of *coverage* of the residues (atoms) of the binding site, i.e. the fraction of residues (atoms) of the binding site found in the cavity, is used. A residue belongs to a cavity if at least one of the surface points close to it belongs to the cavity.

If the binding site of a ligand is known, the *best-coverage cavity* is defined as the cavity with the biggest coverage of the binding site. In discussing the results, only the best-coverage cavity for each complex of the dataset is considered. The best-coverage cavity is indicated simply as cavity in the following.

From line 6 and line 10 of algorithm 2 it follows that $R_{min} \leq R \leq Z_{max}$, where R is the radius of the sphere associated to the blocked point. To assess the optimal values for R_{min} and Z_{max}, the cavity detection algorithm was run with $R_{min} = \{0, 0.5, 1, 1.5, 2\}$ and $Z_{max} = \{5, 10, 15, 20, +\infty\}$ on 30 random proteins from the dataset. $R_{min} = 1$ Å and $Z_{max} = 10$ Å give the highest values of coverage and accuracy for the best-coverage cavities, and while changing Z_{max} to higher values doesn't affect much the results, changing R_{min} or using lower values for Z_{max} gives poor coverage values (data not shown).

The results of this procedure for the whole dataset are available at
`http://www.dei.unipd.it/%7Egaruttic/cavity/cavities07.xls`
Fig.3.2(b) shows the distribution of the best-coverage cavities according to their rank. It can be seen that in most cases the method identifies the binding site in the biggest cavity. Moreover, as shown in Fig.3.2(c), the values of coverage of residues of the binding sites are generally very high, with the majority of cavities achieving a coverage above 90%, which means that most of the times the binding site is completely included in the cavity. Fig.3.2(d) shows the distribution of binding sites by cavity rank and number of atoms of the binding site . The bigger the number of atoms of the binding site, the better the rank of the corresponding cavity. In fact, of the 88 binding sites that have less than 20 atoms, only 17 lie in the biggest cavity, while 63 binding sites are located in a cavity smaller than the fourth. The results improve if the number of atoms of the binding site increase. For instance, all but four of the 29 binding sites that have 80 or more atoms but less than 100 lie in one of the three biggest cavities, and all the 14 binding sites that have 100 atoms or more lie in the biggest cavity.

Table 3.1 shows the comparison between SURFNET-ConSurf and the cavity detection procedure described in this chapter. The top of table 3.1 shows the top 10 cavities found with the method described in this chapter, according to their values of coverage. All these cavities tightly include the binding site, and in the first seven cases they coincide with it. It can be seen that, for these 10 entries, the binding site is located in one of the four biggest cavities on 7 cases out of 10, which is competitive with the 3 out of



(a) (b)

(c) (d)

Figure 3.2: Statistics on the nonredundant data set of proteins Glaser *et al.* (2006). (a) histogram of percent of proteins (in white) and binding sites (in black) in the dataset, sorted on the horizontal axes according to their percentage of blocked points. (b) distribution of rank of the best-coverage cavities. (c) coverage of residues of binding sites. (d) distribution of binding sites by cavity rank and # atoms of binding site.

10 of SURFNET-ConSurf. Moreover, in all the entries but one, the procedure find that the best-coverage cavity has rank less than or equal to that of SURFNET-ConSurf. The only exception is for protein 1p6o with ligand HPY-411, but it can noted that this protein has several cavities with similar dimensions, and thus the ranking can be significantly different even with similar algorithms. The results at the bottom of table 3.1 show the biggest cavities. They all have rank one, high coverage, and a considerable number of atoms (more than 600). Five of the cavities found with SURFNET-ConSurf have rank higher than one, which suggests that these cavities are smaller than ours. This analysis suggests that these results are close to those of SURFNET-ConSurf, with a fast and still accurate geometrical method, without including any information about residues conservation.

Given a protein, the definition of its cavities is not unique. This is due to the fact that a protein is a closed 3-D surface, that can't be expressed using an analytical function for the whole surface. Hence, the cavity detection algorithm itself provides an operative definition for what a protein cavity is. Since the definition of cavity is not unique, cavity detection algorithms are not benchmarked on a dataset of well-known cavities, but rather on their ability to find cavities that contain the protein binding sites, relying on the property that ligands usually bind to cavities. For example, in CASTp a cavity is defined using Delaunay triangulation, alpha shape and discrete flow; in this case, the procedure fails to recognize a cavity whose Delaunay triangulation produces obtuse triangles with a discrete flow that goes to the outside or infinity, i. e. a cavity with a lateral wall that smoothly degrades into a flat region. This procedure identifies the blocked points of a protein, and then clusters them according to the overlapping spheres generated by the blocked points. This method might fail in identifying shallow cavities that don't have any blocked points, as well as those atoms that are surrounded by other atoms that belong to a cavity but that happen to have no blocked points. In the former case it's unusual to miss a binding site, since ligands most often bind into one of the biggest cavities; in the latter case, future extensions of the method that, in addition to blocked points, include unblocked points surrounded by blocked points, may solve the problem.

PdbID	Chain	Rank	Rank SURFNET-ConSurf	Cov	#Atoms of the b.s.	#Atoms of the cavity	#Atoms of the ligand	Ligand name
1ejj	A	4	> 4	1.00	24	24	11	3PG::601
1fw9	A	2	4	1.00	25	25	10	PHB::199
1h2r	SL	> 4	> 4	1.00	16	16	8	NFE::1004
1l9g	A	3	> 4	1.00	25	25	8	FS4::201
1p6o	AB	2	> 4	1.00	18	18	8	HPY::410
1p6o	AB	2	1	1.00	18	18	8	HPY::411
1qft	A	2	> 4	1.00	27	27	8	HSM::173
1otw	AB	> 4	> 4	1.00	42	46	24	PQQ::501
1p0z	A	2	4	1.00	38	42	13	FLC::1632
1o7d	ABCDE	> 4	> 4	1.00	26	29	8	TRS:A:2
1jv1	AB	1	1	0.95	62	1499	39	UD1::901
1jv1	AB	1	3	0.97	60	1499	39	UD1::902
1l3i	ABCD	1	1	0.97	62	1080	26	SAH::803
1l3i	ABCD	1	1	0.97	58	1080	26	SAH::802
1l3i	ABCD	1	1	0.96	57	1080	26	SAH::804
1l3i	ABCD	1	1	0.93	57	1080	26	SAH::801
1m98	AB	1	3	1.00	103	775	42	HEQ::351
1m98	AB	1	2	0.98	105	775	42	HEQ::350
1m98	AB	1	2	0.74	35	775	23	SUC::401
1nek	ABCD	1	3	0.88	141	766	77	CDN::308

Table 3.1: The ten cavities with the best values of coverage (top block) and biggest number of cavity atoms (bottom block). *PdbID* is the ID of the complex in the PDB. *Chain* is the chain used in the experiment. *Rank* is the identifier of the cavity of the protein with the best-coverage of the binding site. *Cov* and *# Atoms of the cavity* refer to the best-coverage cavity. *Cov* is the coverage expressed in terms of atoms. *# Atoms of the b.s.*, *# Atoms of the ligand* and *Name of the ligand* refer to the ligand as indicated in the PDB. *Ligand name* is expressed in the format `resname:chain:seqnumber`.

