Università degli studi di Milano

Scuola di Dottorato in Medicina Molecolare

Dipartimento di Scienze e Tecnologie Biomediche

Curriculum in Genomica, Proteomica e Tecnologie Correlate

Ciclo XXIII

Anno Accademico 2009/2010

Area disciplinare: scienze mediche, MED/46

Matricola: R07687

Dottorando: Paolo Alessandro COZZI

PROTEIN SURFACE SIMILARITIES EVALUATION FOR FUNCTIONAL ANNOTATION STUDIES

Direttore della Scuola: Prof.ssa Maria Luisa VILLA

Tuture: Prof.ssa Paola COMI

Correlatore: Dott. Luciano MILANESI

SOMMARIO

Uno degli obiettivi più complessi della bioinformatica è quello di provare a predire la funzione di proteine a funzione ignota basandosi sull'identificazione di omologie con proteine a funzione nota. Diversi approcci sono disponibili attualmente: la scelta del più adatto dipende dalla distanza evolutiva che separa la proteina di interesse e la sua proteina omologa. Recentemente l'interesse si è concentrato verso le superfici molecolari in quanto esse non dipendono dal particolare tipo di struttura tridimensionale e permettono quindi di evidenziare delle similarità difficilmente identificabili con altri metodi. Inoltre le superfici molecolari rappresentano l'interfaccia delle interazioni tra molecole, per cui poter descrivere le loro caratteristiche geometriche e fisiche consentirebbe di poter comprendere i processi di riconoscimento molecolare, in quanto la componente geometrica gioca un ruolo fondamentale nella prima fase della formazione di un complesso. Questo aspetto in particolare avrebbe conseguenze rilevanti nel drug-design e nella comprensione degli effetti collaterali dovuti alle interazioni proteina-proteina.

In questa tesi si è sviluppato e ottimizzato un protocollo per l'identificazione di similarità a livello della superfici molecolari. In questo processo le superfici molecolari vengono prima calcolate secondo il modello di Lee Richards, per poi essere rappresentate attraverso mesh triangolari. Successivamente, mediante una tecnica propria della computer vision, le superfici vengono trasformate in una serie di immagini object oriented. Questo tipo di rappresentazione ha il vantaggio di essere indipendente dalla posizione nello spazio degli oggetti rappresentati, per cui superfici simili possono essere descritte da immagini simili. La ricerca di similarità avviene poi andando a individuare correlazioni tra coppie di immagini simili, filtrando le corrispondenze sulla base di criteri geometrici e raggruppando poi tali corrispondenze in gruppi ad alta similarità. Tali gruppi vengono poi utilizzati per reallineare le superfici in modo da poter valutare la qualità dei risultati sia a livello visivo, sia attraverso degli appositi indici. Il processo può essere utilizzato sia nell'ambito dell'annotazione funzionale, attraverso l'identificazione di similarità tra superfici di proteine omologhe, sia nello studio delle interazioni proteina-proteina, individuando delle complementarietà tra superfici di proteine interagenti.

L'intero processo di riconoscimento di similarità dipende dalla configurazione di 15 parametri che bilanciano i tempi necessari per effettuare i calcoli con la qualità dei risultati trovati. L'ottimizzazione dei parametri è stata affrontata mediante l'utilizzo di un algoritmo genetico, in cui le diverse configurazioni di parametri sono rappresentate come una popolazione in cui gli individui in grado di allineare le superfici in modo soddisfacente venivano premiati con un punteggio di fitness più alto. L'efficacia dell'algoritmo è stata poi migliorata dall'introduzione di un'euristica di vicinanza delle corrispondenze che ha permesso di ridurre i tempi di calcolo necessari per il clustering delle correlazioni lungo la superficie.

Particolare attenzione è stata posta nella visualizzazione dei risultati e nella costruzione di indici che potessero quantificare la qualità dei risultati. Per quanto riguarda la visualizzazione, si è implementato un sistema basato sulle librerie del Visualization ToolKit in modo da rappresentare le superfici allineate come oggetti in uno spazio tridimensionale, dando la possibilità all'utente di interagire con la scena rappresentata cambiando il punto di osservazione o ingrandendo dei particolari della scena.

Per quanto riguarda la scelta degli indici con cui valutare i risultati, due indici hanno avuto un ruolo determinante. Il primo indice, denominato overlap, misura la percentuale dei vertici di due superfici allineate che si trovano a distanze minori di 1 A°. Tale indice è utile per stimare la similarità di superfici, in quanto superfici simili ben allineate avranno un gran numero di vertici sovrapposti. Il secondo indice, detto RMSD, valuta lo scarto quadratico medio dei carboni alfa di due superfici allineate nel caso della ricerca di complementarietà. Questo indice permette di valuare quanto la proteina allineata si allontana dalla posizione corretta assunta nel complesso proteina-proteina. Nell'ambito della valutazione dei risultati, si è visto come l'introduzione della valutazione del potenziale elettrostatico permetta di assegnare buoni punteggi in caso di forti similarità geometriche nell'ambito dell'annotazione funzionale, facilitando l'identificazione di superfici effettivamente omologhe.

Il metodo è stato poi validato sia per quanto riguarda la ricerca di similarità che per la ricerca di complementarietà. Per quanto riguarda la ricerca di similarità si è analizzato un campione di 13 proteine con domini prosite noti per verificare se l'algoritmo era in grado di reidentificare sulla superficie la presenza di tali domini. Per fare questo, si è per prima cosa ridotto il numero di strutture della Protein Data Bank ad un gruppo di strutture rappresentative. Si è calcolata la superficie molecolare di ciascuna proteina rappresentativa, per poi realizzare un dataset ricavando la superficie molecolare in corrispondenza dei domini funzionali prosite. Il test è stato poi eseguito cercando di allineare le superfici delle 13 proteine note con il dataset delle patches dei domini funzionali. I risultati hanno dimostrato che nella maggior parte dei casi, l'algoritmo è riuscito ad allineare correttamente un dominio funzionale ad una superficie che presentasse lo stesso dominio, e che questa evidenza fosse facilmente identificabile sia attraverso i parametri utilizzati per valutare i risultati, sia ispezionando visivamente i risultati dell'allineamento.

Il metodo è stato successivamente testato per la ricerca di complementarietà, cercando di ricostruire i complessi proteina-proteina presenti in un dataset utilizzato per validare i metodi di docking. Mentre nel caso di ricerca similarità è importante descrivere le superfici in modo dettagliato in modo da aumentare la specificità del metodo, nel caso di ricerca di complementarietà un'alta precisione è controproducente, in quanto l'interazione tra proteine non è determinata solo da caretteristiche geometriche ma comporta anche la formazione di interazioni elettrostatiche favorevoli e riarrangiamenti della catene laterali. Per questo motivo, le superfici sono state calcolate con un modello smooth, per cui dei dettagli di rappresentazione vengono persi a favore di una maggiore similarità tra superfici in grado in interagire. I risultati hanno dimostrato che l'algoritmo è in grado di allineare i complessi con risultati paragonabili rispetto ad alcuni programmi attualmente a disposizione. Per come è stato scritto il disegno sperimentale e per via del fatto che il metodo non tiene ancora conto di considerazioni di tipo energetico, il risultato ottenuto è particolarmente interessante anche perchè il metodo proposto fornisce un insieme di conformazioni più ampio di quelli proposti dagli altri algoritmi, su cui poi è possibile estendere le analisi per identificare una migliore predizione.

In conclusione, il sistema proposto è in grado di identificare similarità a livello della superficie molecolare attraverso l'analisi di immagini di descrizione locale. I risultati ottenuti dimostrano che il programma è efficace nell'identificare superfici simili nel contesto dell'annotazione funzionale. Per quanto riguarda la ricerca di *complementarietà il programma ha interessanti prospettive, anche se i complessi proposti come migliori non sono sempre biologicamente corretti. Da questo punto di vista, si dovranno compiere ulteriori analisi per migliorare il metodo in modo da essere informativo nel campo dello studio di interazioni tra proteine.*

ABSTRACT

One of the main targets of bioinformatics is to assign functions to proteins whose function is unknown relying on homologies identifications with proteins with known functions. Several approaches are currently available: the best choice depends on the evolutionary distance that separates the protein of interest from its homologous. Recently attention has been focused on molecular surfaces since they do not depend on the three-dimensional structure and allow similarities to be identified which other methods can't identify. Furthermore, molecular surfaces are the interface of interaction between molecules, and their geometrical and physical descriptions will lead to the comprehension of the molecular recognition process, since the geometrical component has a fundamental role in the early stage of complex formation. This particular aspect would have a major impact in the field of drug design and in the understanding of the side effects due to interactions between proteins.

During this thesis a protocol for similarities identification on molecular surfaces has been developed and optimized. In this process, molecular surfaces are calculated according to Lee Richard's model, and then are represented through triangular meshes. Successively surfaces are transformed into a set of object oriented images using a computer vision approach. This type of representation has the advantage of being independent from the position of the objects represented, and thus similar surfaces can be described by similar images. The search for similarities is then performed by indentifying correspondences between pairs of similar images, by filtering matches relying on geometrical criteria and then by clustering correspondences in high similarity groups. These groups are then used to align surfaces in order to evaluate results both by visual inspection and through appropriate indexes. This process can be applied in the field of functional annotation, through the identification of similarities between surfaces of homologous proteins, and in study of interaction between proteins, through the identification of complementary areas between interacting proteins.

The whole process of similarities detection depends on the configuration of 15 parameters that balance the time needed to perform calculation with the quality of results found. The problem of parameters estimation has been addressed using an implementation of genetic algorithm, which allowed representing different configuration parameters as a population in which individuals that are able to align surfaces satisfactory are rewarded with an high fitness score. The effectiveness of the algorithm was then improved by the introduction of neighbor heuristic which reduced the computational time required for correspondence clustering on surfaces.

Particular interest was placed in results displaying and in the construction of indices that can quantify the quality of results. Regarding the visualization problem, a display system was implemented based on the Visualization ToolKit libraries in order to represent surfaces aligned as objects in three-dimensional space, enabling the user to interact with the scene represented by changing the point of view or enlarging details of the scene represented.

Regarding the definition of useful indexes for results evaluation, two indexes had a fundamental role. The first one, called overlap index, measures the percentage of vertices of two surfaces that are closer than 1 A° after the alignment. This index in particular is useful for evaluating the surface similarity since similar aligned

surfaces will have a large number of vertices closer than this distance. The second index, called RMSD, is important because it evaluates the Root Mean Square Deviation of alpha carbons of two aligned proteins in the case of a complementary search. This index allows evaluating how the aligned protein is distant from the correct position in the crystal complex. Concerning results evaluation, we have noticed that the consideration of electrostatic potential allows assigning good scores in case of strong geometrical similarity in context of functional annotations, thus facilitating the identification of homologous surfaces.

This method has been validated both in the search of similarities and in the search of complementarities. Regarding the search of similarities, we tried to analyze a sample of 13 known proteins with a prosite domain in order to identify the presence of such domains on molecular surfaces. For doing this, we first reduced the number of structures present in the Protein Data Bank to a group of representative structures. Then we calculated the molecular surfaces for each representative protein and we created a dataset of patches corresponding to the prosite functional domain. The test was then performed trying to align the surface of the 13 known proteins to the patches dataset of functional domains. The results showed that in most cases we are able to properly align a functional domain to a protein surface with the same functional domain, and that these evidence was easily identifiable both by the parameters used for results evaluations, both by visually inspecting the results of the alignments.

The method was then tested for complementary research, trying to reconstruct the protein-protein complex present in a well known dataset used to validate docking methods. In the case of searching for similarities it is important to describe surfaces in details in order to increase the accuracy, but high precision when searching for complementarity is counterproductive, since the interaction between proteins is not only determined by geometrical features but also involves the formation of favorable electrostatic interactions and rearrangements of side chains. Thus molecular surfaces were calculated using smoothed surfaces, where most details are lost but allowing to detect more easily interacting surfaces. Results showed that the algorithm is able to align complexes with comparable scores than the programs currently available; Considering this experimental design and that the method does not take into account the electrostatic potential, we can assume that the results obtained are particularly interesting since the proposed method provides a wider set of conformations than other algorithms, upon which we can extend the analysis in order to identify a better prediction.

In conclusions the proposed system is able to identify similarities on molecular surfaces through the analysis of images of local description. The results show that the system implemented is effective in identifying similar surface areas in the context of functional annotation. In regards to the search for complementarities, the algorithm seems to have an interesting perspective, even though the best complex proposed is not always biologically correct. From this point of view, we have to do more analysis in order to improve the methods in protein interaction studies.

INDEX

1 INTRODUCTION

1.1 The genomics challenge

With the refinement of experimental techniques and the exponential increase in the number of sequences identified by genomics projects, we are facing a lot of data from which information has to be investigated or assigned. Where are genes and what do genes do? Annotation of a genome involves assigning functions to gene products, but it would be unthinkable to perform a functional assay for every uncharacterized gene in every genome. Moreover, it is impossible to keep up with the influx of data by manually curated annotation. For such reasons, scientists have been turning to sophisticated computational methods for assistance in annotating the huge volume of sequence and structure data being produced [1].

Even the same concept of function is difficult to define. Proteins perform the most important tasks in organisms, such as catalysis of biochemical reactions, transport of nutrients, and the recognition and transmission of signals. All of the aspects associated with a particular protein are referred to as its “function”. However, protein function is not a well-defined term; instead, function is a complex phenomenon that is associated with many mutually over-lapping levels: biochemical, cellular, organism-mediated, developmental and physiological [2]. Consider, for example, a protein kinase: it can be related to different cellular functions (such as cell cycle) and to a chemical function (transferase). The same kinase may also “misfunction”, thereby causing disease. Moreover the same kinase doesn’t act alone in vivo, but may be part of a signaling pathway, where this protein both phosphorylates, and is phosphorylated by interacting partners. We have to say that biological function is an interpretation and not a property of objects, and it is strongly related to the context in which the molecules are [3].

Moreover, when assigning functions to proteins there are problems: the annotation of a protein is written in a human language and this includes the function and the experimental evidence which supports it, the research history, the group that carried out the annotation and other characteristics. Furthermore, as in a human language, many terms are synonymous. This implies much more difficulty in designing applications that are able to do automatic annotation [1]. For such reasons a controlled vocabulary has to be defined: a remarkable solution has been proposed in the ideation of the Gene Ontology Project. The Gene Ontology Project [4] is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data. This project is a major step forward in the standardization of functional annotation, however it is necessary to describe protein function only through defined ontological terms, trying to make the most accurate description.

All these aspects complicate the problem of functional annotation and must be taken into consideration both by those who do functional annotation, and by those who create the instruments for doing functional annotation. In this thesis lies a methodology that can provide useful information in the field of functional annotation. The aim of the work is not to achieve a comprehensive approach in order to replace existing methodologies, but to support functional annotation in identifying structural evidence, since protein function depends on amino-acids disposition in a three-dimensional space.

This chapter describes an overview of methods for making functional annotation. All these methods are based on the identification of pieces of evidence which can be similarities between structures or local characteristics that identify well-defined regions on proteins which are successfully used in the field of structural genomics to assign function to unknown proteins.

1.2 Structural genomics

Structural genomics seeks to describe the three-dimensional structure of each protein, since structure is closely linked with protein function. Understanding protein functions has great implications in understanding the mechanism of diseases and at the same time in identifying potential targets for drug design. To accomplish this task, structural genomics involves taking a large number of approaches to structure determination, including experimental methods or modeling-based approaches based on sequence or structural homology with a protein of known structure.

However, it is not convenient to determine a three-dimensional structure of each protein, since proteins with a high degree of similarity will present identical structures. Indeed, the aim of structural genomics is to create a representative set of experimental structures, in order to obtain all the remaining structures through a combination of experimental and modeling approaches. This has been carried on by improvements in target selection, in order to choose structure with non structural similarity with other proteins present in the Protein Data Bank [5]. On the other hand, a significant proportion of proteins structures are proteins of unknown function, annotated merely as “hypothetical proteins”, since they don’t have any close similarity to proteins of known function. Although these provide a valuable contribution to our knowledge of protein structure, their worth can be significantly enhanced by knowing the biological roles that they play in cell.

In recent years, a lot of effort has gone into the development of methods for deriving functional clues directly from the three-dimensional structure relying on fold similarities and on amino-acid around an active site, but the biochemical function of most hypothetical proteins remains unclear. This is because either the folds are completely new or the folds are so widely distributed in so many protein families that it is difficult to identify their functions [6]. Moreover, it is also useful to try and re-annotate annotated structures: the identification of new elements in already known structures can provide further explanations to the molecular mechanism as well as useful information for drug design.