3.5.2 Cavity Matching

The benchmark algorithm 3 is done with an initial set of experiments on five pairs of proteins or chains (1atp with 1phk, 1csn, 1mjh chain B, 1hck and 1nsf) binding ATP from the representative set chosen in (58) and also used in the study by (9). As it has been observed also by (62), the ATP binding pockets in different proteins show great structural variability, although their size is about the same. The solutions are analyzed using the measure of coverage, i.e. the fraction of residues of the binding site found in the solution, and of accuracy, i.e. the fraction of residues in the solution that belong to the active site. A residue belongs to a solution if at least one of its surface points belongs to the solution. Table 3.2 shows the values of coverage and accuracy obtained when comparing the cavity with rank one of the Catalytic Subunit of cAMP-dependent Protein-Kinase (pdb:1atp, chain E) with those of proteins 1phk, 1csn, 1mjh, 1hck and 1nsf. Also the values of coverage of the binding site obtained by MolLoc are shown. Chapter 2 doesn't show the accuracy values for MolLoc; although the solution regions had a significant overlap with the binding sites, they spanned areas much larger than the binding sites. Indeed the goal of MolLoc was to identify similar regions on protein surfaces, not to find binding sites. For the proteins 1atp and 1csn, which both bind to the ligand ATP, the two most similar regions on each protein are part of the binding site and this explains also the high values of coverage for MolLoc. In both proteins, the binding sites are located in the top cavity. The new method improves on coverage while at the same time obtaining a good accuracy for all pairwise comparisons. The execution time is drastically reduced w.r.t. MolLoc. While MolLoc took about two hours to execute, the new method took less than two minutes.

From the observations in the section 3.5.1 about the difference in size of different binding sites for the same ligand, it is evident that any matching procedure based on purely geometric criteria will fail to recognize binding sites for those cases. Nevertheless, if more than two proteins share similar regions in correspondence of the binding sites, then those regions are likely to be conserved structures with a functional characterization. The next sections show how collecting the information of different matchings, by means of clustering techniques, enhances functional recognition.

Pdb ID	# residues in binding site	Coverage	Coverage Cavity comparison	Accuracy Cavity comparison	Sequence Identity
1phk	26(23)	0.69(0.78)	0.90(0.91)	0.76(0.80)	34.3%
1csn	26(23)	0.62(0.70)	0.80(0.78)	0.91(0.75)	19.0%
1mjh:B	25(23)	0.24(0.26)	0.32(0.34)	0.88(1.00)	4.7%
1hck	24(23)	0.42(0.39)	0.58(0.56)	0.87(0.92)	29.6%
1nsf	23(23)	0.35(0.43)	0.43(0.60)	0.76(0.93)	8.3%

Table 3.2: Comparison of 1atp (cAMP-dependent Protein-Kinase) with 1phk (Subunit of glycogen phosphorylase kinase), 1csn (Casein kinase-1), 1mjh:B ("Hypothetical" protein MJ0577), 1hck (Cyclin dependent PK) and 1nsf (Examerization domain of N-ethilmalemide-sensitive fusion protein). In brackets, the values for 1atp. The sequence identity values are obtained with CE (57)

.

3.5.3 All-To-All Pairwise Cavity Comparison

To further test how cavity detection and cavity matching together can be used to identify proteins with similar function, an all-to-all pairwise comparison was performed on a dataset (shown in table 3.3) previously used by (48). This dataset has 40 proteins with low pairwise sequence similarity, divided in 4 sets of 10 proteins that bind different ligands. The proteins of the first three sets bind respectively ATP, NAD and heme, while those belonging to the last set bind five distinct but chemically similar steroids (estradiol, progesterone, equitinin, testosterone and dihydrotestosterone).

Ligand	Pdb
ATP	1asz(AR), 1awm(AB), 1b38(A), 1b76(A), 1d9z(A), 1dv2(A), 1e4g(T), 1e8x(A), 1f9a(A), 1fmw(A)
NAD	1a4z(A), 1ad3(A), 1ahh(A), 1b14(A), 1bmd(A), 1bxk(A), 1bxs(A), 1cer(O), 1e3l(A), 1nff(A)
HEME	102m, 155c, 1a00(A), 1a2f, 1apx(A), 1arp, 1atj(A), 1b7v, 1b80(A), 1bgp
STEROID	1a28(A), 1a52(A), 1cqs(A), 1dbb(LH), 1ere(A), 1i37(A), 1i9j(HL), 1jtv(A), 1kdk(A), 1ogz(A)

Table 3.3: Dataset of proteins for the all-to-all pairwise comparison. The chain used is indicated in brackets.

In the dendrogram in Fig.3.3(a), the higher the number of correspondences between the cavities of two proteins, the more conserved is their structure and the sooner the cavities are clustered together. The dendrogram shows that the protein cavities that are the most structurally similar are those of steroids binding proteins, followed by hemes and by NADs and ATPs.

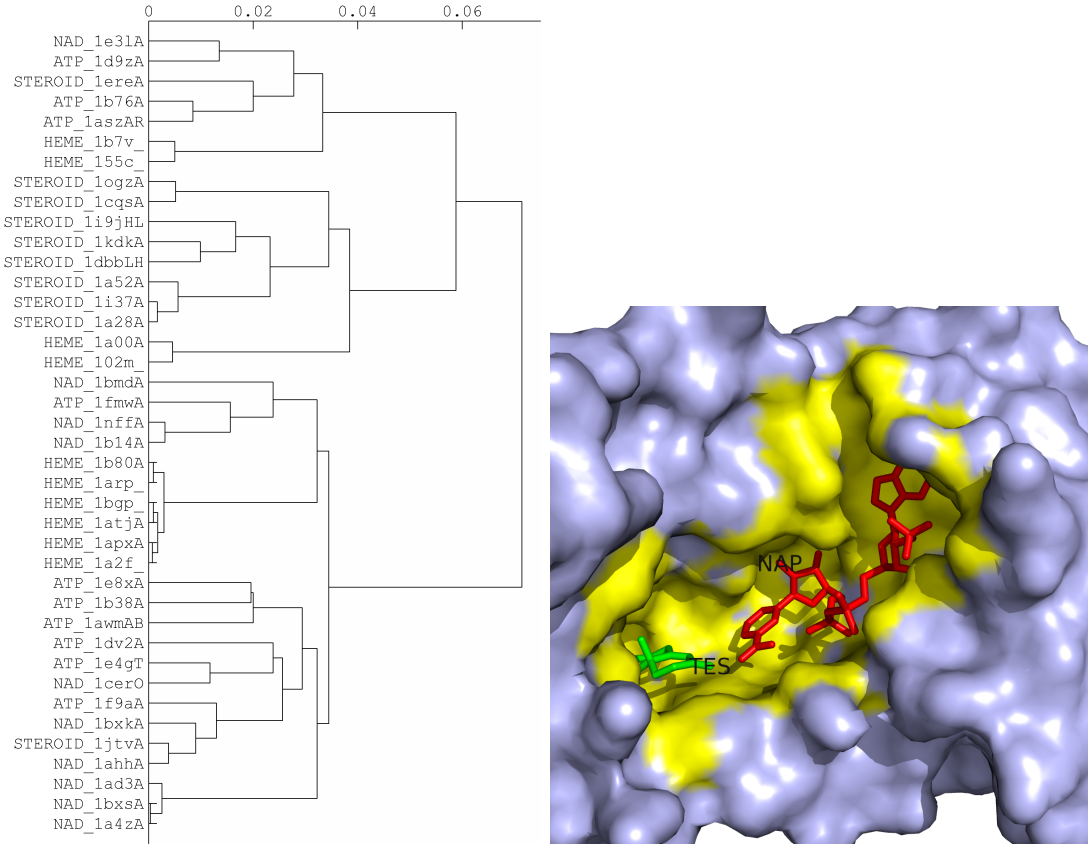
Eight protein cavities that bind steroids cluster together (1ogz, 1cqs, 1i9j, 1kdk,

1dbb, 1a52, 1i37 and 1a28). This strong recognition is due to the fact that the steroids are –relatively– rigid ligands, and thus also their binding sites are rather structually conserved. However, two proteins that bind steroid aren't recognized as such. The first is the estrogen receptor (1ere, chain A), misrecognized also by (48), where the steroid is buried into an internal cavity. Since it is the only protein of the steroids dataset where the binding site lies into an internal cavity, its conformation is significantly different from those of the other steroids binding cavities, and thus the matching fails. The second protein that is not recognized as steroid binding is 17beta-hydroxysteroid dehydrogenase type 1 (1jtv, chain A), that is clustered with six NADs and three ATPs. In this protein, the binding site of the steroid lies into a big cavity, which hosts also a more extended NAD(P) binding site[1] on the opposite part of the cavity (see Fig.3.3(b)). Since the pairwise comparison returns the most extended similar regions, the comparisons with the NAD binding proteins have the highest number of correspondences. Furthermore, 17beta-hydroxysteroid dehydrogenase type 1 is an NAD(P)-binding Rossmann-fold domain, and thus the most extended similar region, again, happens to be the functional region of the protein.

Six hemes binding proteins tightly cluster together (1b80, 1arp, 1bgp, 1atj, 1apx and 1a2f) and the other four form two isolated pairs (1b7v and 155c, 1a00 and 102m). This represents an improvement to (48), where a cluster of six hemes contains also an ATP, and two isolated hemes are paired with two ATPs.

In regard to NADs and ATPs binding proteins, they cluster together in the large cluster at the bottom of the dendrogram that includes the 17beta-hydroxysteroid dehydrogenase type 1 (1dv2, 1e4g, 1cer, 1f9a, 1bxk, 1jtv, 1ahh, 1ad3, 1bxs, 1a4z), in a cluster of four cavities(1bmd, 1fmw, 1nff and 1b14), a triplet(1e8x, 1b38 and 1awm) and two pairs(1b76 and 1asz, 1e3l and 1d9z). The reason why NADs and ATPs don't participate in separate clusters is that both ligands are extremely flexible, with the exception of an adenine ring that is common to both structures. Moreover, the slight preponderance of NADs than ATPs in the large cluster reflects NAD narrower range of possible conformations(Stockwell and Thornton, 2006).

[1] the pdb id of the complex with NAD(P) is 1a27



(a) (b)

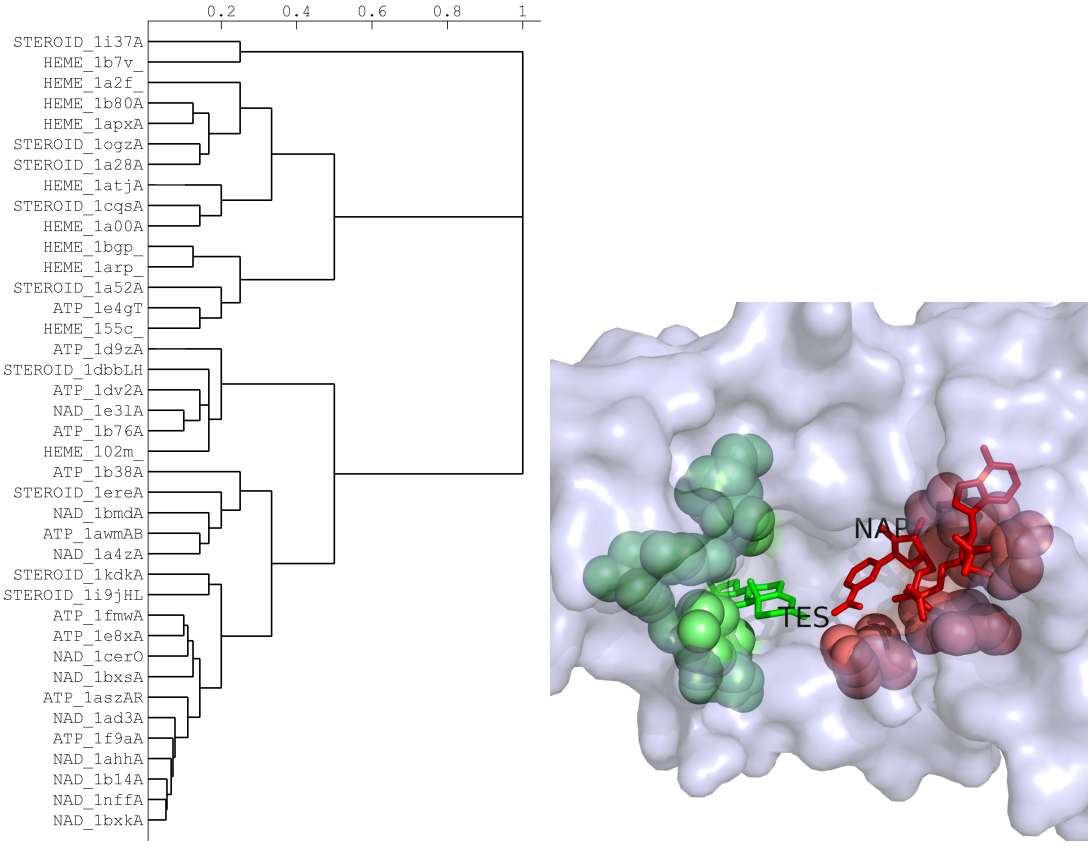
Figure 3.3: (a) Hierarchical clustering with the complete linkage (furthest distance) in the all-to-all pairwise comparison. (b) The large cavity of 17beta-hydroxysteroid dehydrogenase type 1 hosts both a steroid (TES, in 1jtv) and a NAD(P) (NAP, in 1a27).