1.3 Homology identification supports functional assignment

In the absence of experimental data, the function of a gene is usually inferred by establishing homologies with well-annotated genes. In biology, two traits can be considered homologous when they have a common evolutionary origin. According to this hypothesis, two genes derived from the same ancestor may present different mutations but may have preserved the same function. For this reason, finding similarities between two proteins may support the assignment of homology. In such a way, a series of methods relying on finding homology by sequence similarity has been developed. In fact, the more similar the sequence, the more similar the function is likely to be, although there are cases of proteins having 100% sequence identity which perform different functions according to where they are expressed [7],[8].

When sequence similarity is low, protein structure represents a powerful means of discovering function, because structure is well conserved over evolutionary time, and it provides the opportunity to recognize homology that is undetectable by sequence comparison. If the protein has a known fold, then that protein may have a function similar or identical to other proteins with the same fold [9]. Moreover, the structure of a protein is much more informative than the amino acid sequence alone: knowing the structure allows us to explain the biochemical mechanism of the protein of interest.

However, sometimes the function of one or both proteins may change during evolution while their folds remain largely the same [10]. In these cases functional differences might be more obvious from structural comparison of functional sites than they would be from comparison of sequences or overall tertiary structure [11]. Considering this, one might ask why not assign function by comparing directly functional sites, bypassing any problems related to sequence and fold similarity. We have to make some considerations: first of all we don’t always have a three dimensional structure of the protein of interest. Despite the fact that in recent years we have improved techniques of crystallization, the Protein Data Bank [5] contains about 70,000 structures, and the aim of the structural genomics project is not to resolve all protein structures but to identify and resolve only a representative set. Moreover, we often deal with sequences that have a high degree of similarity with previously annotated sequences, so sequence based approaches are reliable. Only when the sequence of interest doesn’t have any close sequence similarity to proteins of known function, or when the fold is similar to many proteins with different functions we have to use methods for deriving functional clues directly from three-dimensional structure (if we have one). Thus, the more complicated the hypothesis, the higher the possibility of making mistakes.

1.3.1 Sequence based approach

The most used approach for doing functional annotation is to search in sequence databases using sequence similarity tools such as BLAST [12]. The aim is to find a significant sequence similarity to any other sequence of which its function has been experimentally characterized. Nevertheless even with a high sequence similarity annotation transfer may be erroneous, for example enzymatic function may not be conserved since it specifically may depend on a few amino acids [13]. A related form of erroneous annotation transfer is due to domain shuffling: during evolution functional domains have rearranged and mutated in different sequential location creating proteins with different function from the same “building blocks” [14]. Errors in annotations can be caused when database hits with significant e-values occur, even though the query and hit sequence may have a different overall domain structure.

As a consequence of the explosion of the number of sequences identified by genomic projects we are faced with more new and different sequences and sequence similarity tools are less effective since the number of sequences with no annotation is rapidly growing [1]. For example, at the moment pFam [15] contains about 11,000 family domains, out of which about 5,000 domains have no annotation. Another consideration is the propagation of incorrect annotations: a single error could propagate to any other sequence with an high degree of sequence similarity [16]. Even when dealing with sequences with high similarity, the evolutionary information has to be considered in order to discriminate paralogue genes (genes related by duplication within a genome often with different functions, even if these are related to the original one) from orthologue genes (genes in different species that evolved from a common ancestral gene with the same function).

1.3.2 Sequence pattern

Proteins that share a common function but are otherwise diverse will usually share one or more common sequence or structure patterns necessary to maintain their structure and function. This is because proteins perform their functions using a relatively small part of their structure. Therefore, to predict a protein's function, in many cases there is no need to use annotation transfer from a homologous protein. All that is required is to identify a sequence- or a signature or feature that can be associated with a function. For example, the PROSITE [17] database consists of a large collection of biological signatures described as pattern or profiles. Finding a signature in the protein of interest could help in the assignment of function. Alternatively profile methods like pFam [15] can identify remote homologies. The advantage of those methods are that they provide greater sensitivity compared to simple sequence comparison because the profiles implicitly contain information on both which residues within the family are well conserved and which are the most variable [8].

1.3.3 Structure alignments

When sequence-based methods fail to identify function, useful information can be derived from three-dimensional structures, if available, because they define the biochemical mechanism by which protein implements its functionality and they are much better preserved than sequences. Many proteins with little or no sequence similarity still have a structural similarity [18]. Methods for predicting function from structure can be classified according to the level of protein structure and specificity at which they operate, ranging from analysis of the protein's overall fold to the identification of highly specific three-dimensional clusters of functional residues [8]. The fold based methods attempt to establish an equivalence between two or more polymers relying on their three dimensional conformation. Those equivalences are established after aligning three-dimensional structures: the query protein is aligned towards a dataset of known structures and the attribution of functionalities is effected relying on fold similarity. Examples of tools for doing structure similarity are DALI server [19] or CATH server [20]. A good review of structure similarity methods can be found in [21].

However, caution should be used in assigning homology relying on fold similarity: proteins sharing similar functions often have similar folds, but sometimes the function of one or both proteins may alter during evolution while their folds remain largely unchanged. In such cases the same fold may give rise to two functions [8]. Moreover, proteins can have a common tertiary structure as a consequence of convergent evolution. Finally, analysis carried out on structural databases shows that proteins tend to take a limited number of folds [20]. For example, the TIM barrel fold supports more than one function [10]. This could be due to an experimental bias in the choice of proteins to be crystallized or to crystallization experimental conditions, but this essentially means that finding a fold match between the target protein and one in the PDB does not always provide a reliable prediction of the protein's function: it can only suggest a possible function type [8].

1.3.4 **Structure patterns**

When structure alignment fails in assigning function unambiguously, or when the protein has a novel fold or a low similarity to a known fold it is possible to obtain information by analyzing structure patterns of the protein. The rationale for structure patterns is the same as for sequence-based pattern: identifying unique markers associated with a function [1]. For example considering serine proteases: the class of the enzyme is due to a few amino-acids in a specific geometric conformation regardless of the fold assumed [22]. Moreover functional differences might be more obvious from the structural comparison of functional sites than they would be from comparison of sequences or overall tertiary structure. For example functional annotation of hypothetical protein YBL036C through structural similarity alone would bring several different conclusions. This protein has a fold typical of many proteins, known as Tim Barrell [23]. Through research with the FSSP database is possible to show a high structural similarity in two different classes of enzymes, alanine racemase (EC: 5.1.1.1) and ornithine decarboxylase (EC 4.1.1.17). Subsequent analysis allowed assigning the function of this protein to alanine racemase on the basis of similarity to the catalytic site [24].

Structure patterns are best described as identifiable spatial regions within the protein's structure, which leaves much more room for descriptive methods. Those range from identification of 3D shapes completely dissociated from the amino acids, to a string of characters representing amino acids and their physical environment [1].

1.3.4.1 Residue template methods

The function of certain types of proteins is due to a small number of residues found in a localized region of the three-dimensional structure. For example, in enzymes the catalytic function is performed by a small "constellation" of residues located in the active site, while residues responsible for binding the substrate are not as vital to the catalytic function of the enzyme and can change through evolution, sometimes allowing the enzyme to accommodate new substrates [25]. For DNA-binding proteins instead, the residues located on the surface may be responsible for the specificity of binding to a particular sequence of DNA [26]. Often, the specific arrangement and conformation of catalytic sites are crucial to the function and remain strongly conserved over evolutionary time, even the protein's sequence and structure undergo major changes [8].

Residue template methods depend on identifying and compiling the conformations of the crucial residues, and scanning them against any novel protein structure. Database of templates can be created manually or generated automatically. For example Catalytic Site Atlas [25] contains two dataset of templates: the first one is hand annotated and includes only residues which are directly involved in catalytic reaction. The second one is derived by homology with the first dataset. Another example is PDBSiteScan [27] in which we can search for functional sites by pair wise structural comparison of a protein with a collection of functional sites obtained analyzing the SITE records of Protein Data Bank [5] files. Another interesting example is SuMo [28] in which protein structure is described as a set of stereochemical groups that are defined independently from the notion of amino acid and similarities are found using graphs of similar chemical groups.

In general, finding a match to one of these functional templates may suggest the protein’s function. However, we have to take care to ensure that the match is correct, since these methods are particularly prone to return false positive matches [8].

1.3.4.2 Protein clefts and binding pockets

Methods that recognize patterns of residues that are conserved in their 3D positions and in their amino acid identities are not always applicable. There are also biological examples of proteins that can bind the same binding partners without sharing any conserved patterns of amino acid residues [29]. In such cases, analysis of clefts and pockets localized on protein surfaces can help us to understand many things about proteins functionalities. For example, cofactors, substrates and regulatory elements tend to bind in clefts on the surface or in a region between separate interacting protein chains, preferably in one of the two largest clefts on the surface [30]. Moreover the size of surface clefts can be small or large, depending on the geometry of binding pockets, whereas methods based on a three-dimensional template are often limited to the number of residues that can be included. The last consideration about these methods is that similar binding pockets can also be found in protein with different sequences and folds. For such reasons analysis and comparisons of annotated clefts can be more informative than template based methods.

There are some useful tools for analyzing molecular cavities and binding sites such as the web-server pevoSOAR [31], which provides an online resource to identify similar surface clefts on protein structures taking into account evolutionary information specific to the binding surfaces. The web-server SMAP-WS [32] is designed for the comparison and similarity search of binding pockets, Whereas SiteEngine [33] has an efficient algorithm for finding similar binding sites on proteins by describing relevant regions with their physic-chemical properties.

Limitations of these methods arise when binding sites undergo significant conformational change upon substrate binding; another issue is when proteins undergo significant conformational change due to allosteric control: in such cases the ligand-free form of proteins will have a very different site that may be difficult to even recognize as the binding site [8].

1.4 Functional site identification

When it’s not possible to identify similarities between well known proteins and the protein under examination, local characteristics can be identified on protein structures in order to support the following phases of annotation. In recent years, substantial effort has been directed towards characterizing the properties that distinguish the parts of a protein that are involved in molecular recognition [34],[35],[36]. The reasons behind this are twofold. Firstly, there is the scientific goal of understanding the physical principles that underlie the exquisite molecular-recognition processes that permit fidelity in processes such as signal transduction. Secondly, there is the more applied goal of using structures to contribute to the functional annotation of genomic projects. In this context the identification of ligand binding sites and interaction sites allows the assignment of function to unknown proteins since these sites mediate the ability of proteins to recognize molecules for transport, signal transduction and catalysis.

Identifying the binding sites means understanding the biological role of a protein, and consequently may be useful for developing new drugs. As previously said, enzyme active sites are located in large and deep clefts on protein structure, and the need for significant favorable interactions between ligands and proteins usually means that this bind occurs in surface depressions. On the other hand, protein-protein interactions typically involve large, accessible and mostly planar sites in which the solvation potential, interface propensities and protrusion of residues cannot be easily distinguished from the rest of the protein's surface [37].

Some interesting methods have emerged in order to distinguish these two different types of molecular regions: methods based on evolutionary information, methods based on physicochemical features and approaches based on correlated mutations.

1.4.1 **Evolutionary trace methods**

Active-site and ligand-binding site residues are more conserved than general surface residues across many different protein families [38]. This result is perhaps not surprising considering that the precise arrangement of residues required for catalysis and ligand binding is expected to impose strong constraints on the evolution of sequences and structures. Furthermore, predictive studies have shown that clusters of residues that make up active sites or ligand-binding sites are invariably more conserved than clusters of residues defined elsewhere on the surface of a protein. These results show that conservation analysis is of predictive value in the identification of active sites and ligand-binding sites [39].

The evolutionary trace method is powerful enough to identify significant functional site residues using phylogenetic information to rank the residues according to functional importance and map them onto the 3D structure. This technique is based on two assumptions: the first one is that functional sites evolve through variations of a conserved architecture. From this hypothesis, while residues that are important to preserve the architecture will be mostly invariant, the residues important for functional specificity can undergo many substitutions, each associated with a functional variation. The second assumption identifies residues whose variations correlate with changes in function using sequence identity trees. According to this assumption, sequences with greater identities have diverged more recently than sequences with lesser similarity, and therefore have had less time to functionally diverge: residues that are important for functionalities will display a high level of conservation, while residues that are important for the specificity will only be conserved to sequences evolutionarily related [40].

1.4.2 **Physicochemical approaches**

Physicochemical approaches assume that functional sites have physical and chemical features which are different from the rest of the protein. Several mechanisms, such as polar complementarity or reduction of steric and electrostatic strain contribute to protein stabilization, while active sites have destabilizing properties. For example hydrophobicity can be statistically demonstrated to be preferentially expressed at protein–protein interfaces[35]. This aspect can be used for functional site prediction. Electro-statically unfavorable residues could be a putative functional site [41] and direct evaluation of the stability change caused by point mutations, combined with other structural features that differentiate the active site from a non-active site, would provide a good prediction of possible active sites [42].

Whereas many biologically relevant protein–protein interactions derive their affinity from the burial of the hydrophobic surface, electrostatics have been shown to play a key role in determining specificity and, in some cases, the thermodynamics and kinetics of macromolecular association [43]. Calculation of the electrostatic potential of protein-protein complexes has led to the general assertion that protein-protein interfaces display “charge complementarity" and “electrostatic complementarity" [44]. Even this information helps in understanding the mechanism by which proteins act.

1.4.3 **Correlated mutations**

If functional sites are preserved because they have to maintain their specificity to perform their functional activities the role of conservation is less clear for protein–protein interfaces [38]. The generally accepted model for the variation of the rate of evolution of proteins is one in which the rate of evolution increases (i.e. conservation decreases) from the catalytic site to the protein core, to substrate-specificity sites and finally to surface regulatory regions.

Genome wide analyses have demonstrated that interacting protein pairs tend to evolve in a correlated manner. More precisely, in heterodimers the interface is more conserved than the rest of the molecular surface and in homodimers the interface is less conserved. Irrespective of which factor contributes the most significantly, these results may be rationalized in terms of the comparison between the conceivably more complicated co-evolutions of two independent proteins that are constrained to maintain an heterodimeric interface and the conceivably simpler evolution of a single sequence constrained to maintain the homodimeric interface [39]. Considering this, the identification of correlated mutations should suggest some information about the location of protein-protein interfaces. An interesting approach analyzes correlated mutations between pairs of residues in multiple sequence alignments of two proteins [45]. Although this method can’t identify the pair of proteins that are likely to interact, it has the potential to identify interaction sites and also interacting residues [46].

1.5 Molecular surfaces

Despite the effectiveness of methods for identifying biological information, many functional sites are not concave and fold based methods or three dimensional patterns identifications cannot identify them. Moreover some strategies for rational drug design require the study of functional domain outside the largest cavity. For example, the design of inhibitors of protein-protein interactions requires the study of binding sites that do not normally bind small molecules [47]. Furthermore an approach that doesn’t involve the creation of three-dimensional templates will be useful to identify functional sites which are not characterized by patterns, such as interaction domains or non-catalytic domains.

Recently, interest has been focused on molecular surfaces, where most protein functions occur. More precisely, surface geometry [48] and electrostatic properties [49] are considered to be essential for molecular recognition. Proteins utilize common surface motifs to create precise chemical environments designed to perform specific functions. These motifs are not restricted to a single protein scaffold but can be found within different protein folds or at domain/domain and subunits interfaces. While biochemical activity can be attributed to a few key residues (e.g catalytic triads), the broader surrounding environment (i.e. auxiliary residues in spatial proximity) often plays an equally important role in fine-tuning molecular recognition and/or catalysis. Moreover molecular surfaces are less sensitive than the corresponding atom disposition, in that such analysis can detect convergent evolution of functional sites.

In many cases, the identification of similarity in binding sites can suggest novel relationships that would go undetected when using traditional sequence or backbone structure comparisons [50]. Such analysis can be useful because recognition of similarity in the binding pattern of a well-known protein may help in gaining a better understanding of its function and activation mechanism. Furthermore proteins with similar binding sites may bind similar drugs and this can provide an explanation in the understanding of side effects. These considerations are crucial for the development of targeted drug leads like inhibitors.

Another difficult task of proteomics is to understand protein interactions, since molecular mechanisms involve systemic effects produced by interactions between proteins with different functions. In this context the study of molecular surfaces can be interesting because molecular surfaces are involved in molecular recognition [49], and their description can shed light on how proteins interact and how molecular processes take place.

However, these approaches have an inherent limitation caused by the dynamic nature of proteins. For example, upon ligand binding the functional site or also the entire molecule may be subjected to conformational change. For such reasons, molecular surfaces should be described through a method able to handle different conformations of side chains. Moreover, similarity searches of this nature can be very expensive computationally: for this reason there is a need to balance surface description and the choice of fast algorithms to identify surface similarities. Finally these methodologies are unsuitable when the functional site is determined by the interaction of multiple molecules. Despite this, surface methods can provide useful information for functional annotation undetectable by other methods.