3.5.4 Background Cavity

Looking at the dataset, the natural choice for a protein background cavity to test the procedure described in algorithm 5 is 1jtv, which binds TES and NAD(P) in two different regions of the same cavity. The dendrogram for 1jtv best-coverage cavity is shown in Fig.3.4(a), and it identifies four distinct clusters with no residues in common on the background protein because the complete linkage distance equals 1. There is a cluster at the bottom of the dendrogram with seven NADs and ATPs (1asz, 1ad3, 1f9a, 1ahh, 1bxk, 1nff and 1b14), and just two steroids pairs (1ogz and 1a28, 1kdk and 1i9j). Fig.3.4(b) shows how this procedure finds both binding sites; the common residues identified by the cluster of the seven NADs-ATPs cavities belong to the NAP binding site, while the common residues identified by one of the two pairs of steroids (1ogz and 1a28 in this figure) belong to the TES binding site.

3.6 Conclusions

This chapter presented a method for binding site recognition that is effective and fast. It uses only geometric criteria and a description of the protein surfaces by means of a collection of two-dimensional arrays, the spin images, each describing the spatial arrangement of the protein surface points in the vicinity of a given surface point. As mentioned, there are cases where the recognition procedure fails to identify the correct binding sites. When a ligand binds different proteins at sites that vary significantly in size and shape, most of existing approaches are inadequate to identify the binding location. However, the all-to-all pairwise comparison approach groups together structurally similar cavities when dealing with a large collection of proteins. Moreover, the comparison of a background cavity with a dataset of protein cavities can identify multiple binding sites on the background cavity. Physico-chemical properties can be easily incorporated to the presented geometrical methods, by adding a labeling to the points during the comparison phase and comparing only points with the same label. Alternatively, they can be used in the final stage of the matching to prune from the final geometrical solution the correspondences that link two points with different properties.



(a) (b)

Figure 3.4: (a) Hierarchical clustering with the complete linkage (furthest distance) in the pairwise comparison with the background cavity 1jtv. (b) Both binding sites on 1jtv are now identified using the clustering procedure with 1jtv as background cavity; in red NAD(P) ligand and the common residues identified by the cluster of seven NADs-ATPs cavities, and in green the steroid TES and the common residues identified by 1ogz and 1a28.

4

MolLoc Web Server: a Tool for Local Molecular Surface Alignment

MolLoc stands for ***Mol****ecular* ***Loc****al alignment* and is the name of the method described in Chapter 2 as well as the name of a web server for the local alignment of molecular surfaces. The surfaces may be restricted to cavities, binding sites or any residue selection of a complete protein, RNA or DNA. The server determines the most extended similar regions of the two selected surfaces. This application can be particularly useful when the user is interested in inferring functional information for a molecule, be it a protein, RNA or DNA. MolLoc accepts files in PDB format, which is described in the documentation of the PDB repository[1]. The web server is available at `http://bcb.dei.unipd.it/MolLoc/` .

4.1 Molecular Surface

The molecular surface representation has been described in 1.3. It is built by rolling a probe sphere with the size of a water molecule (radius 1.4$\mathring{A}$) over all the `ATOM` and `ANISOU` atoms of the chosen chain(s), and discarding the `HETATM` atoms, according to the PDB format. The probe sphere generates oriented points (i.e. points with normals). Then, a spin-image is built for each oriented point. The oriented points and their spin-

[1] www.rcsb.org

images describe the molecular surface. The `HETATM` atoms are excluded from the surface generation routine becouse `HETATM` is the label that is applied to water molecules and to ligand atoms, while most applications in structural biology require to compare the structures of proteins and nucleic acids atoms only. If the application needs to take into account one or more `HETATM` atoms, the user can edit the PDB file and substitute the label `HETATM` with a label `ATOM` for the atoms of interest.

4.2 Input

The first page of MolLoc allows the user to insert a pdb code, or to upload a file in PDB format. Moreover, the user can write his/her email to receive a link to the results page at the end of the computation.

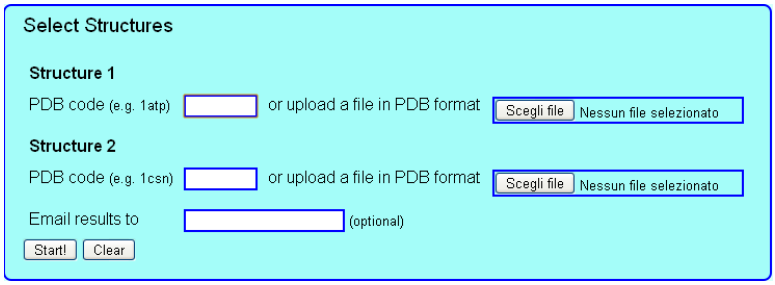


Figure 4.1: Home page of MolLoc

The second page shows the chain(s) selection, and requires the user to choose one or more chains. At this stage, MolLoc builds the molecular surfaces.

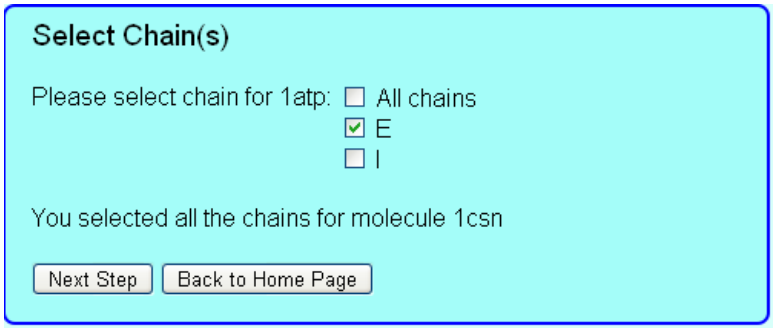


Figure 4.2: Chains selection

The third page shows the selected chains and their ligands. MolLoc shows all the ligands that are closer than $4\mathring{A}$ to at least one atom of the chains that have been

selected from the input structure. With respect to the PDB format, a ligand is defined as a residue of `HETATM` atoms whose residue name is not `HOH` (water).

The user must provide an atom selection for each of the two structures, and the comparison will run on the surfaces belonging to those atoms selections. For each structure the user can choose between:

- one or more binding sites
- the molecular cavities
- an arbitrary set of residues

The user must provide non-empty selections for both structures. A ligand binding site is the set of `ATOM` atoms that belong to the molecular surface and that are closer than $6\mathring{A}$ from at least one ligand atom. Moreover, the user can choose the alignment method. With "Only surface point alignment" the alignment is computed superimposing only molecular surface points. With "Atoms alignment after surface point alignment", an initial alignment is obtained by superimposing the surface points; then the obtained alignment is iteratively refined based on the centers of the atoms with the same atom-type. Section 4.4 decribes the method to find the cavities, while section 4.5 describes the two alignment methods.