1.5.1 **Molecular surface definition**

Molecular surfaces can be described as the outer part of volume occupied by the molecule. An easy way for representing molecular surfaces is obtained by using *Van der Waals* representation, in which atoms are represented as spheres whose size is determined by the radius of *Van der Waals*. The molecular surface is obtained by the union of such spheres. However, this type of representation creates very small regions that cannot be occupied by any other molecule or atom. Those regions must be considered as a part of the molecular surface if the salvation is taken into account. A different model was proposed by *Lee and Richards* in 1971 [51], by which molecular surface is obtained by rolling a water molecule on the *Van der Waals* representation of the protein. They distinguish two different types of molecular surfaces: the *solvent excluded surface (SES)* is the surface obtained by the contact of the probe with the molecule, while the *solvent accessible surface (SAS)* is obtained by the trajectory of the center of the probe along the protein (Figure 1).

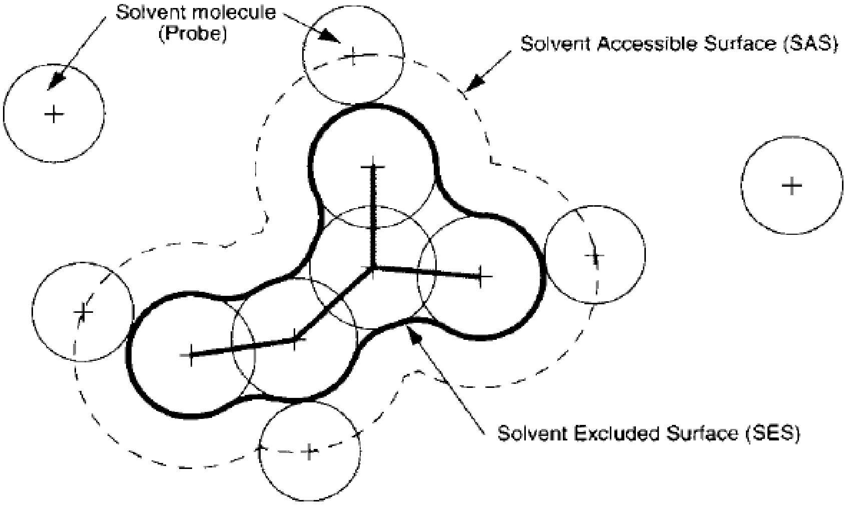

Solvent molecule
(Probe)
Solvent Accessible Surface (SAS)
Solvent Excluded Surface (SES)

Figure 1. This figure represents the solvent accessible surface and the solvent excluded surface.

Molecular surface can be described in an analytical way by the union of different objects (spheres, tori and circular arcs) and in such a way it's possible to generate high definition surfaces. The first computational representation of molecular surface was carried out by Connolly in 1983 [52]. His algorithm was able to distinguish atoms accessible to the solvent from atoms buried by the protein. Rolling the water probe on the molecule, three types of geometrical figures are generated: concave spherical triangles, saddle-shaped rectangles and convex spherical regions (Figure 2).

Once the surface is calculated in an analytical way, those objects are represented by computer programs with triangles (*mesh*) obtained by deriving points from concentric areas on geometrical objects. This type of surface representation is also

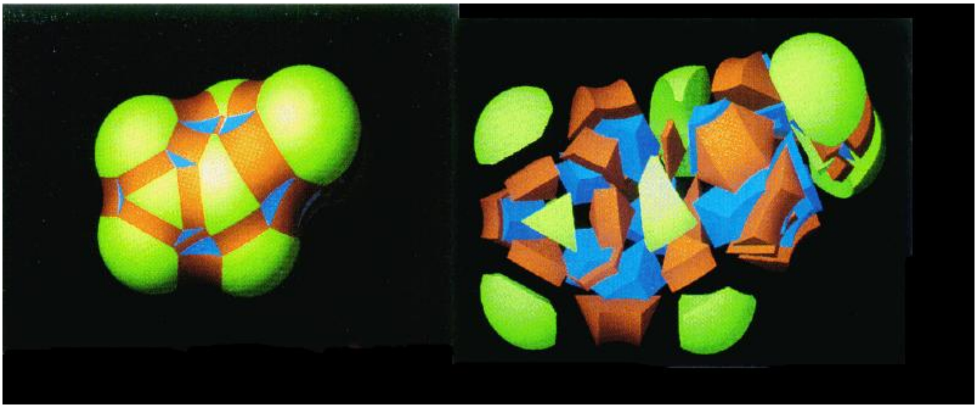


Figure 2. Molecular surfaces are determined by the union of 3D objects.

referred to as "*Connolly representation*", since Connolly was the first who implemented an analytic description of molecular surfaces. From this model, it is possible to study the geometrical features of proteins, as might be the presence of cavities or protruding parts. Nowadays there are many tools for calculating molecular surface, each with its pros and cons, but each one is based on the same analytical concepts.

Among the two different types of surface representation, the *solvent exclude surface* is preferably used for representing molecules and for displaying properties such as atomic charge, electrostatic potential and hydrophobicity [53]. This information is displayed on surface using textures and color maps, such as the representation of electrostatic potential in which the red areas (where the charge is negative) are interpreted as areas which will most favorably interact with blues areas (where the charge is positive) (Figure 3).

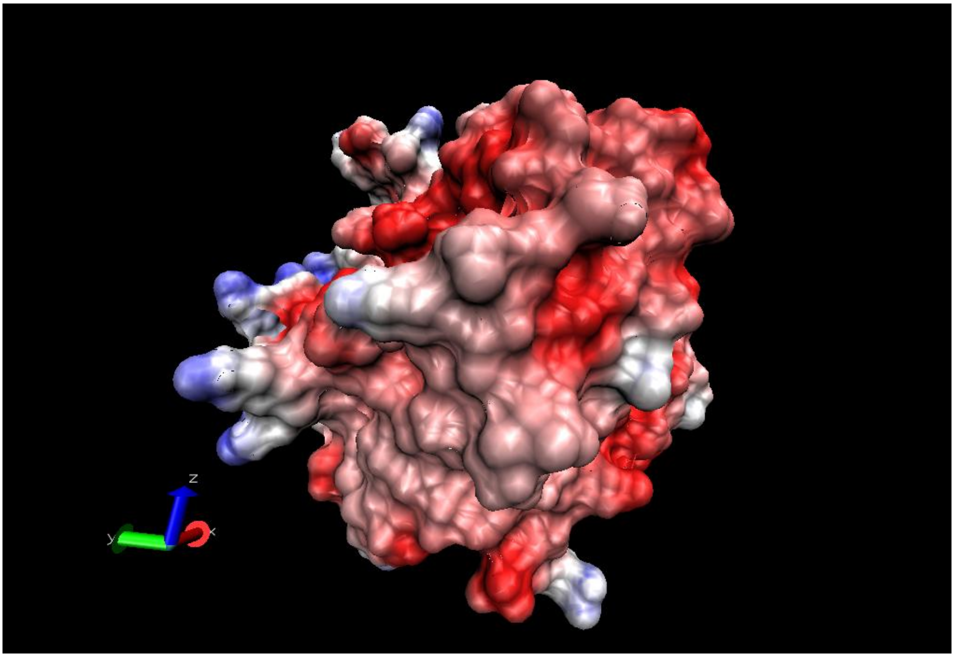


Figure 3. The electrostatic potential of protein C mapped on molecular surface

1.5.2 Surface similarities detection

Since protein surfaces are critically involved in selective binding, recognition and interaction with molecular partners, methods for surface comparison may give new insights into protein function analysis. For such reason in recent years several attempts were made in order to find surface similarities. Ideally, in searching for protein surface similarity, one would like to be able to select similar surfaces even when analyzing proteins crystallized at different resolutions, solved with different methods, in different experimental conditions and sometimes also crystallized both in their bound and unbound states. The ability to describe a protein surface without too many details is therefore also precious in protein surface similarity searches. However the choice of the surface representation method is crucial since the surface description must be sufficiently detailed in order to recognize different binding sites, but must not affect the performance of the surface comparison algorithms. Some attempts to identify similar surfaces avoid describing molecular surfaces by the Connolly representation, for example describing molecular surface with *alpha-shapes* [54],[55] or by describing molecular surfaces with the atoms facing on surfaces [56]. Another interesting approach describes surfaces by alpha carbon [57]. All of these methods work well but they have the disadvantage of not being able to describe the geometric properties of molecular surfaces. More precisely these methods don’t take into account all possible variations of the conformations of the side chains of amino-acids exposed to the solvent. Indeed, these regions have great mobility and a different orientation of side chains within a cavity can alter its volume and therefore the interaction with a substrate [58]. For this reason, a more detailed description could better describe the characteristic of a certain molecular region.

An interesting solution may be the use of *sparse critical points*: by this solution molecular surfaces are described first by the Connolly representation, and then the number of surface points is reduced by finding a point location defined by a projection of a gravity center on the molecular surface. This type of representation was successful applied in recognizing the catalytic triad Ser-His-Asp [59] or in identifying similar binding sites on proteins [60]. However, this solution seems to be computationally efficient, the low number of points may lead to errors since small changes in atoms dispositions generate different point locations.

Other approaches describe molecular surfaces in an analytical way by triangular mesh. In this case the surface representation is deliberately detailed since a simpler description has negative effects on the recognition of similarities because two similar molecular surfaces are represented by different sets of points. If surfaces are described through details, there will be small differences between similar molecular surfaces, and they would more likely be recognized as such. This approach involves creating a large amount of data and recognition performance is only acceptable when having a lot of computing power and deciding in advance which regions of proteins to represent [61]. Such methods have been successfully applied in the detection of DNA binding proteins [61] and functional sites [62].

Another interesting solution can be the description of molecular surface with three dimensional Zernike descriptors (3DZD), by which each molecular surface is described by a vector of descriptors [63] and comparison times for a query protein against the entire PDB take, on an average, only a couple of seconds [64]. Despite the effectiveness of the method in computational terms, this solution seems to describe molecular surfaces in a less detailed way. Moreover, the interest in a method to compare molecular surfaces is not only to compare the entire structures, which can be done using fold based methods, but also to identify similar regions in proteins with different folds.

The last solution cited in this chapter seeks to identify surface similarities with images of local description through a procedure very similar to our strategy [65]. In our opinion, even this description relies on the whole molecular surfaces, making it difficult to recognize small regions, such as functional sites.

1.6 Protein-protein docking

In the context of functional annotation understanding how a pair or proteins might associate is important, since it helps to better elucidate the role of the protein of interest. Indeed protein-protein interactions are integral to many mechanisms of cellular control, including protein localization, competitive inhibition, allosteric regulation, gene regulation and signal transduction. Disruptions of protein–protein interactions can cause biochemical diseases related to these functions. The prediction of protein–protein interactions, if accurate and sufficiently consistent, can greatly increase the amount of structural information available to understand the function of biologically important complexes. This is often referred to as "the docking problem". Several crystallographic structures of protein complexes have now been determined, and these frequently exhibit high degrees of steric and chemical complementarity at the protein-protein interface. Docking algorithms are then developed and refined according to their ability to reproduce these known structures in order to predict the structure of a complex when only the unbound structures of its constituents are known in advance. The hypothesis underlying docking predictions is that the structure of a complex is the lowest free energy state accessible to the system. Thus, docking requires the ability to sample many states and accurately determine a free energy. Resolving the problem of protein-protein docking could have considerable implications in structure based drug design.

However, this problem is generally hard to address computationally, even not considering the presence of the solvent, mainly due to the large number of atoms and the involved degrees of freedom. It is possible to use molecular mechanics to refine the hypothesized complex, but it is unfeasible to apply these functions since the computational load is too high, and the calculation of a correlation function that assesses the degree of overlap and penetration upon relative shifts of the macromolecules in three dimensions is computationally very intensive. An important paper from Camacho argues that from the physical point of view, macromolecular interactions occur in two different stages, in particular when proteins are involved: There is a first stage of molecular recognition, where molecules diffuse near each other until the interface patches come sufficiently close. Then the binding stage begins, when high affinity atomic interactions are formed by modification of the side-chain and backbone conformations of the macromolecule [66]. For this reason most protein-protein docking algorithms assume rigid bodies, hence limiting the problem to the search for the most favorable roto-translation of one protein towards the other [67]. Even in this case, it is not uncommon for docking methods to test hundreds or thousands of distinct rigid-body relative orientations. A solution to this problem is to perform an initial analysis of the molecules in order to locate significant geometric features, such as cavities or knobs. The basic idea is that there is a kind of complementarity between the interacting molecules. This complementarity should be both geometric and electrostatic [39]. In this context the study of molecular surfaces can be interesting because molecular surfaces can describe the geometry of the protein and their description can shed light on how proteins interact and how molecular processes take place. However, proteins are flexible, and rigid methods have been accompanied by methods that consider a kind of flexibility at the surface of the protein.

1.6.1 **Rigid docking**

Some previous works are available in relation to the possibility of matching surfaces looking for complementarities as the first step of a docking procedure. Here we'll take a look at three different methods of performing rigid body docking. A famous approach relies on the computation of the Fast Fourier Transform in three dimensions for predicting complementarities in possible complexes between macromolecules of known structures [68]. The algorithm starts with an automated procedure for the digital representation of the molecules. Then, the calculation of a correlation function that assesses the degree of molecular surface overlap and penetration upon relative shifts of the molecules in three dimensions is computed using the Fast Fourier Transformation (FFT). The procedure is repeated many times to scan all the possible relative orientations of the molecules in three dimensions. A complete suite of programs for docking has been implemented on top of this algorithm, which is called FTDOCK [69]. Another example of a docking algorithm that relies on the computation of the FFT for searching all the possible binding modes for a protein by evaluating shape complementarity is ZDock [70]. Another approach for searching spatial complementarities in macromolecules has been recently proposed within an important bioinformatics framework for structural analysis, which is called ROSETTA [71]. The proposed method explores the three dimensional space starting with a random orientation of each macromolecule and a translation of one of them along the line of protein centers to create glancing contact between the proteins. Then the algorithm employs a rigid-body Monte Carlo search, translating and rotating one partner around the surface of the other through 500 random move attempts. Step sizes are adjusted continually to maintain a 50% move acceptance rate, with initial Gaussian perturbation sizes of mean value 0.7 A° along all three axes. This low-level analysis produces a score of the probability of the correctness for each conformation. Generally, these methods works well when trying to orientate the proteins present in the crystal structure of the same complex, while it is more difficult to build a complex starting from proteins in their native forms. The reason for this is that proteins in native form must rearrange their side chains in order to optimize the interaction energy during complex formation. Since these methods are not able to consider side chains rearrangement, they meet some difficulties in considering protein in their native conformations.

1.6.2 **Scoring functions**

Scoring functions are not yet accurate enough to discriminate native and near-native structures and rank models appropriately for all docking pairs. Their development continues to be a subject of considerable study within the docking community. Clearly, the ability to develop a universally applicable function would greatly enhance the docking methodology. The functions can be broadly classified according to whether they are empirical or knowledge-based in origin: functions in the former category are based on a parameterized force field model (e.g. including Van der Waals, electrostatic, entropy and solvation terms), whereas functions in the latter category are based on a statistical analysis of observer contacts in the structural database [46].

Given the difficulty in limiting the number of obtained binding modes and the ranking of the so-called "correct" docked configurations, it is possible to impose additional constraints. More precisely, it would be advantageous to incorporate some information about the chemical properties of atoms, or groups of atoms involved in the interactions into the docking procedure. For example, if the active site is known in advance and if residues playing a role in the binding are known, recent advances in the discussed algorithm greatly improve the ranking of the near native complexed configuration.

A remarkable development in protein-protein interactions modeling is the establishment of CAPRI (Critical Assessment of PRotein Interactions), a sort of competition in which individual groups that develop docking procedures, predict the three dimensional structure of a protein complex from the known structures of the components. The predicted structure is subsequently assessed by comparing it to the experimental structure determined most commonly by X-ray diffraction, which is deposited with CAPRI prior to publication. The predictions are thus made blindly, without prior knowledge of the correct answer, and the evaluation is carried out by an independent team that has no knowledge of the identity of the predictors [72]. From the CAPRI experience, we can see that many different methods start to evaluate protein-protein interactions relying on geometrical properties. This usually happens in early stages of a protein–protein docking, where the goal is to discard solutions that are obviously wrong and reduce the number of potential binding modes. Then other considerations are made, for example physical and chemical properties are taken into account, or models are refined with molecular mechanics in order to optimize interaction between residues.