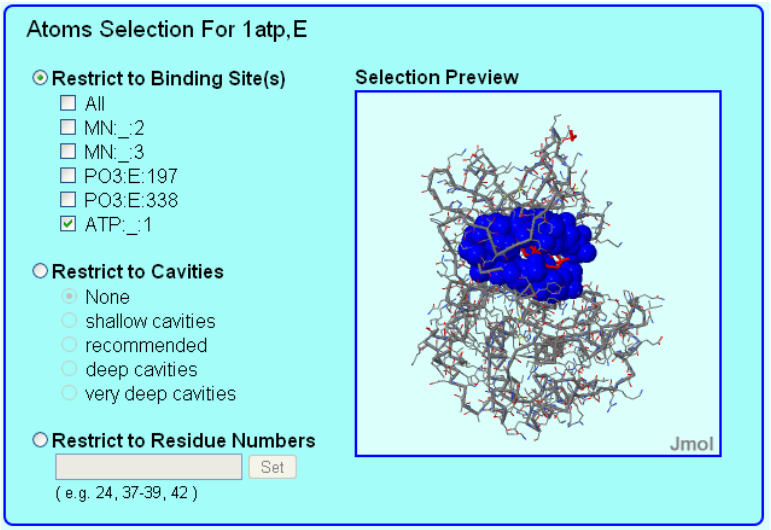
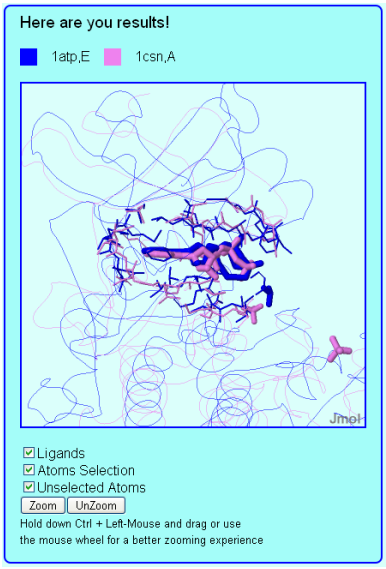


Figure 4.3: Atoms selection

4.3 Output

The fourth page (not shown) has a summary of the selection and the status of the current computation, while the fifth page contains the results. The user can download the results in various formats, view the results on JMol, highlight the atoms correspondences and optionally leave a comment.



41 atoms correspondences RMSD = 0.111284

Molecule 1			Molecule 2				
RES.	RES.NUM.	ATOM	RES.	RES.NUM.	ATOM	TYPE	☐
VAL	57	CG1	ILE	26	CG2	AL	☐
VAL	57	CB	ILE	26	CB	AL	☐
VAL	123	CG1	LEU	88	CD2	AL	☐
VAL	123	CG2	LEU	88	CD1	AL	☐
VAL	57	CG2	ILE	26	CG1	AL	☐
TYR	122	O	LEU	87	O	AC	☐
THR	51	O	GLU	20	O	AC	☐
THR	51	C	GLU	20	C	PI	☐
THR	51	N	GLU	20	N	DO	☐
SER	53	O	SER	22	O	AC	☐
SER	53	OG	SER	22	OG	DA	☐
SER	53	N	SER	22	N	DO	☐
PHE	54	N	PHE	23	N	DO	☐
PHE	54	O	PHE	23	O	AC	☐
PHE	54	C	PHE	23	C	PI	☐
MET	120	SD	ILE	85	CD1	AL	☐
LYS	168	NZ	LYS	133	NZ	DO	☐
LYS	72	CD	LYS	41	CD	AL	☐

(a) (b)

Figure 4.4: Results page in MolLoc

4.4 FastCav: fast cavity detection

Chapter 3 shows how limiting the surface comparison of two structures to their binding sites and their cavities can improve the recognition of similar functional regions. As mentioned in the Introduction, a study on 175 enzymes indicates that the residues of the binding site are very often found among the 5% of residues closest to the enzyme centroid (4). Nevertheless, this is often not true for those proteins that are not globular (ex. protein fibrils), or for binding sites of small ligands (ex. zinc fingers). MolLoc implements a routine for cavity detection that is fast and that retrieves cavities at different depth levels. The routine is described in algorithm 6

Algorithm 6 FastCav

Input: protein P, parameters R,s

Output: set of cavity atoms C

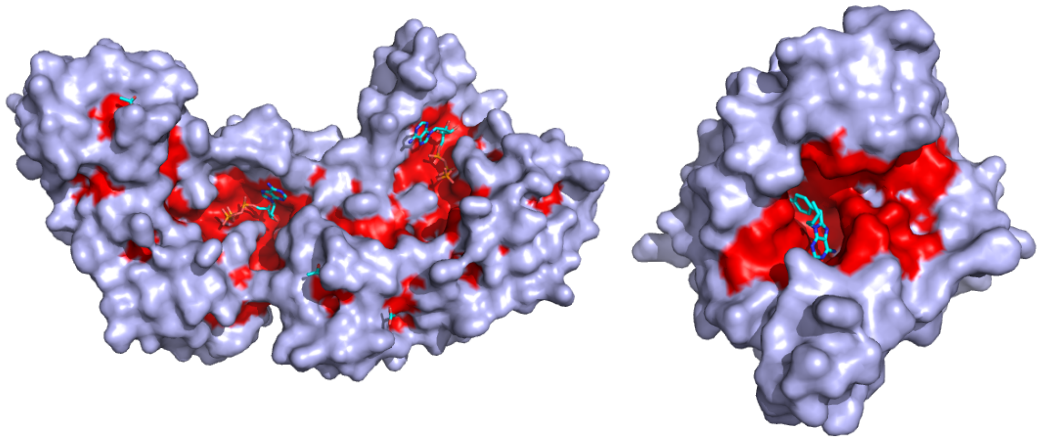
1: find the set S of surface atoms, $S \subset P$
2: **for all** the atoms $a_i \in S$ **do**
3: define $N(a_i) = \left|\left\{a_j : a_j \in P, \|a_i - a_j\|_2 < R\right\}\right|$
4: **end for**
5: define $\mu = \frac{1}{|S|}\sum N(a_i)$ and $\sigma = \sqrt{\frac{1}{|S|}\sum (N(a_i) - \mu)^2}$
6: **return** set of cavity atoms $C = \{a_i \in S : N(a_i) > \mu + s\sigma\}$

The procedure can be summarized as follows. For each surface atom a_i, count the number $N(a_i)$ of neighbours, i.e. of atom centers (including non-surface atoms) that lie within a radius R from the center of a_i. The distribution of $N(a_i)$ has mean μ and standard deviation σ. Then, an atom belongs to a cavity if the number of its neighbours is larger than $\mu + s\sigma$. The value of R determines how deep the cavities are, while the value of s determines their size. This algorithm relies on the fact that internal atoms are tightly packed (26), and that cavity atoms are less exposed to the solvent than flat regions or convex regions.