1.7 Aim of the work

Generally, approaches to identify similar surfaces have demonstrated that similar functions can be found when similar surfaces can be found. For this reason we think that molecular surfaces are interesting, and that analysis can be useful for functional analysis, by finding similarity between surfaces of different proteins, and for interaction study, by establishing complementarities between proteins. However, the goal to be achieved is to describe molecular surfaces accurately enough in order to perform correct identification of similarities, but at the same time to be flexible enough to identify similarities between similar molecules with different orientation of side chains, such as the open or closed form of an enzyme. In particular our aim is to identify regions of local similarity, such as the functional domains or regions able to interact with other molecules. Even the computational time is important to consider: the description must be precise enough to avoid excessively long computing time.

In these terms we want to develop a methodology to analyze molecular surfaces in order to find similarities or complementarities between molecular surfaces. To accomplish this task we proposed a method of computer vision originally oriented to robotics, which is able to describe surfaces through a set of images of local description. This representation enables the correlation of the surface simply by analyzing the similarity between the images from the two macromolecules. More precisely, the core of the matching procedure consists in establishing point-to-point correspondences between meshes by correlating images that describe the local topology of surfaces. The surfaces can then be aligned by minimizing the distances between points from which the corresponding images were defined. In the last step of our methodology the alignment between surfaces must be evaluated, in order to have an idea of surface similarity.

In order to test the algorithm, before trying to formulate a functionality prediction, we have to demonstrate that we are able to correctly identify similar surfaces. To accomplish this validation it is necessary to compare protein surfaces with known functions with protein surfaces with a known function. If the algorithm works well, we should be able to align surfaces with the same functions. The same concept is applied even in protein-protein interaction prediction: we have to validate the algorithm by rebuilding protein complexes as we find them the crystal form.

In conclusion: the aim of the work is to develop a method for recognizing molecular surfaces, and assess it in the context of structural genomics.

2 THE SURFACE MATCHING ALGHORITM

The problem of comparing similar surfaces is a very complex concept since it cannot be based on traditional methods such as template matching or searching for fold similarity. Since molecular surface is a geometrical concept, it can be represented by the aid of computer graphics and analyzed by algorithms of 3D shape matching. However, the difficulty lies in definition of the shape matching problem, since the solution is not unique. Indeed, given two surfaces *S* and *M*, it is difficult to define a matching function which associates to any point of *S* with a corresponding point of *M* relying on surface similarity. In this context the difficulty lies in finding the matching function for an optimal solution: considering for example the computational cost, it would be unthinkable to evaluate each solution exhaustively. Despite this, the exploitation of shape matching techniques applied to macromolecular structures has the advantage of disregarding the internal atomic structure and to find non trivial similarities that can validate the attribution of homology. For example the analysis of clefts by comparison with proteins of known function can provide fair indicators of protein function. Moreover, the study of molecular surfaces can be interesting because molecular surfaces are involved in molecular recognition, and their description can shed light on how proteins interact and how molecular processes take place.

This chapter will discuss the matching algorithm. In our approach, the search for surface similarity can be done by representing two given surfaces using triangular meshes and trying to superimpose them relying on the correspondences found by comparison between surfaces. This comparison cannot be performed trivially by means of the Euclidean distance assessment between meshes, since two surfaces could be similar but not identical. Furthermore, considering two mesh instances of the same object, points could be positioned in different places, and this elucidates how the matching procedure can be difficult for non identical surfaces. Moreover, it is necessary to make the algorithm independent from the reference system to allow fast comparisons for any mutual orientation. For such reasons we proposed a method of computer vision originally oriented to robotics [73] to compare molecular surfaces by describing them using a set of images of local description. This representation enables the correlation of the surface simply by analyzing the similarity between the images from the two macromolecules. More precisely, the core of the matching procedure consists in establishing point-to-point correspondences between meshes by correlating images that describe the local topology of surfaces. All these correspondences are then filtered, using specific similarity thresholds, and clustered, employing different methods, for providing consistence to full matches. The surfaces can then be aligned by minimizing the distances between points from which the corresponding images were defined. In the final stages of the process alignments are evaluated by measuring some parameters or by comparing the results with biological structures.

2.1 The PDB file

In order to study molecular surfaces we need to have three-dimensional coordinates of atoms that compose proteins. These coordinates are derived from nuclear magnetic resonance (NMR) or by X-rays diffraction. Furthermore some models may be derived from homology modeling, relying on structures of the most similar protein sequence, but such models are less precise. All protein structures are available from the worldwide protein data bank (wwPDB) [74], a project which consists of organizations that act as deposition, data processing and distribution of macromolecular structures. All the newly resolved structures are submitted to one of those organizations (RCSB, PDBe and PDBj), and then the community synchronizes all the databanks so that the content is always the same in each database.

The most used standard for recording structural coordinates is the PDB file, a column separated file in which the coordinates of each atom are recorded along with information about function, secondary structure, citations and other information. Each PDB file is identified by a four character code in which the last three letters are assigned arbitrarily while the first character is always a number that identifies the revision of the model. A PDB file is characterized by a series of records and a particular syntax, as described in wwPDB documentation [75]. This is necessary to allow the development of computer programs that are able to analyze macromolecular structures. Each PDB file is the starting point for the algorithm of similarity recognition: from each ATOM record, atom's coordinates will support the definition of molecular surface.

2.2 Surface construction

Molecular surfaces are calculated starting from their three dimensional structures by MSMS program developed by Sanner [76]. This program allows computing a molecular surface starting from a union of spheres centered on atomic coordinates specified by the PDB file and having a radius equal to the *Van der Waals* radius of the atoms. A more precise description of this concept can be found in 1.5.1. A key observation is that if the structure lacks certain atoms, the computed surface will be affected by this issue: for example, it is difficult to derive the position of all the hydrogen from MNR structures . This problem will also affect the phase of surface comparison. To address this problem, the program allows calculating molecular surface relying on an *atom-united* model [77]. With this model, hydrogen atoms are considered implicitly by enlarging the *Van der Waals* radius of carbon, oxygen and nitrogen atoms. As a consequence spheres of carbons, nitrogen and oxygen incorporate the hydrogen atoms internally, and knowing the position of hydrogen atoms is no longer necessary (Figure 4).

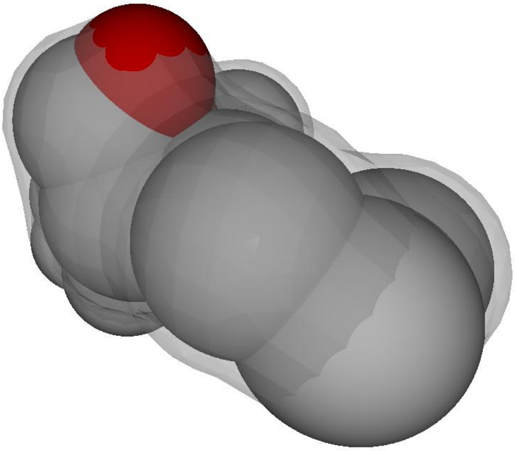
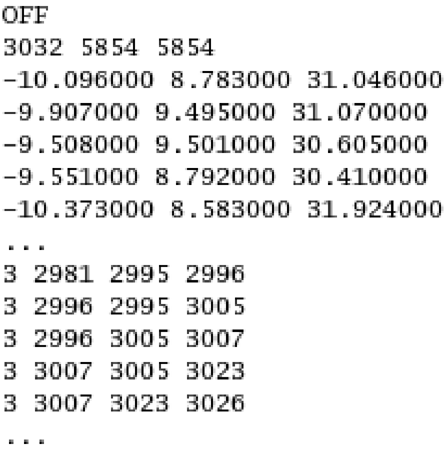


Figure 4. This figure presents the transparent molecular surface derived with atom-*united model* and the Van der Walls spheres of its atoms. Notice that this model incorporates the hydrogen atom in red.

Generally surface is often represented by polygons. In particular, the output of the MSMS program is surface represented by triangular mesh, geometrically represented by a points cloud and topologically described in terms of the connectivity. These meshes can be coded in different formats, but for our purposes we chose the Object File Format (OFF) [78], which consists simply of two matrices, the former describing point's position and the latter representing their aggregation in polygons. An example of the OFF file is presented in Figure 5.



```
OFF
3032 5854 5854
-10.096000 8.783000 31.046000
-9.907000 9.495000 31.070000
-9.508000 9.501000 30.605000
-9.551000 8.792000 30.410000
-10.373000 8.583000 31.924000
...
3 2981 2995 2996
3 2996 2995 3005
3 2996 3005 3007
3 3007 3005 3023
3 3007 3023 3026
...
```

Figure 5. The OFF file format

2.3 *The matching process*

The developed system for matching surfaces is summarized in Figure 6. For simplicity henceforth we will call the surface that we want to align *scene (S)* and the surface on which we want to do the alignment *model (M)*. The matching process is summarized as follows. The first step is related to the “shape conversion” aimed to transform the surface triangular mesh representation into a set of local surface descriptors with object-centered coordinates. This type of representation allows describing surface in a coordinate independent way, thus facilitating the

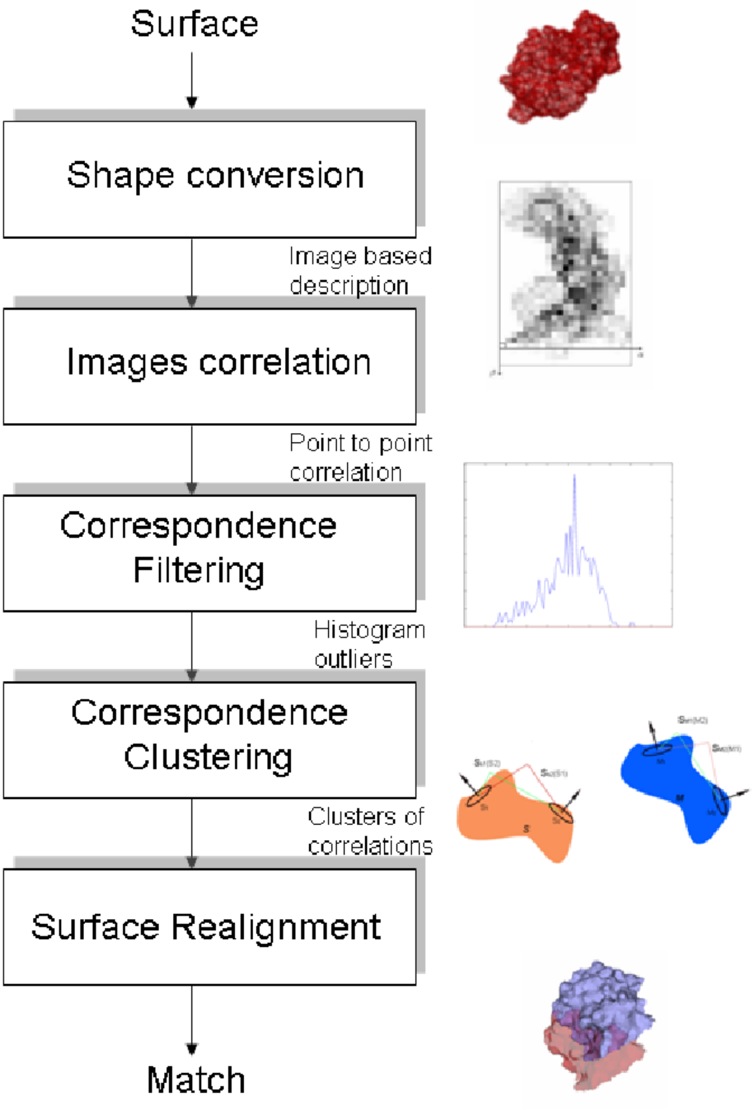


Figure 6. General schema of the system

subsequent stages of surface matching. The second step concerns the “image correlation”, which relies on a method for establishing point-to-point correspondences between surface images. This identification of such correspondences between surface points may lead to the identification of local regions of similarities. In the third step, “correspondence filtering”, the best correspondences are identified. The remaining correspondences are processed in the next step with a two-level “correspondence clustering”, in order to obtain consistent groups of correspondences which indicate local regions of similarity. In the final step, the “surface realignment”, the surfaces are aligned on the basis of the correspondences groups identified in the previous step. This phase is very important since we can establish a definitive measurement of the matching quality.

2.3.1 Shape conversion

The first step of our algorithm is the transformation of the surface triangular mesh representation into a set of local surface descriptors with object-centered coordinates. Object-centered representations have the advantage of providing an independent view description of the surface that allows direct matching, without aligning the surfaces first. Unfortunately, the definition of an object-centered view is difficult because these coordinate systems are generally based on global properties of the surface. Moreover, univocal rules for the identification of specific coordinates systems are crucial to provide similar decompositions of similar objects, improving the possibility of finding valid matches.

A suitable solution for this task has been identified in a method of computer vision oriented to robotics applications, which describes shapes using object-centred systems of reference [73]. Through this local system of coordinates, centred in mesh vertices, it is possible to obtain a set of images for each macromolecule, which represents the surface exhaustively. The key components for the generation of the images are the oriented points, which are surface vertices with an associated direction. An oriented point *O* is defined through the position of a mesh vertex *p* and the orientation of its surface normal *n*, computed as the average of the normals of the triangular faces that insist on such vertex. Through the definition of an oriented point it’s possible to define the location of all other vertices defining a value α, which represent the distance of the vertex from the normal, and a value β which represents the distance of the vertex from the tangent plane P (Figure 7). In this way all the points that describe molecular surfaces are represented in a cylindrical reference system. Using an oriented point basis *O*, we can define a spin-map S_O as the function that projects three-dimensional points *x* to the two-dimensional coordinates (α,β) of a particular basis *(p,n)* corresponding to oriented point *O*.

$$S_O(x) \rightarrow (\alpha,\beta) = \left(\sqrt{\|x-p\|^2 - \big(n\cdot(x-p)\big)^2}, n\cdot(x-p)\right)$$

Although α cannot be negative, β can be positive or negative. The term spin-map comes from the cylindrical symmetry of the oriented point basis.

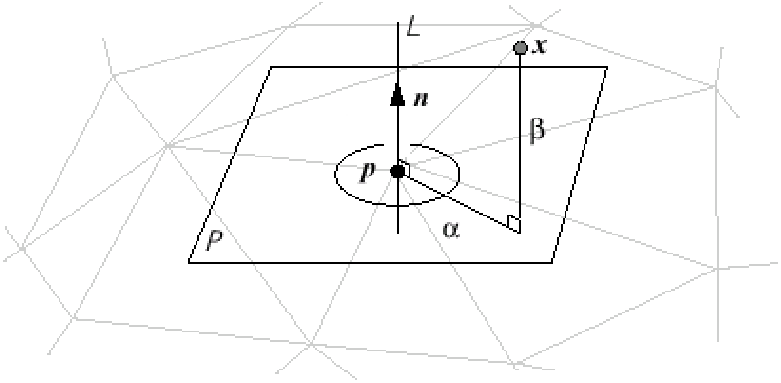


Figure 7. The cylindrical coordinate system defined by an oriented point

The image of local description is then obtained by grouping the vertices projected on the map in a matrix. More precisely, in this matrix a vertex is projected relying on its *α,β* coordinates and in this matrix vertices belonging to a local piece of surface are projected. To count the density projection in each zone of the map, the tangent plane is divided in many sectors: the corners of each sector represent counters, called bins, which are updated through a bilinear interpolation (Figure 8). This solution permits one to update the counters adequately according to contribution of each projected vertex, making the image less sensitive to variation. Two meshes even representing very similar surfaces can have vertices in slightly different positions, which make the matching very difficult. This problem is overcome through the blurring effect introduced by the bilinear updating system of the image matrix.

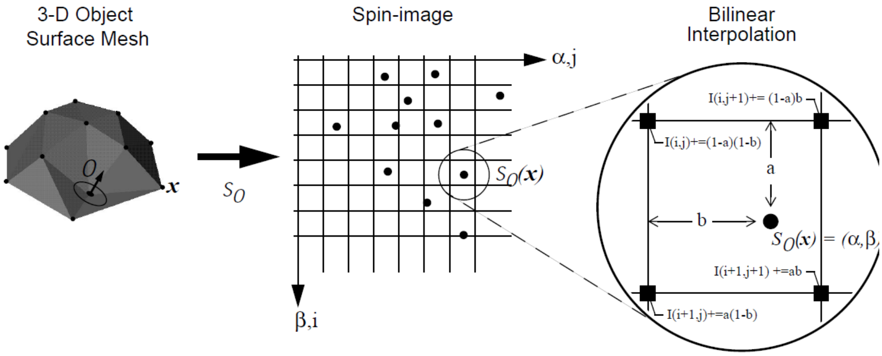


Figure 8. The addition of a point to the matrix representation of a spin-image

The normal orientation generally defines the outer layer of the surface and it is very important because it is crucial even to exploit the dual approach of similarities and complementarities analysis. In fact, if the algorithm works for searching similarities, the surface normals have to be oriented outwards for both meshes. On the other hand, to capture complementarities the orientation of one surface has to be complemented and the easiest way to accomplish this task is to invert the direction of the surface normals inwards (Figure 9). This operation can be easily performed by defining an orientation for the surface mesh, and consequently the issue of having an algorithm applicable to both the situations is achieved.

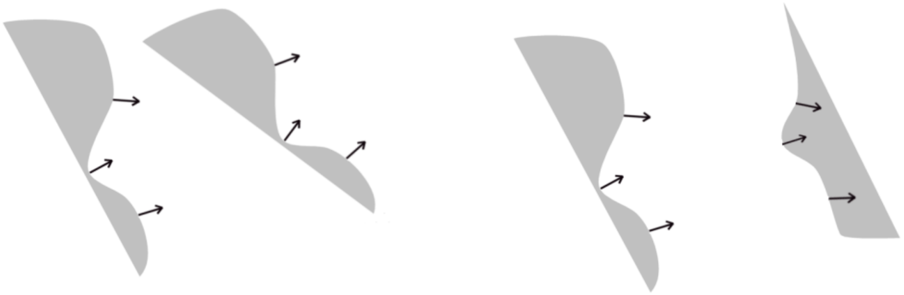


Figure 9. The normal orientation in search of similarities and complementarities

2.3.2 Image correlation

The object-centered description obtained by calculating spin images on surface vertices has the property of being independent from rigid transformations. This means that two instances of the same object in two different positions will have identical images. On the other hand the comparison of all the possible pairs of images belonging to two different objects allows one to determine point to point correspondences. If similar surfaces generate similar images, finding similar images means that surfaces are similar. For this reason we can use image processing functions to correlate surface points. In particular two images from corresponding vertices on similar surfaces are expected to be linearly related, because the number of points that fall in corresponding bins will be similar. For such reason we use *normalized linear correlation coefficient* as a similarity measure to compare all possible pairs of images. This index can vary between -1 (anti-correlate images) and 1 (completely correlated images): two images with a high similarity measurement are likely to come from corresponding points.

2.3.3 Correspondence filtering

Once linear correlation coefficient is applied to evaluate image similarity, the correspondences are then filtered in order to maintain only significant correlation. Two different filters are applied during this stage. The first one attempts to maintain significant matches by removing those correspondences that have similar correlation coefficient over the entire surface. More precisely giving an image on surface *S*, all correspondences between this image and all the images on surface *M* are calculated to generate a histogram of all correlation coefficient. Then we calculate the distance F_s between the first and the third quartile of the histogram, to determine a threshold by increasing the median by 1.5 F_s (Figure 10). All correspondences with a correlation coefficient higher than this threshold are considered outliers and their correlations correspond to image pairs which are

more similar than the others, which indicate plausible corresponding points between the two macromolecules. On the other hand, if no outliers are found, it means that there are no significant correlations between images on surface *S* and all other images on surface *M* and so all these correlations can be discarded.
The second filter applied in this stage reduces the number of correspondences relying on the value of their correlation. The combination of these two types of filtering ensures that correlations between images have significant values higher than average.

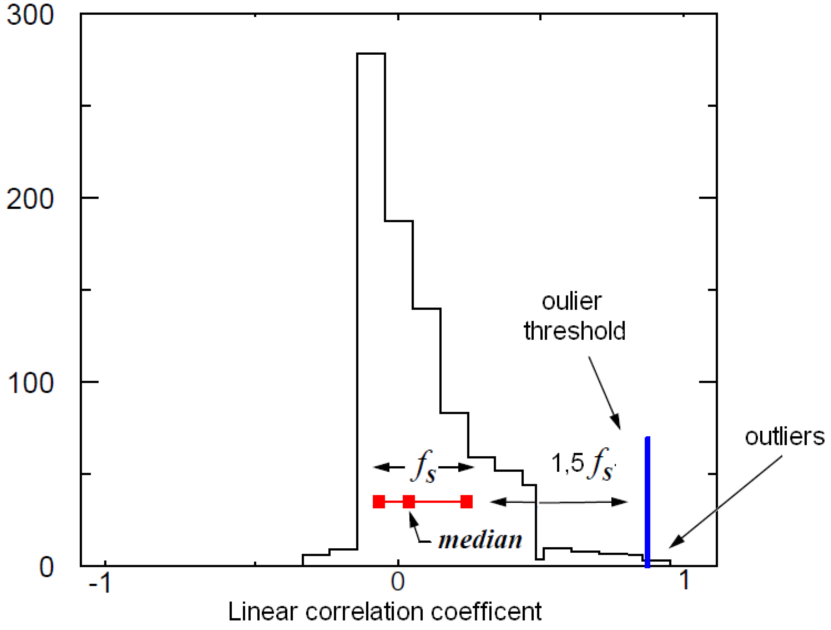


Figure 10. Correspondences are filtered by median values

2.3.4 Correspondence clustering

In order to identify a 3D transformation that’s able to realign surfaces according to a matching prediction, at least three geometrically consistent correspondences need to be established. Clearly, the correspondences provided by the system may belong to different partial matches within the same surface comparison analysis and in order to provide only a few consistent matches a double clustering procedure is adopted in the algorithm. The first clustering is related to the geometrical coherence of the vertices correlation: several punctual correspondences are grouped using geometric consistency to calculate a few univocal transformations able to realign surfaces. Then, another clustering is performed, according to the necessity of working on small patches, for organizing correspondence points on the two surfaces in order to provide groups of correlations localized in restricted and well defined regions.

2.3.4.1 Geometric filtering

To avoid combinatory explosion, geometric consistency is used to determine a group of correspondences from which plausible transformations can be computed and to eliminate correspondences that may affect the matching process negatively. Indeed during the matching process a single point can be matched to more than one point for two reasons. The first reason is that symmetry in data and in spin image generation may cause two points to have similar spin images. Second, spatially close points may have similar spin images. This may be caused during the matching process when some points from a surface *(scene)* that doesn't overlap another surface *(model)* may be incorrectly matched. In general, these incorrect correspondences will have low similarity measures and will be geometrically inconsistent when compared to the rest of the correspondences.

The geometric consistency is a measurement of the likelihood that two correspondences can be grouped together to calculate a transformation of model to scene. If a correspondence is not geometrically consistent with other correspondences, then it cannot be grouped with other correspondences to calculate a transformation, and it should be eliminated. Considering two correspondence $C_1=[s_1,m_1]$ and $C_2=[s_2,m_2]$ between two spin images, we can evaluate the geometric consistency by comparing the spin map coordinates of the point m_1 in the coordinate system of m_2 and s_1 in the coordinate system of s_2, and vice versa (Figure 11). Then the geometric consistency is calculated by the formula:

$$d_{gc}(C_1, C_2) = \frac{\left\|S_{m_2}(m_1) - S_{s_2}(s_1)\right\|}{\left(\left\|S_{m_2}(m_1)\right\| + \left\|S_{s_2}(s_1)\right\|\right)/2}$$

d_{gc} measures the distance between the correspondences in spin-map coordinates normalized by the average length of the spin-map coordinates. Spin-map coordinates are used to measure geometric consistency because they are a compact way to measure consistency in position and normals. Distance between spin-map coordinates is normalized by the average spin-map coordinates of the correspondences so that the geometric consistency measure is not biased toward correspondences that are close to each other. When d_{gc} is small the *scene* and *model* points in C_1 are the same distance apart and have the same angle between surface normals as the scene and model points in C_2. By this solution it is possible to filter out bad correspondences based on the properties of the correspondences taken as a group and then the matching process will be more robust. This solution does not require reasoning about specific point matches to decide which correspondences are the best.

2.3.4.2 Agglomerative clustering

Finally, correlations have to be clustered to achieve a few significant transformations, which should be able to realign the matching surfaces. The possibility to perform partial matches is very important to correlate specific macromolecular regions which have significant functions. A particular issue we tackled in relation to the biological application domain is that the algorithm should be able to establish correspondences working on small surface patches, which are effectively involved in a biological interaction, in order to avoid the superimposition of the structures. Indeed, if within a cluster most of the correlations are located in a specific area, there are a few others in different places, and the algorithm tends to

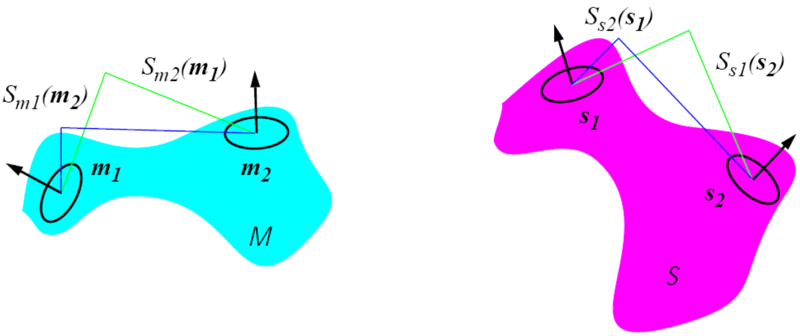


Figure 11. Mutual projections of reference vertex between surface S and M. When coordinates α and β between the oriented point bases of two correspondences are similar on both surfaces, the correspondences are geometrically consistent.

move meshes one inside the other to minimize the distance among correspondences, causing clashes between the surfaces. To avoid this effect, we try to remove correspondences which are out of the area delimited by most of the correlations from each group.

The approach adopted is based on agglomerative clustering, which at first considers each element as a single cluster and then tries to join similar clusters in an iterative manner. Using agglomerative clustering, each object is initially placed into its own group. Therefore, if there are *N* objects to cluster, the algorithm starts with *N* groups. Each of these groups only contains a single object, which is known as a singleton. Before starting the clustering, the threshold distance T_{ac} for groups joining has to be fixed. Then, all pairs of groups are compared and the closest pair is selected. If the distance between these two clusters is lower than the threshold distance T_{ac}, the groups are merged and compared again with the others, otherwise the clustering procedure stops. It is worth noting that if the threshold value is too small, many groups will remain at the end of the clustering procedure, and many of them will be singletons. Conversely, if T_{ac} is too large, objects that are not very similar may end up in the same cluster.

To perform an agglomerative clustering, a method for measuring the distance between two objects has to be defined. In our algorithm, the choice was to measure the distance between two correspondences in terms of the Euclidean distance of the coupled points on both the surfaces. The problem of considering distances on both of the surfaces has been addressed by reassessing distances on the first surface considering the correlated point positions on the second surface. Some options for the implementation of the agglomerative clustering are simple linkage, average linkage and complete linkage, which differ on how the distance, well defined between two objects, is effectively computed among clusters. In this algorithm we chose to use the simple linkage approach, which means to join clusters in consideration of the distance between the two closest elements in the two clusters. By introducing this clustering approach, correspondence points which are far from the other on at least one of the two surfaces are excluded by the correlation set and the result is a close localization of the matching points.

2.3.5 Surface realignment

The last step of the algorithm concerns the realignment of the matched surfaces on the basis of the clusters recognized in the previous stage. The realignment plays a crucial role in giving a final evaluation of the matching quality, because it enables a definitive representation of the predicted matching configurations. To accomplish this task each cluster of correspondences has to be considered in order to compute a 3D transformation of coordinates able to realign the two surface patches according to the elements of the cluster. Although the geometrical consistency is granted by the first clustering level, a 3D transformation of coordinates has to be calculated by solving the non-linear equation system which describes the distance between the corresponding vertices of the two macromolecules. A plausible rigid transformation T between the two surfaces is calculated from each group $[m_i; s_i]$ of correspondences by minimizing:

$$E_T = \sum \|s_i - T(m_i)\|^2$$

This issue has been solved by computing the root mean square distance between couples of correlation points using the Levenberg-Marquardt algorithm [79], a well known method for solving non linear problems of minimization. Using this approach the procedure is able to rank the different matches using the effective Euclidean distance between the first macromolecule and the rearranged spatial configuration of the second one.

2.4 Results evaluation

Once the surfaces are aligned, there is a need to make a quantitative assessment of the quality of the alignment. For example in functional annotation, where a query surface is aligned against surfaces associated to different functions, it is useful to have a ranking of the results in order to have an idea on the most likely function of the query surface. The same concept applies in search for complementarities: we have to find a way to identify the best alignment within the solutions space. Moreover we have to remember that surfaces are not only geometric objects, they also have physical-chemical features and it would be interesting to evaluate them in the context of search for similarity / complementarity.

We use substantially four methods to evaluate the quality of the alignments. The first parameters is the *overlap* value, which expresses the percentage of aligned surface points that are at a distance less than 1 A° from the other surface. This parameter works well in context of search for surface similarities. Indeed, extracting a patch from a surface which corresponds to a functional domain and trying to align it on another surface with the same functional domain, the overlap values should have a high score since functional domains may have similar geometric features. This parameter can't be used: in the context of a search for surface complementarities because it doesn't necessarily identify the best alignment. Indeed the overlapping area of two interacting surfaces can be small and can have worse scores than alignments with no biological sense. Moreover sometimes it can bring us to the wrong conclusions: if the overlap value is not high perhaps the corresponding alignment is not what one would expect to find. In such cases it is better to visually inspect the result of the alignment.

To assess the alignment of two complementary surfaces we calculate the *root mean square deviation (RMSD)* on the protein backbone between the aligned protein and the protein in native conformation. This parameter is certainly useful in the testing phase, but becomes unusable in prediction phase, when we want to determine the proper conformation. Moreover this parameter tends to penalize small shifts in the case of large structures. Always in the context of complementary surface alignment, the index of compenetration may be useful to eliminate bad alignment solutions. Indeed, as results of the complementary search we can obtain hundred of results where many of these make no biological sense, for example when structures compenetrate themselves. For this reason, we evaluate the degree of compenetration in order to eliminate such solutions in which atoms are closer than their Wan der Waals radii. However, it is difficult to implement a program to measure the degree of compenetration as the volume subtended between two complex geometric objects such as molecular surfaces. In theory, it would be more correct to assess the degree of compenetration through a scoring function able to measure the contribution of *Van der Waals* interactions between all the atoms of the two proteins, but even this solution is difficult to implement. To address this problem, we implemented and heuristic solution in which we measure how points belonging to one surface fell within the volume of atoms used to generate the second surface. However this parameter is effective in filtering solutions where the degree of compenetration is very high, considering that even the correct solution has a certain degree of compenetration and it is difficult to discriminate this solution from the others in which the value of compenetration is lower since structures are aligned outside of the site of interaction.

The last parameter that we evaluated concerns the evaluation of the electrostatic potential of both the aligned surfaces. First of all we evaluate the electrostatic potential for the two proteins that we want to align. This is achieved using APBS [80], a program in which the protein is put in a three dimensional grid and the electrostatic potential is evaluated for each node of the grid. Then we assign a precise electrostatic potential value to each vertex of the protein surface through trilinear interpolation of surface points on the electrostatic grid. Then we identify the nearest vertices on the two surfaces similarly to how we evaluate the overlap value. Finally, from these vertices we evaluate the linear correlation coefficient of the two potentials. In the context of search for similarity these vertices are linearly correlated, since if the functional sites are similar then we will expect to find the same physic-chemical features. In the context of search for complementarity these vertices are expected to be anti-correlated, since the fact that if surfaces are able to interact they will present opposite charges. Even this parameter suffers from some limitations: computational analyses of the surface properties of protein-protein interfaces [39] have demonstrated that interfaces display electrostatic complementarity, and this generally means that vertices on surfaces are anti-correlated. However the measured value makes it impossible to assign electrostatic complementarity unambiguously. With regards to the search of similarity the correlation is stronger when similarity is high. In the case of low similarity these values can vary significantly.

3 PARAMETER ESTIMATION

The process of recognition of similarity is very complex since different parameters affect the functioning and the performance. The optimal choice of these parameters becomes a crucial problem: on one hand we want that the recognition process works at the best of its possibility, on the other we want to prevent this process from becoming computationally expensive. In this chapter these two aspects are considered. Firstly, the approach to indentify the best parameter configuration is described. Then the techniques to optimize the recognition process that entails a decrease in computational time are presented.

3.1 Parameters that affect the search of similarity

As mentioned earlier, the process of similarity detection is affected by 15 parameters. Some of them concern image generation, while others affect the number of image correlations, which are used to match similar surfaces. The choice of these values has important implications on the entire procedure: choosing parameters that are too tight in the early stage of the algorithm implies a negative impact on results even if the other parameters are set appropriately. For this reason it becomes essential to understand the role of each parameter in the recognition process.

3.1.1 **Parameters related to spin-image generation**

The first important parameter is the density of points that represent molecular surfaces. As mentioned above, molecular surfaces are determined in an analytic way and then are represented by triangular meshes. Thus a surface can be represented with more or less detail (Figure 12) and this has consequences on the number of images that will be considered in the following steps of the algorithm. Generally, representing low definition surfaces has positive effects on computational time but also a negative effect on the similarity recognition process. Indeed, a high definition surface minimizes errors due to the different generation of points that describe the molecular surface. If surfaces are described in detail, there are minor differences in point generation and will be easier to recognize two similar surfaces as such.