The optimal for the two parameters are $R = 8$ and $s = 0.5$. They were experimentally determined by running FastCav on 30 random proteins from the dataset described in (25) with the values $R = 4, 8, 12, 20$ and $s = 0, 0.5, 1, 1.5$. While different values of R allow detection of binding sites of ligands with different sizes, extreme values of s deteriorate the cavity detection for all ligands. In particular, for $s = 0$ the number of selected atoms is about half of the overall atoms and therefore the putative cavities are not informative; for $s = 1.5$ the coverage of the binding sites is less than 40% for all the ligand binding sites of the 30 random proteins.

Figure 4.5(a) shows the cavities of cystic fibrosis transmembrane conductance regulator (CFTR)[1] that identify the ligands binding sites with $R = 12$, and figure 4.5(b) shows the deep cavity of the human chaperon hsp90[2] that identify the binding site of the purine-based inhibitor with $R = 20$. The values used in MolLoc are $R = 4, 8, 12, 20$ to offer to the user the possibility to create cavities ranging from shallow cavities ($R = 4$)

[1] Pdb id: 1r10
[2] Pdb id: 1uy7



(a) (b)

Figure 4.5: (a) cavities of cystic fibrosis transmembrane conductance regulator with $R = 12$; (b) cavity of the human chaperon hsp90 with $R = 20$.

to very deep cavities (R=20).

4.5 Alignment Methods

In MolLoc, the user can choose between two alignment methods: *only geometrical* and *geometrical+atomtypes*. The first method builds upon the method described in chapter 2, and can be summarized as follows. Given two structures, MolLoc builds their molecular surfaces, i.e. one cloud of oriented points for each structure. Then, the method considers an atom selection for each structure, and builds a spin-image representation for each molecular surface point belonging to a selected atom. Next, each spin-image of the first structure is compared with each spin-image of the second structure, and their 2D correlation is computed. Given an oriented point P belonging to the first structure and a point Q belonging to the second structure, P and Q are put in correspondence if they are similar, according to the conditions described in Section 2.4. This produces a set of points correspondences. This set is pruned using a heuristic that keeps only a subset of correspondences that are geometrically consistent. This reduced subset is used as input in Horn's method to produce a rototranslation that superimposes the surface of the first structure on the surface of the second structure.

The second method uses the "only geometrical" to obtain a first alignment. Then,

Donor (DO)	
ARG	NE,NH1,NH2
ASN	ND2
GLN	NE2
LYS	NZ
TRP	NE1,CZ1,CZ3,CH
all	N
Acceptor (AC)	
ASN	OD1
ASP	OD1
ASP	OD2
GLN	OE1
GLU	OE1,OE2
all	O
Mixed Donor/Acceptor (DA)	
HIS	NE1,NE2
SER	OG
THR	OD1
TYR	OH

Hydrophobic Aliphatic (AL)	
ALA	CB
ARG	CB,CG,CD
CYS	CB,SG,
ILE	CB,CG1,CG2,CD1
LEU	CB,CG,CD1,CD2
LYS	CB,CG,CD,CE
MET	CB,CG,SD,CE
PRO	CB,CG,CD,
THR	CD2
VAL	CB,CG1,CG2
Aromatic (PI)	
HIS	CG,ND1,CD2,CE1,NE2
PHE	CG,CD1,CD2,CE1,CE2,CZ
TRP	CG,CD1,CD2,NE1,CE2,CE3
TYR	CB,CD1,CD2,CE1,CZ
all	C

Table 4.1: Table with atomtypes as in (54)

the method refines the alignment based on atoms with the same atomtype that are close after the superimposition. Thus, the method finds the selected atoms of the first structure that are closer than 2.5$\mathring{A}$ to any selected atom of the second structure, and that have the same atomtype (54) as shown in table 4.1. These atoms are put in correspondence, and atoms that are in correspondence with more than one atom are discarded. The result is a set of one-to-one correspondences, that is used as input in Horn's method to produce another rototranslation. Again, the atoms with the same atomtype that are close after the rototranslation are put in correspondence and a new rototranslation is computed, iteratively. The iterations stops either after 10 steps, or if the set of correspondences has a cardinality that is inferior to that of the previous rototranslation.

Also electrostatics can be taken into account to some extent. Two interfaces that are structurally similar and align atoms with the same atomtypes can be analyzed with tools such as VASCo (61) that visualize differences in electrostatic paths between aligned surfaces.

4.6 Conclusions

MolLoc web server is a tool for the local alignment of molecular surfaces, that allows comparison of regions that are functionally relevant for a molecule, be it a protein, RNA or DNA. The web server provides visualization of the regions that are compared, and visualization of the results, within a JMol applet. The user can choose between a method that aligns the surface atoms only, and a method that uses addictional information on the atomtypes to refine the alignment.

5

Conclusions

The way proteins interact with the other molecules in the cell determines their function and therefore the processes that regulate life.

Discovery of a similar region on two protein surfaces can lead to important inference about the functional role or molecular interaction of this region for one of the proteins if such information is available for the other. This work proposes a new characterization of protein surfaces based on a spin-image representation of the surfaces that facilitates the simultaneous search of the surface of each of two proteins for a matching region.

The method described in Chapter 2 finds regions of similarity on two protein surfaces and produced good results when tested on known families of proteins. Restricting the local surface comparison to functional regions, like binding sites and cavities, improves the quality of the results and speeds–up the computation.

All-to-all pairwise comparison are used to group together structurally similar cavities when dealing with a large collection of proteins. Moreover, the comparison of a background cavity with a dataset of protein cavities can identify multiple binding sites on the background cavity.

Nevertheless, even if shape is fundamental in determining the protein's function (50), also physico-chemical properties and electrostatics play a crucial role.

Physico-chemical properties can be easily incorporated to the presented geometrical methods, by adding a labeling to the points during the comparison phase and comparing only points with the same label, as well as by pruning from the final geometrical solution the correspondences that link two points with different properties. A modified version of the purely geometrical method described in Chapter 2 has been implemented in MolLoc

web server, which is a tool for the local alignment of molecular surfaces, that allows comparison of regions that are functionally relevant for a molecule, be it a protein, RNA or DNA. In MolLoc web server, the user can choose an alignment method that refines the rototranslation obtained from the surface points, by using information on the atomtypes.