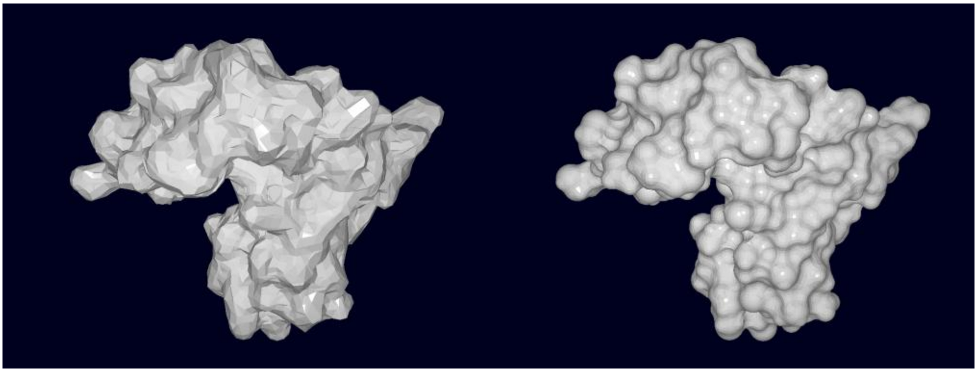


Figure 12. The same molecular surface represented with two different level details

Even the choice of points on which calculate spin images is important. It is reasonable to assume that representing surface in detail involves two adjacent points to generate similar images so we can reduce redundant information avoiding generating images on vertices adjacent to those used for images generation. For this reason all points in the neighborhood of a reference vertex are marked as not available for further map creation.

Another criterion proposed to limit the number of images generated is related to surface curvature, which is considered to be an important indication of the information that each single point can bring to the representation. In literature there are few examples that document that protein-protein interactions occur in flat surfaces, while ligand binding is related in molecular cavities [30]. Moreover small binding sites seem to have rough surfaces in order to maximize interactions between atoms [47]. An important parameter of our algorithm is the interval of curvature values for the selection of the reference vertices, which depends on the analysis context. By this parameter we can decide to create spin images by selecting points on smooth or high curvature areas. In this way we can create images only for points that actually bring important information for the matching procedure. This solution avoids a redundant representation of surface patches, and therefore reduces the computational time without any impact on the quality of results (Figure 13).

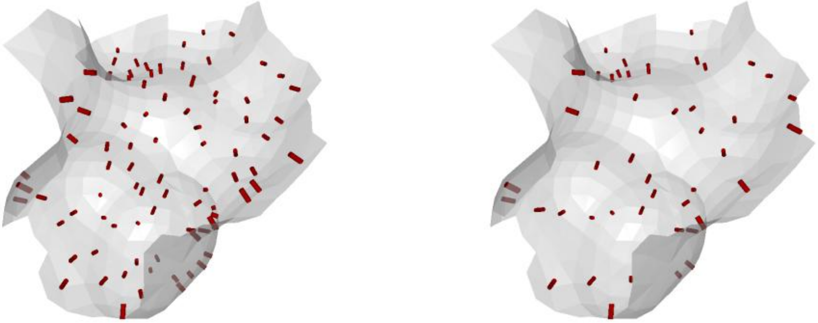


Figure 13. This figure presents the effect of reducing the number of points for spin images generation

The same principle of avoiding redundancy in the information representation can be adopted for the surface projection area in each image. For every cylindrical reference system only a limited piece of the surface is projected into the corresponding image. This region is selected by searching iteratively all the vertices that are in the neighborhood of the reference point, following the edges starting from it. This choice is important in order to reduce the overlapping of different local details on the same image, improving the accuracy and the locality of the searching algorithm. Consequently, an important issue related to the number of radial iterations performed for selecting points to project in each reference system. However, this parameter depends on the image size and the bin size (the dimension of bins used for bilinear interpolation, see section 2.3.1), which together define the image coverage. During this step of the algorithm, points that are projected in an image are marked as not available for the creation of a new cylindrical reference system. As a consequence, we achieve a uniform surface coverage using the images of local description, because maps can be overlapped only partially, and a suitable representation of the mesh can be obtained. In Figure 14 we can see the effect of bin size and of the iterated projection of nearest point on the same image. By enlarging the bin size, the image is condensed, while raising the number of iteration the image focuses around the oriented point.

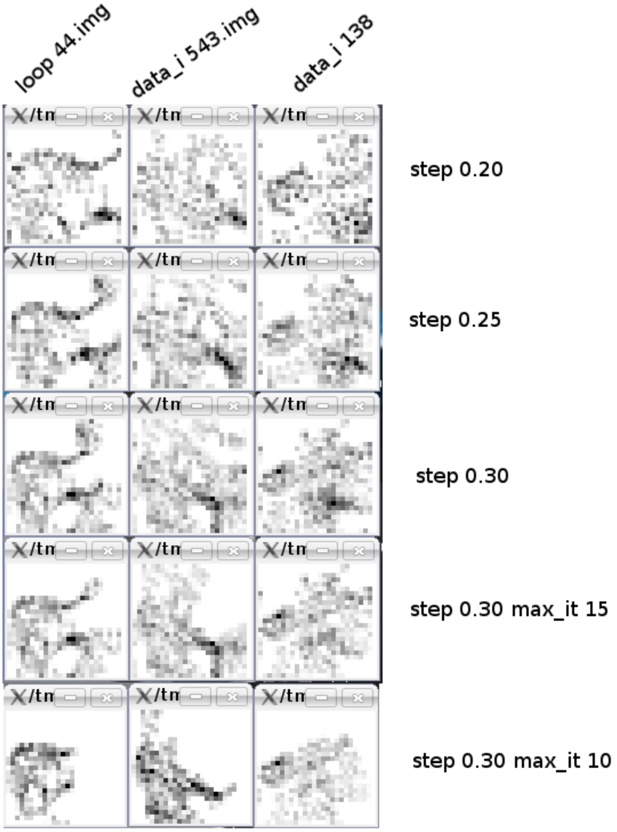


Figure 14. The effect of the bin size and iteration of nearest points in spin image creation.

3.1.2 Parameters related to similarity detection

All the remaining parameters of the algorithm concern the recognition of similarity. These parameters are critical because they regulate the most computational expensive process and a wrong choice can compromise the whole recognition. The calculation time required for image generation is negligible when compared with the calculation time to find correspondences between images.

The first bottleneck is the time needed to evaluate all the correspondences between similar images. Figure 15 reports the time needed for calculating image correlation. The time needed to calculate this step grows in a polynomial way with the number of correlations to evaluate. For such reasons it is better to contain the number of images generated in previous stages. Even to calculate geometric filtering takes time, and also in this step the complexity of calculation is polynomial with the number of correspondences achieved (Figure 16). However, it is possible to limit the number of images that arrive at this point by setting a threshold value for correlation between images. In this way we can assess the geometric consistency only for pairs of very similar images.

The real bottleneck is the geometric clustering of correlation, as every time an item is added to a group, all distances are recalculated. Even at this stage it is possible to limit the number of correspondences by setting a high threshold value for the distance d_{gc} (see paragraph 2.3.4.1). However, even limiting the number of matches the computational time required is still too high since the algorithm is complex. Indeed the clustering algorithm iterates over three basic steps: first of all it starts considering all the elements as groups composed by one element. Then it measures the Euclidean distance between all groups and finally it merges the two closest elements in the same group. The algorithm iterates for several steps, until the lowest distance between two groups is higher than a certain threshold. Regarding the complexity of calculation, the clustering algorithm has to calculate a total of N^2 distances at each step, where N is the number of elements. Even in the final steps of the algorithm, when there are a few groups, all distances between point have to be established: if the clustering process continues until a single group is reached, all these steps should be repeated N times. In this way the time needed for clustering is higher than the other steps.

3.2 Clustering performance

To reduce the computing time needed to perform the clustering procedure, heuristics and special solutions were used. The first solution is to consider the distance matrix as symmetrical, the distances to calculate in each passage are N^2 / 2, since the distance between *i* and *j* is the same of *j* and *i*. In this way we can halve the time required for evaluating distances in each step of the algorithm. We also tried to calculate all distances for all elements, since in each clustering step distance to be calculated is always the same. However, this solution was

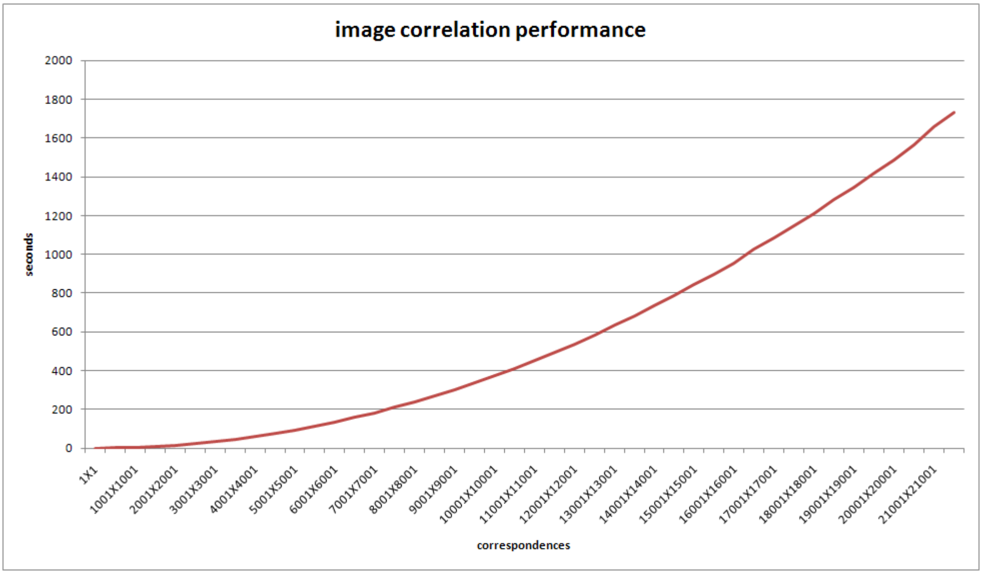


Figure 15. Time needed for image correlation

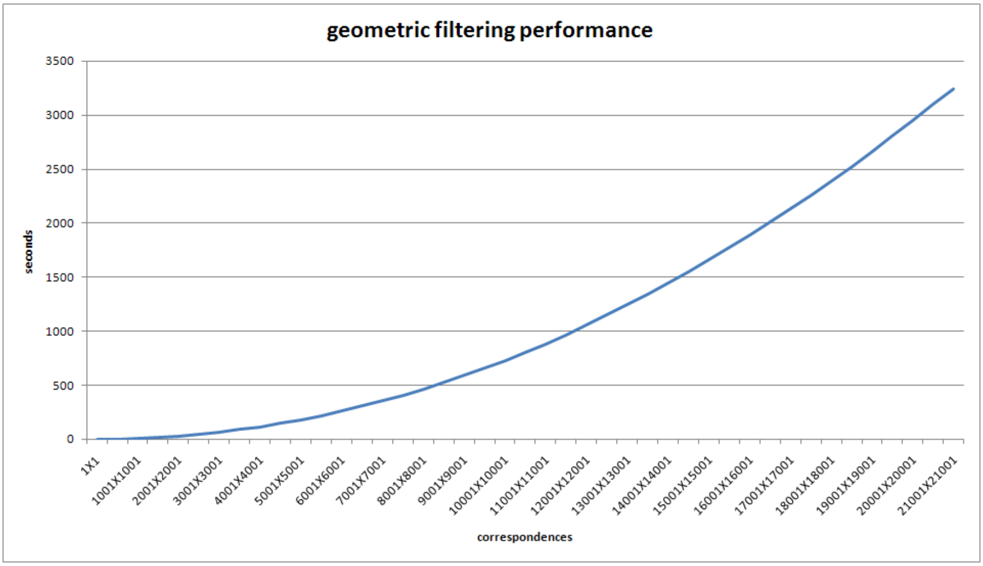


Figure 16. Time needed for geometric filtering

abandoned since it was too expensive in terms of memory and distances have to be evaluated in each clustering step.

To further restrict calculation time we tried to implement clustering with the neighbor heuristic. According to this approach, it is not necessary to recalculate all the distances at each step. In the first step each point stores its nearest neighbor and its distance. Then the two nearest points are joined in the same group and the nearest neighbor for the new group is determined. By this heuristic, all the neighbors of the others points remain unchanged, since they are not included in any group in that step. The only distances to be calculated in each clustering step are those of newly formed groups, in order to identify the nearest neighbor of the new group.

The last heuristic introduced consists in indexing points relying on their spatial coordinates, in order to calculate distances only for those points with index compatible. Indeed, if we divide points in cubes with sides greater than clustering threshold, we can avoid calculating distances between points of two nonadjacent cubes since their distance will be higher than clustering threshold and they will never be included in the same cluster.

Since the introduction of these solutions, the time required for geometrical clustering has been significantly reduced. In Figure 17 we can appreciate how the computational time has been improved in respect to the first implementation of the algorithm.

3.3 Parameters estimation with genetic algorithm

In order to optimize the algorithm, which depends on 15 parameters, we have to estimate its behavior in different conditions. However, it is not convenient to find a parameter configuration that optimizes solutions by hand. Moreover we cannot explore all the possible combination of parameters since the complexity of the algorithm and the computational time required for the calculations shown in the previous paragraph suggest that some parameter configuration require a lot of time to be evaluated. A solution to this optimization problem is to find a satisfactory approximate solution through machine learning techniques.

For this reason, we chose to use Genetic Algorithms to identify a good parameter configuration so that we can successfully apply the surface similarity evaluation to the largest possible spectrum of structures. The Genetic Algorithm is a computational model of biological evolution used to find approximate solutions to optimization problems. By using this model, all candidate solutions are represented by individuals composed by different sets of genes (chromosomes). Starting from a population of randomly generated individuals, the evolution of the population happens in generations: in each step of the algorithm a fitness function is evaluated for all the individuals and only the best individuals are selected for the following generations. Those individuals are recombined and mutated in order to generate a new population with higher fitness. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the current population.

Each model implemented with a genetic algorithm requires a genetic representation of the solution domain and a fitness function to evaluate the solution domain. The fitness function is defined over the genetic representation and measures the quality of the represented solution and it is always problem dependent. This means that its definition is important, since if properly defined. it allows satisfactory solutions to be found for the initial problem.

In our case, the genetic representation of the problem is achieved by considering our parameters as genes, and the fitness of each individual is evaluated by inspecting the quality of the similarity match achieved by each parameter configuration. More precisely, the simulation consisted in aligning two similar surfaces, and giving a fitness score proportional to the percentage of overlap between surfaces measured after the alignment as a measure of surface similarity.

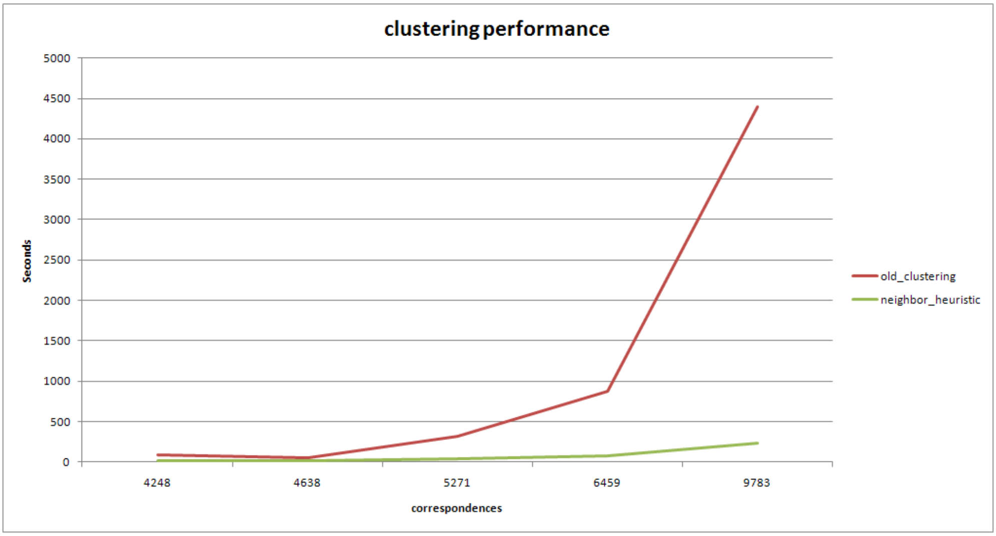


Figure 17. Comparison of times required for the two different clustering techniques

Once we have the genetic representation of the problem and the fitness function is defined, the genetic algorithm proceeds to initialize a population of solutions randomly, and then improves it through repetitive application of mutation, crossing-over and fitness function evaluations. The repeated evaluation of the fitness function for parameter estimation is very time consuming, since each similarity measure can take from minutes to hours on a single CPU. Whereas many steps of the algorithm have a polynomial complexity, it is possible to estimate how much time is needed to complete the calculations, in order to stop expensive simulations and to assign them a low fitness score. This solution means giving up possible solutions, but at the same time resizing the search space towards acceptable solutions in terms of computational time. Even considering this, the evaluation of fitness function remains the most costly operations in term of computational time. For this reason, we have developed a parallel implementation of a genetic algorithm in order to execute it on a computer cluster to improve the fitness function evaluation.

In our implementation the main process handles the population by generating new individuals and selecting the best configuration of parameters. The fitness evaluation is managed by the master process which acts as a server in the master node of the cluster and sends individuals to clients running on the other nodes through socket connections. In this way the fitness evaluation of each individual can be done in parallel by each client process. When one client finishes its calculation, it writes a result file and asks the server for a new individual to evaluate. After each generation, the master process reads all the results files produced by the clients and prepares a new population relying on the fitness to be evaluated by the client processes in the next step of the algorithm. A schematic representation of this solution is presented in Figure 18.

3.3.1 Application in surface similarity detection

With the help of our parallel implementation of this Genetic Algorithm, we tried to superimpose similar surfaces and to evaluate the similarity level by using the image based approach. Molecular surfaces were calculated and represented by triangular meshes: each client process evaluates the similarity level by inspecting the percentage of vertices nearer than 1A° from two aligned surfaces. In our model we chose to evaluate each parameter in 10 steps, to initialize a population of 100 individuals and to perform 200 generations by evolving individuals with two points crossing over.

After about 50 generations on average, solutions converged to the specific fitness values, therefore we decided to terminate the simulation after two weeks. During this period we computed about 100 CPU days, by running each simulation on 20 client processes. Even the time needed for the calculation of the slowest individual affects the time needed to calculate the entire generation we estimate a speedup of about 7 times, proving that our approach is scalable.

Finally we evaluated the quality of the parameter configuration identified by searching for surface similarity for 6 proteins randomly chosen towards a database of patches derived from the extraction of prosite domains from the whole molecular surfaces of the non redundant chain set of protein data bank. In each case we correctly identified similarities with prosite patches in all the proteins used for this test. We concluded that the parameter configuration adopted for these test cases

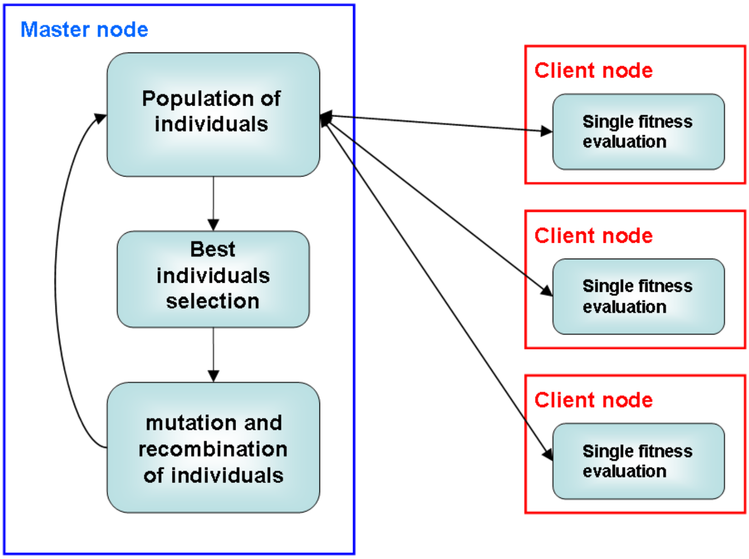

Master node
Population of individuals
Best individuals selection
mutation and recombination of individuals
Client node
Single fitness evaluation
Client node
Single fitness evaluation
Client node
Single fitness evaluation

Figure 18. The main process on the master node generates the population by selecting individuals with higher fitness and sending each individual to the client nodes in order to evaluate the fitness

also applies to other cases and it was adopted as the default parameter configuration of our algorithm.

3.3.2 Application in surface complementarity detection

The parameter estimation with genetic algorithm was also applied in order to identify configurations of parameter to be applied in complementarities detection. In this case to avoid the over fitting of parameters in only one case, the simulation was designed in order to optimize more than one alignment at the same time. We accomplished a test using a well-known non-redundant data set of bounded and unbounded protein complexes. In particular, we performed a wide range test on the Chen data set of protein-protein interactions [81], which consists of 59 co-crystallized structures. First of all, six complexes were isolated and used for parameter estimation. Those complexes have been split and one component was moved randomly in space in order to try to re-identify the interaction surfaces using our system. In this case the fitness function was modified in order to estimate the Root Mean Square Deviation (RMSD) of the aligned structure towards the structure present in the crystal form of the complex. Again, it was considered a population of 100 individuals for 200 generations and we had performed about 2081 CPU days of computation in about 105 days. The configurations of parameters identified were used to realign all the structures of the Chen dataset. For each simulation different matching was identified and among them the best match was selected according to the effective conformation of the co-crystallized structures. According to the evaluation criteria proposed in the CAPRI experience [82], which identifies a threshold of 10 A° for acceptable results, in half of our simulations we were able to find an acceptable solution. These values of RMSD are encouraging considering that our algorithm works at this point without any energetic minimization relying in chemical-physical considerations: we can conclude that we implemented a fairly good solution for an initial fast screening of the possible interactions.

The application of genetic algorithms allows us to identify a satisfactory configuration parameter, but there are some contraindications. The first is that fitness function evaluation has a high computational cost. In this case we decided to sacrifice the computationally expensive solutions with the risk of losing satisfactory solutions. Another point is that the best solution we choose is the best solution among all those found: knowing when to stop a simulation once a satisfactory solution is identified, is not an easy problem to solve. Moreover genetic algorithm has a tendency to converge to local optima rather than the global optimum of the problem. This means that it does not "know how" to sacrifice short-term fitness to gain longer-term fitness. To address this problem we can increase the rate of mutation or we can use selection techniques that maintain a diverse population of solutions. A common technique to maintain diversity is to impose a "niche penalty", wherein, any group of individuals of sufficient similarity (niche radius) have a penalty added, which will reduce the representation of that group in subsequent generations, permitting other (less similar) individuals to be maintained in the population. At the moment, we have only imposed a mutation rate high enough to ensure diversity. Indeed diversity is important in genetic algorithms (and genetic programming) because crossing over a homogeneous population does not yield new solutions. Nevertheless, it is difficult to set the mutation rate properly since a very small mutation rate may lead to genetic drift, where a high recombination rate may lead to premature convergence of the genetic algorithm or may lead to loss of good solutions. However, when individuals begin to converge on a precise set of parameters, it becomes difficult to explore other configurations. To deal with this problem, we are planning to introduce other heuristics such as the niche penalty, or increasing the probability of mutation when the solution quality drops (called triggered hypermutation). Another solution is to occasionally introduce entirely new, randomly generated elements into the gene pool (called random immigrants).

3.4 Calculations execution through the GRID infrastructure

The last consideration to make in regards to the performance of the similarity detection pipeline is that for determining surface similarity among different proteins the entire process starting from surface generation to surface alignment has to be completed, which is a computationally intensive task. This means that the time needed for evaluating a protein surface towards a database of molecular surfaces is equal to the time required for evaluating each single case. Nevertheless, the analysis of each single case is independent from all the others and this means that all the work can be split up on different machines, scaling the time required for doing all the evaluations. To accomplish this task, we decided to use the Grid platform in order to execute all the analysis on multiple CPUs.

We considered in particular the use of EGEE grid [83], which consists in a network of several Computing Elements, which constitute the gateways for the computer clusters on which jobs are performed, and an equal number of Storage Elements, that implement a distributed file system on which data is stored. All the elements of the grid infrastructure are deployed as Grid Services, which envelope the main provided functionality enabling secure communications among the grid components. In particular, the gLite middleware [84] has to be installed on a local server within the institution that wants to join the EGEE grid in order to establish secure communications between the user itself and the distributed infrastructure. This local server, which is usually called User Interface, enables the submission of jobs, to monitor the state of advancement of the jobs, to retrieve the outputs when the computations have a normal termination and to resubmit jobs in case of failure.

The gLite middleware provides some services for job submission, monitoring and results retrieval, but the manual management of each job for a big challenge can be daunting for scientists [85]. In our experience, we have shown that the implementation of an efficient coordination system for submitting and monitoring jobs on the top of the EGEE grid environment increments the efficiency of large scale computations. A further advantage is to exploit the Computing Elements even for providing the post process of the results. From the computational point of view, the major problem of the Grid is the dynamic behaviour of the available resources. Due to network and system errors, or depending on the global computational workload, the available resources are continuously reshaped and the rate of failure in computations is quite high. The crucial point is then the creation of a fault tolerant infrastructure that, by tracing the status of the job constantly, allows an automatic resubmission of every task that presents problems or inconsistent status [86].

For this reason we used an automatic system to coordinate the job submission and monitoring system provided by the EGEE grid. This infrastructure hides the complexity of managing data by automating the whole process of challenge coordination and allowing a user with no computer skills to take advantage of the computing resources provided by grid. This solution has been adopted in the context of the surface analysis, but it could also be useful in other situations.

3.4.1 The challenge coordinator system

The system proposed was implemented to work on the top of the gLite middleware and it is able to completely coordinate a challenge running on a single User Interface: the whole computation can be divided into a set of small jobs in order to execute them on different Computing Elements. All the necessary operations are fully automated, allowing a user with limited computer skills and no knowledge on the grid environment to use the potential of calculation provided by the grid in the simplest way possible.

The Challenge Control System implemented can be viewed as a double-layered infrastructure in Figure 19. The first layer is designed to control the execution of each single job: it’s composed by the execution manager and a monitoring database and its role is to manage job execution, submitting jobs and monitoring their status using the instruments supplied by the grid itself. A key aspect of this system layer is that it acts on the working directory at the moment of job submission by packing the data files and submitting them on the Storage Elements. Then a “job description” file with all the information needed to rebuild the working directory and to execute users calculations is created and submitted to the grid environment. Once the results have been completed, the data is recorded on the Storage Elements and downloaded locally by the System.

The second layer coordinates the distribution of the whole challenge by creating a directory containing all the files needed by the calculations. This layer also prevents submitting too many jobs on the grid environment at the same time. The two layers can communicate through a relational data base that stores the challenge crucial information: the monitoring activity of jobs and files distribution, for the first layer, and the Input/Output status of the application and the parsed results, for the second layer. When using this approach a strict control of data integrity can be performed by checking the consistency between the information of the two layers. More precisely when a job is executed on the Computing Element, it performs the post processing of the results exploiting the computing element and then records all the data produced in the Input database.

When results are downloaded from the grid by the first layer, they are automatically checked for consistency with the database and in case of inconsistence the job is considered as failed and resubmitted to the grid by the first layer. In such a way the database of the Challenge Coordinator ensures the consistence of the I/O data. Through the two layer system proposed all the processes needed for the correct challenge execution and for check the integrity of the results are completely automated and hidden, allowing the utilization of resources of the grid with minimal intervention by the user.

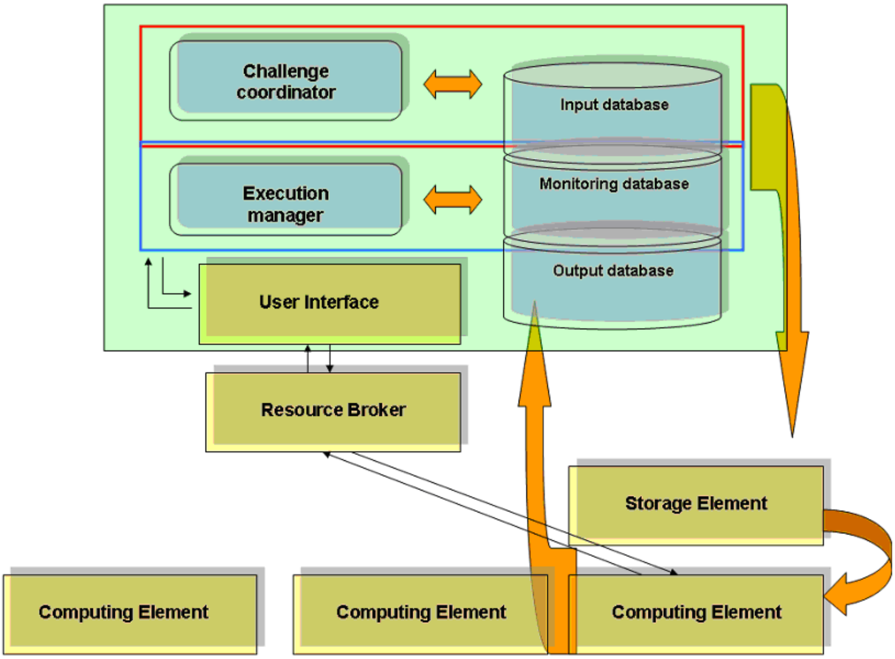


Figure 19. The system proposed to perform functional genomics computation on Grid platforms like the EGEE project infrastructure. The software is structured into two different layers: the topmost one is highly application oriented while the bottom one is closely connected with the implementation of the Grid. The system relies on a data base that allows the grid Input/Output coordination and the monitoring of the job status.

4 THE PROBLEM OF VISUALIZATION

The development of a complex protocol as the surface matching pipeline needs the development of a methodology able to view the results of the alignment. Consider for example a typical output of the recognition process: the result of the process is not a solution like “similar / not similar”. Like many other tools used in bioinformatics we have several indexes for evaluating surface similarity. It’s clear that an overlap of 100% suggests that surfaces are identical. But how can we evaluate different values? In this context, a method able to visualize results of alignments is very useful both in testing the application and in verifying the surfaces similarity. To accomplish this task, a reliable surface representation is needed: vertices and polygons have to be properly visualized in the three-dimensional space and the observer should look at the scene from a significant point of view to understand how surfaces are similar. Moreover, it is difficult for the observer to have a good idea of the three dimensional object represented by observing a static image. In this case the disposition of shadows and lights may give the observer the impression of a three-dimensional object. Even the opportunity to interact with the scene, changing the point of view or enlarging and reducing some details, or even the possibility of rotating the objects works in this direction. Therefore, an application for representing surfaces and their correlation in a graphic way can be very helpful. There is a lot of software available for molecular visualization [87],[88],[89], but none that can easily handle surface data or correlations between vertices. More precisely these programs have their internal surface representation and often we can’t export surface data for our analysis. Moreover, these types of software are not able to handle data generated by the matching process and we want to visualize surface and possibly to tag the vertices used for the alignment. In this way, we can’t use current software for molecular visualization, and we have to look for another solution. Clearly, rather than developing a new software for molecular visualization a better approach is to represent surface data in a well-established framework, which should provide enough flexibility to highlight the corresponding vertices on surfaces.

4.1 Surface data files

In the process of surface recognition there are mainly two types of files, the *.off* file and the *.corr* file. The first file is produced during the calculation of molecular surface with the program MSMS and contains information about surface. The *.corr* file is produced as output of the matching algorithm: when two surfaces are matched the correspondences between vertices are recorder in this data file, indicating the number of vertex on each surface and the correlation score between them. This file is then used to align surfaces and to generate a surface file for each aligned surface.

In the process of visualization these files have a fundamental role: while surface files are used for displaying surfaces, the *.corr* file is used to indicate which vertices support the alignment. If two surfaces are similar they must have correspondences with similar vertices. This evidence is useful for evaluating the similarity detection and the alignment itself.