Bibliography

[1] R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, and M. Magrane. **UniProt: the Universal Protein knowledgebase.** *Nucleic Acids Research*, **32**:D115, 2004. 45

[2] W.P. Arend. **Interleukin-1 receptor antagonist**. *Advances in Immunology*, **54**:167227, 1993. 2

[3] J.A. Barker and J.M. Thornton. **An Algorithm for Constraint-Based Structural Template Matching: Application to 3D Templates with Statistical Analysis**. *Bioinformatics*, **19**(13):1644–1649, 2003. 18, 37

[4] A. Ben-Shimon and M. Eisenstein. **Looking at Enzymes from the Inside out: The Proximity of Catalytic Residues to the Molecular Centroid can be used for Detection of Active Sites and Enzyme–Ligand Interfaces**. *Journal of Molecular Biology*, **351**(2):309–326, 2005. 4, 60

[5] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. **The Protein Data Bank**. *Nucleic Acids Research*, **28**(1):235–242, Jan 1 2000. 25, 45

[6] T. A. Binkowski, A. Joachimiak, and J. Liang. **Protein surface analysis for function annotation in high-throughput structural genomics pipeline**. *Protein Science*, **14**(12):2972, 2005. 37

[7] T.A. Binkowski, L. Adamian, and J. Liang. **Inferring functional relationships of proteins from local sequence and spatial surface patterns**. *Journal of Molecular Biology*, **332**(2):505–526, Sep 12 2003. 37

[8] M.E. Bock, G.M. Cortelazzo, C. Ferrari, and C. Guerra. **Identifying Similar Surface Patches on Proteins using a Spin-Image Surface Representation**. *Proceedings CPM*, pages 417–428, 2005. 18

[9] M.E. Bock, C. Garutti, and C. Guerra. **Discovery of Similar Regions on Protein Surfaces**. *Journal of Computational Biology*, **14**(3):285–299, 2007. 1, 36, 51

[10] M.E. Bock, C. Garutti, and C. Guerra. **Effective Labeling of Molecular Surface Points for Cavity Detection and Location of Putative Binding Sites**. *Computational Systems Bioinformatics Conference*, pages 263–274, 2007. 1

[11] M.E. Bock, C. Garutti, and C. Guerra. **Cavity Detection and Matching for Binding Site Recognition**. *Theoretical Computer Science*, **408**:151–162, 2008. 1

[12] G.P. Brady and P.F.W. Stouten. **Fast prediction and visualization of protein binding pockets with PASS**. *Journal of computer-aided molecular design*, **14**(4):383–401, 2000. 37

[13] A. Brakoulias and R.M. Jackson. **Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching**. *Proteins*, **56**(2):250–260, Aug 1 2004. 37, 38

[14] J.M. Carsi, H.H. Valentine, and L.T. Potter. **m2-Toxin: A Selective Ligand for M2 Muscarinic Receptors**. *Molecular Pharmacology*, **56**(5):933–937, 1999. 2

[15] B.Y. Chen, D.H. Bryant, V.Y. Fofanov, D.M. Kristensen, A.E. Cruess, M. Kimmel, O. Lichtarge, and L.E. Kavraki. **Cavity-Aware Motifs Reduce False Positives in Protein Function Prediction**. *Computational Systems Bioinformatics Conference*, pages 311–323, 2006. 18, 38

[16] M. Comin, C. Guerra, and G. Zanotti. **PROuST: A Comparison Method of Three-Dimensional Structures of Proteins Using Indexing Techniques**. *Journal of Computational Biology*, **11**(6):1061–1072, 2004. 7

[17] M. Comin, C. Guerra, and G. Zanotti. **Proust: a Comparison Method of Three-Dimensional Structures of Proteins using Indexing Techniques**. *Journal of Computational Biology*, **11**(6):1061–1072, 2004. 29, 33

[18] M.L. Connolly. **Analytical Molecular Surface Calculation**. *Journal of Applied Crystallography*, **16**(5):548–558, 1983. 25, 46

[19] L.L. Conte, C. Chothia, and J. Janin. **The atomic structure of protein-protein recognition sites**. *Journal of Molecular Biology*, **285**(5):2177–2198, 1999. 3

[20] K.D. Corbett and J.M. Berger. **Structural basis for topoisomerase VI inhibition by the anti-Hsp90 drug radicicol.** *Nucleic Acids Research*, **34**(15):4269, 2006. 4

[21] D.G. Covell and A. Wallqvist. **Analysis of protein-protein interactions and the effects of amino acid mutations on their energetics. The importance of water molecules in the binding epitope**. *Journal of Molecular Biology*, **269**(2):281–297, 1997. 4

[22] L.P. Ehrlich and R.C. Wade. **Protein-Protein Docking**. *Reviews in Computational Chemistry*, **17**:61–98, 2001. 3

[23] M. Gerstein. **A resolution-sensitive procedure for comparing protein surfaces and its application to the comparison of antigen-combining sites**. *Foundations of Crystallography*, **48**(3):271–276, 1992. 7

[24] F. Glaser, R.J. Morris, R.J. Najmanovich, R.A. Laskowski, and J.M. Thornton. **A Method for Localizing Ligand Binding Pockets in Protein Structures**. *Proteins*, **62**(2):479–488, Feb 1 2006. 18, 36, 37, 39, 45

[25] F. Glaser, Y. Rosenberg, A. Kessel, T. Pupko, and N. Ben-Tal. **The ConSurf-HSSP database: The mapping of evolutionary conservation among homologs onto PDB structures**. *Proteins*, **58**(3):610–617, 2005. 4, 36, 61

[26] M. Hao, S. Rackovsky, A. Liwo, M.R. Pincus, and H.A. Scheraga. **Effects of Compact Volume and Chain Stiffness on the Conformations of Native Proteins**. *Proceedings of the National Academy of Sciences*, **89**(14):6614–6618, 1992. 61

[27] L. Holm and C. Sander. **Protein Structure Comparison by Alignment of Distance Matrices**. *Journal of Molecular Biology*, **233**:123–123, 1993. 7

[28] B. Huang and M. Schroeder. **LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation**. *BMC Structural Biology*, **6**:19, Sep 24 2006. 37

[29] A.E. Johnson. *Spin-Images: A Representation for 3-D Surface Matching*. PhD thesis, Carnegie Mellon University, Pittsburgh, August 1997. 10, 12, 13

[30] A.E. Johnson and M. Hebert. **Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21**(5):433–449, 1999. 9, 22

[31] K. Kinoshita, J. Furui, and H. Nakamura. **Identification of Protein Functions from a Molecular Surface Database, eF-site**. *Journal of Structural and Functional Genomics*, **2**(1):9–22, 2002. 18, 37

[32] G.J. Kleywegt. **Recognition of Spatial Motifs in Protein Structures**. *Journal of Molecular Biology*, **285**(4):1887–1897, Jan 29 1999. 18, 37

[33] N. Kobayashi and N. Go. **A Method to Search for Similar Protein Local Structures at Ligand-Binding Sites and its Application to Adenine Recognition**. *European Biophysics Journal*, **26**(2):135–144, 1997. 7, 18, 37

[34] R. Kolodny, P. Koehl, and M. Levitt. **Comprehensive Evaluation of Protein Structure Alignment Methods: Scoring by Geometric Measures**. *Journal of Molecular Biology*, **346**(4):1173–1188, 2005. 5

[35] I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge, and T.E. Ferrin. **A geometric approach to macromolecule-ligand interactions.** *Journal of Molecular Biology*, **161**(2):269–288, 1982. 37

[36] Y.Y. Kuttner, V. Sobolev, A. Raskind, and M. Edelman. **A Consensus-Binding Structure for Adenine at the Atomic Level Permits Searching for the Ligand Site in a Wide Spectrum of Adenine-Containing Complexes**. *Proteins Structure Function and Genetics*, **52**(3):400–411, 2003. 25, 37

[37] R.A. Laskowski. **SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions**. *Journal of Molecular Graphics*, **13**(5):323–30, 307–8, Oct 1995. 36, 37

[38] R.A. Laskowski. **PDBsum: summaries and analyses of PDB structures**. *Nucleic Acids Research*, **29**(1):221–222, 2001. 45

[39] R.A. Laskowski, J.D. Watson, and J.M. Thornton. **Protein Function Prediction Using Local 3D Templates**. *Journal of Molecular Biology*, **351**(3):614–626, 2005. 18

[40] A.T. Laurie and R.M. Jackson. **Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening**. *Current Protein & Peptide Science*, **7**(5):395–406, Oct 2006. 37

[41] B. Lee and F.M. Richards. **The interpretation of protein structures: Estimation of static accessibility**. *Journal of Molecular Biology*, **55**:379–400, 1971. 8

[42] D.G. Levitt and L.J. Banaszak. **POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids**. *Journal of Molecular Graphics*, **10**(4):229–234, 1992. 37

[43] J. Liang, H. Edelsbrunner, P. Fu, P.V. Sudhakar, and S. Subramaniam. **Analytical Shape Computation of Macromolecules: I. Molecular Area and Volume Through Alpha Shape**. *PROTEINS: Structure, Function, and Genetics*, **33**:1–17, 1998. 37

[44] J. Liang, H. Edelsbrunner, P. Fu, P.V. Sudhakar, and S. Subramaniam. **Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins**. *Proteins Structure Function and Genetics*, **33**(1):18–29, 1998. 37

[45] L. Lo Conte, C. Chothia, and J. Janin. **The Atomic Structure of Protein-Protein Recognition Sites**. *Journal of Molecular Biology*, **285**(5):2177–2198, 1999. 18, 37

[46] J.A. McCammon and S.C. Harvey. *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press, 1987. 4

[47] K. Mehlhorn and S. Näher. *LEDA: A Platform for Combinatorial and Geometric Computing*. Cambridge University Press, 1999. 33

[48] R.J. Morris, R.J. Najmanovich, A. Kahraman, and J.M. Thornton. **Real Spherical Harmonic Expansion Coefficients as 3D Shape Descriptors for Protein Binding Pocket and Ligand Comparisons**. *Bioinformatics*, **21**(10):2347–2355, 2005. 18, 32, 36, 37, 38, 52, 53

[49] R. Najmanovich, N. Kurbatova, and J. Thornton. **Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites.** *Bioinformatics*, **24**(16):i105, 2008. 7

[50] R. Najmanovich, N. Kurbatova, and J. Thornton. **Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites**. *Bioinformatics*, **24**(16):i105, 2008. 65

[51] R. J. Najmanovich, A. AllaliHassani, R. J. Morris, L. Dombrovsky, P. W. Pan, M. Vedadi, A. N. Plotnikov, A. Edwards, C. Arrowsmith, and J. M. Thornton. **Analysis of binding site similarity, small-molecule similarity and experimental binding profiles in the human cytosolic sulfotransferase family**. *Bioinformatics*, **23**(2):e104, 2007. 37

[52] D.W. Ritchie. **Recent Progress and Future Directions in Protein-Protein Docking**. *Current Protein and Peptide Science*, **9**(1):1–15, 2008. 4

[53] M. Rosen. **Molecular Shape Comparisons in Searches for Active Sites and Functional Similarity**. *Protein Engineering Design and Selection*, **11**(4):263–277, 1998. 18

[54] S. Schmitt, D. Kuhn, and G. Klebe. **A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology**. *Journal of Molecular Biology*, **323**(2):387–406, 2002. 63

[55] M. Shatsky, A. Shulman-Peleg, R. Nussinov, and H.J. Wolfson. **The Multiple Common Point Set Problem and Its Application to Molecule Binding Pattern Detection**. *Journal of Computational Biology*, **13**(2):407–428, 2006. 37

[56] I.N. Shindyalov. **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path**. *Protein Engineering Design and Selection*, **11**(9):739–747, 1998. 7

[57] I.N. Shindyalov and P.E. Bourne. **Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path**. *Protein Engineering*, **11**(9):739–747, Sep 1998. 29, 33, 52

[58] A. Shulman-Peleg, R. Nussinov, and H.J. Wolfson. **Recognition of Functional Sites in Protein Structures**. *Journal of Molecular Biology*, **339**(3):607–633, Jun 4 2004. 18, 25, 27, 28, 37, 38, 51

[59] V. Sobolev. **Automated Analysis of Interatomic Contacts in Proteins**. *Bioinformatics*, **15**(4):327–332, 1999. 28, 46

[60] I. Sommer, O. Muller, F.S. Domingues, O. Sander, J. Weickert, and T. Lengauer. **Moment invariants as shape recognition technique for comparing protein binding sites**. *Bioinformatics*, **23**(23):3139, 2007. 37

[61] G. Steinkellner, R. Rader, G.G. Thallinger, C. Kratky, and K. Gruber. **VASCo: computation and visualization of annotated protein surface contacts**. *BMC Bioinformatics*, **10**(32), 2009. 63

[62] G.R. Stockwell and J.M. Thornton. **Conformational diversity of ligands bound to proteins**. *Journal of Molecular Biology*, **356**(4):928–944, Mar 3 2006. 2, 51

[63] S. Subbiah, D.V. Laurents, and M. Levitt. **Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core**. *CUrrent Biology*, **3**:141–141, 1993. 7

[64] A. Via, F. Ferré, B. Brannetti, and M. Helmer Citterich. **Protein Surface Similarities: a Survey of Methods to Describe and Compare Protein Surfaces**. *Cellular and Molecular Life Sciences*, **57**(13):1970–1977, 2000. 18, 37

[65] M. Weisel, E. Proschak, and G. Schneider. **PocketPicker: analysis of ligand binding-sites with shape descriptors**. *Chemistry Central Journal*, **1**:7, 2007. 37

[66] K.A. Xavier and R.C. Willson. **Association and Dissociation Kinetics of Anti-Hen Egg Lysozyme Monoclonal Antibodies HyHEL-5 and HyHEL-10**. *Biophysical Journal*, **74**(4):2036–2045, 1998. 4

[67] H. Yao, D.M. Kristensen, I. Mihalek, M.E. Sowa, C. Shaw, M. Kimmel, L. Kavraki, and O. Lichtarge. **An Accurate, Sensitive, and Scalable Method to Identify Functional Sites in Protein Structures**. *Journal of Molecular Biology*, **326**(1):255–261, 2003. 18, 37