4.2 The ray tracing solution

To address the problem of visualization, ray tracing techniques can represent molecular surfaces in a very informative way. Ray tracing techniques simulate the interaction of light with objects by following the path of each light ray from the viewer's eyes into the three dimensional space, and then generates an image like a photo. When a ray intersects an object, it has to be determined if that point is being lit by a light source, typically by tracing a ray from the point of intersection towards the light source. In this way surfaces are represented by vertex and polygons and these objects are placed in a three dimensional world. The user has to define a significant point of view and at least one light source. Then the algorithm simulates the interaction between object and light sources by following the path of each light ray from the camera towards the objects. The object can be visible if there is an intersection between a light ray and the viewer's ray (Figure 20)

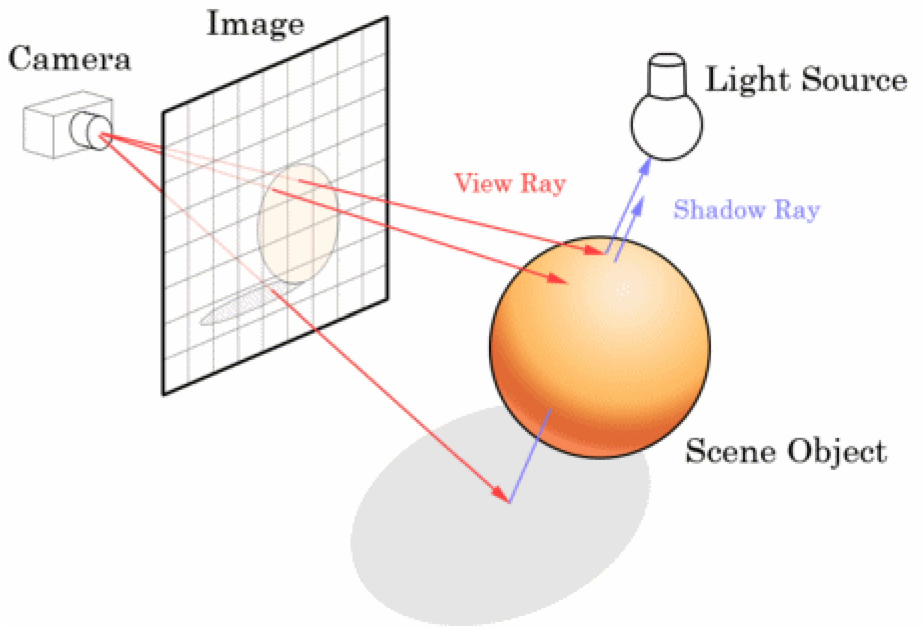


Figure 20. The ray tracing model

Nevertheless, ray tracing techniques are computationally intensive and it is difficult to define a good point of view focused on surface's similarity. Moreover, it is difficult for the observer to have a good idea of the three dimensional object represented since the images obtained are static and no interaction between the scene and the observer is possible, such as rotation or enlargement of the scene represented. These problems have been resolved by developing an application on top of the Visualization ToolKit (VTK), an open source system for 3D computer graphics and scientific data visualization [90].

4.3 The Visualization Toolkit (VTK)

The Visualization Toolkit is an open source collection of C++ class libraries with different interface layers, which allow users to exploit the fast function libraries implemented in C++ with the flexibility and the extensibility of a scripting language like python. It implements methods for rendering objects with ray tracing and to interact with the scene represented with mouse and keyboards, giving the user the possibility to better understand the three dimensional object by changing the point of view or by enlarging some details [91].

Simply stated, the visualization implies that data is properly transformed and represented. Transformation is the process of converting data from its original form into graphics primitives and computer images. Representation includes both the internal data structures used to store the data and the graphics primitives used to display data. The process of converting raw data in structures and their transformation in graphics primitives is usually referred as the visualization pipeline. In Figure 21 all the processes of transformation and representation are presented schematically. Each arrow in this figure is a transformation process, while circles are objects used for representing data. Surface's data as vertices and polygons are stored in VTK special object, while local correspondences between similar surfaces can be recorded to make textures on the surfaces. Afterwards objects have to be transformed in graphics primitives (the vtkPolyDataMapper) and finally an actor (vtkActor) can be generated, a key element in the visualization pipeline. Several vtkActors can be passed to the render window, where lights and shadows are calculated for each object in the scene. In the last step of the visualization pipeline, the render window object is provided to the user, enabling the possibility to interact with the scene and to modify the point of view of the world represented.

4.4 Implementation

On top of VTK, we implemented a system which creates classes and specific methods to read surface data and to manage objects in a transparent way. All process needed for data visualization are done automatically by a python interface layer, which implements a few classes and methods for converting raw data in vtkActors and changing attributes such as color or opacity. This solution simplifies access to the VTK complex structures, providing a more flexible and user friendly environment for dealing with surface data. This simplified collection of methods can be called directly by the python terminal or by a user developed script.

The implemented system provides a specific class to deal with each particular type of data involved in the surface analysis. Surface data is recorded in the OFF file format, a specific text file which defines vertices and connections which compose the polygonal mesh. Correlations between surfaces are stored in text files as vertices coordinates and values of corresponding correlations. The appropriate class for input data is automatically instantiated according to the provided file.

All the process for reading and processing the file, instantiating the VTK object and filling them with data are done automatically. A more generic class inherits methods from surface's and correlation's classes, and provides some additional methods to interact with the scene represented, like changing the color and the opacity of the object or highlighting and labeling the similar vertices directly on surfaces to improve the visibility of the scene and to better understand the surface similarity. The last class allows the user to instantiate the VTK render window and to represent the surface data as VTK actors, providing the possibility to interact with the scene represented by enlarging a detail or by changing the point of view. In this way, we can render surfaces data and tag similar vertices. There is a result of this representation in Figure 22.

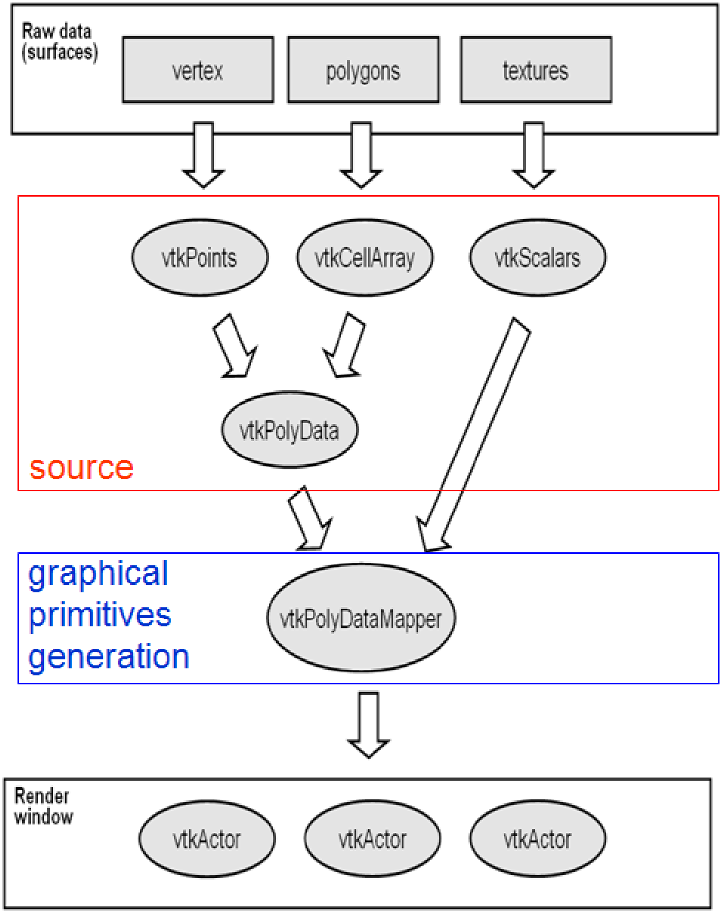


Figure 21. The visualization pipeline

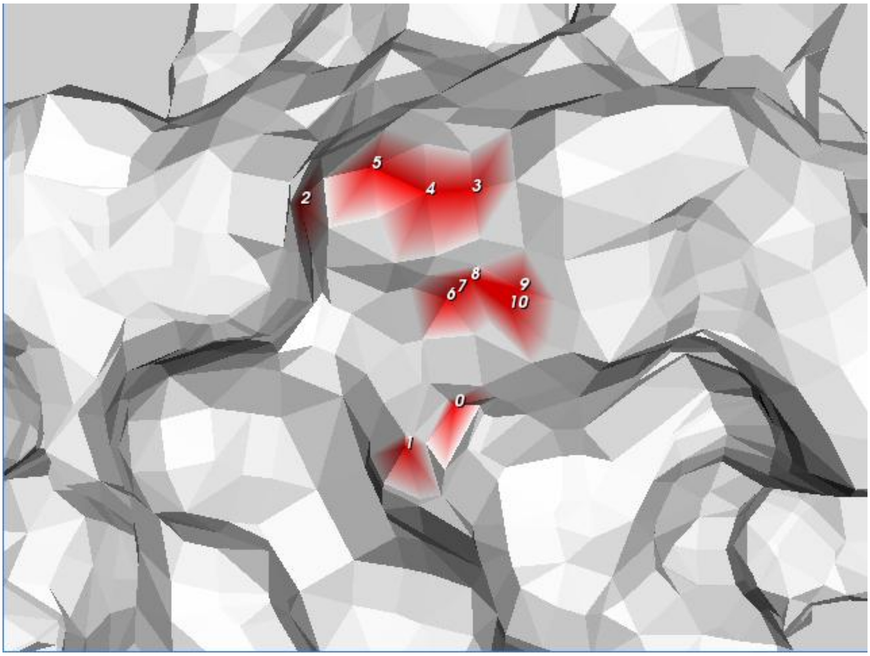


Figure 22. Surface representation with VTK

4.5 Applications

The methods implemented in this framework are particularly suitable for working on data generated by a correlation processes between surfaces. In particular, when two surfaces are matched, a data correlation file is produced which reports the correspondences between vertices of the two surfaces. This file can be processed by the system and the matching vertices can be highlighted on the surface. This type of representation facilitates the evaluation of results while searching for surface similarities. For example, we have studied the similarity between the surface of the lysozyme (PDB:1UIB) and a dataset of 1009 patches obtained by calculating the molecular surfaces in correspondence of the *Prosite* domains identified by running *Interproscan* [92] on the non redundant chain set of protein structures from the Protein Data Bank [74] that didn't include the query protein. The five highest ranking solutions found are reported in Table 1. The *pop* column shows how many vertices between two surfaces seem to be similar, the *overlap* column shows the percentage of vertices of the patch near to the vertices on protein surface, the *RMSD* column shows the distance between aligned vertices and the *vertices* column shows how many vertices supported the alignment. The first two solutions for the lysozime surface seem to have a good alignment, but inspecting results with the system implemented more detailed conclusions can be drawn. In detail, the first solution found is related to the prosite domain PS00128 obtained from a homologous protein (PDB:3LZT): this domain correctly identifies lactalbumine proteins and lysozime enzymes. Once similarity between surfaces has been computed, the user can evaluate the results of the alignment by parsing the data and building a vtkActor calling only a few methods of the system implemented, and all the operations needed for the visualization are done automatically.

Prosite_Name	Description	RMSD	Mean	Vertices	Overlap
PS00128	LACTALBUMINE_LYSOZIME	0.526	0.946	100	71.669
PS00206	TRANSFERRIN_2	1.350	0.917	10	55.483
PS00128	LACTALBUMINE_LYSOZIME	0.540	0.941	7	48.384
PS00790	RECEPTOR_TYR_KIN_V_1	1.000	0.942	7	46.059
PS00678	WD_REPEATS	1.668	0.951	10	44.021

Table 1. Top ranking solutions for the similarity analysis of the protein 1UIB (lysozime), sorted by the percentage of patch vertices closer to 1 ˚A from the protein vertices (Overlap column). For each solution the prosite identifier for the domain, the prosite description, the RMSD distance between aligned vertices, the mean value of correlation between vertex's images and the number of vertices found as similar is reported.

In Figure 23 we can see the results of the alignment. In the image on the left we can see the prosite domain and the vertices found as similar. In the image on the right we can see the two surfaces aligned, with the vertices of the protein in evidence. The transparent blue surface is the prosite domain, while the protein is represented in white. We can notice that the vertices on the patch surface are placed on similar position as on the protein surface. On the other hand, in Figure 24 we can see that the second solution is not as good as the first one, despite the fact that it has an appreciable overlap value.

Another example of the application of the Visualization Toolkit for visualizing molecular surfaces concerns an analysis of complementarities between two proteins. In this case, surfaces for protein thermitase (2TEC E) and its inhibitor englin C (2TEC I) have been calculated and compared searching for complementarities. The inhibitor was aligned on the enzyme's surface by similarities found on the surfaces. The results of the alignment were compared by

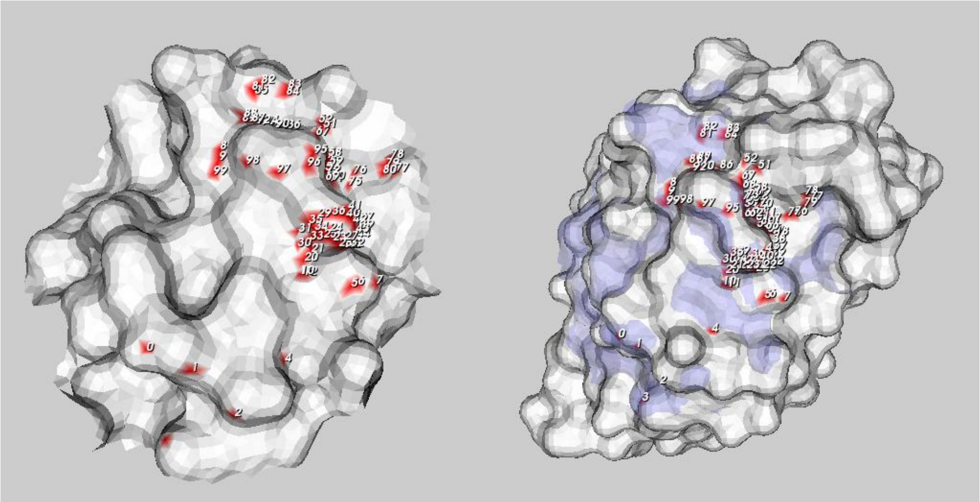


Figure 23. The alignment of prosite domain PS00128 on lysozime (PDB:1UIB)

measuring the root mean square deviation between the aligned surfaces and the surface computed from the crystallographic complex structure. The results of the best alignment are presented in Figure 25 on the left. On the right side of the same figure, we can see how the inhibitor is well aligned to the crystal form, and how this kind of presentation is useful to evaluate the results of the alignment.

In both cases, we can evaluate the results of the search for similarity by inspecting the aligned surfaces with this extension of Visualization Toolkit. This kind of visualization allows the graphical rendering of complex data, such as surface data and correlated vertices. The possibility to display the surfaces aligned allows the user to evaluate the quality of the similarity found by inspecting the alignment, and to check vertices tagged as correlated in order to discriminate bad solutions from the good ones. The user can interact with the scene represented by varying the opacity of surfaces in order to make them transparent to improve the visibility of the surfaces and obtaining a more informative representation of the scene.

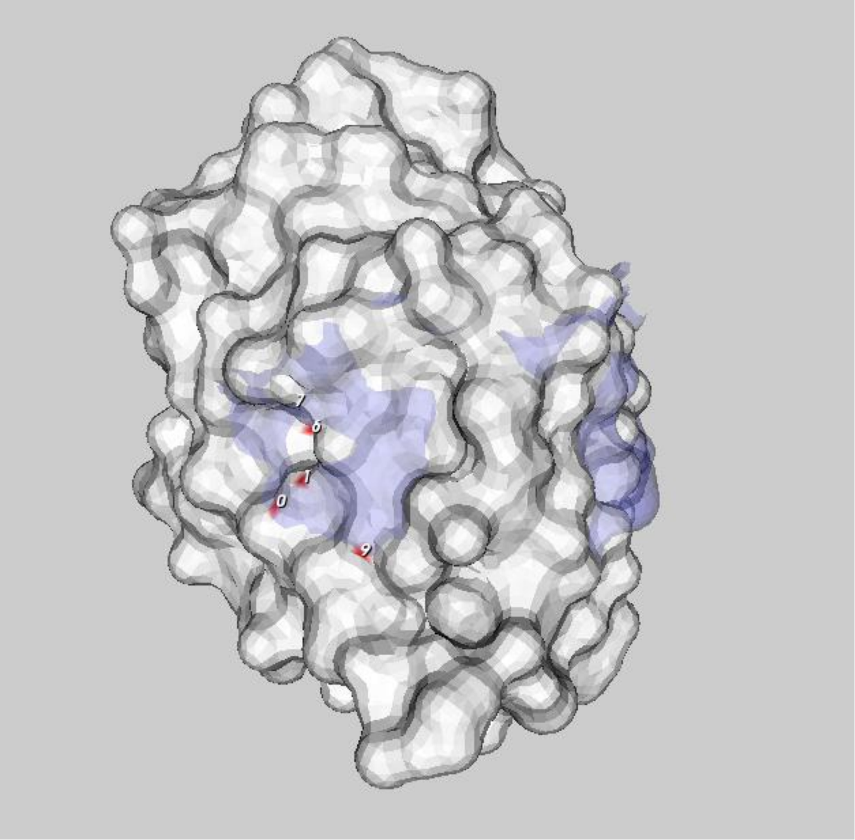


Figure 24. Here the alignment between surface of protein 1UIB (in white) and the prosite domain PS00206 is presented. By observing this alignment we can see that this solution is not as good as the first solution found (PS00128).

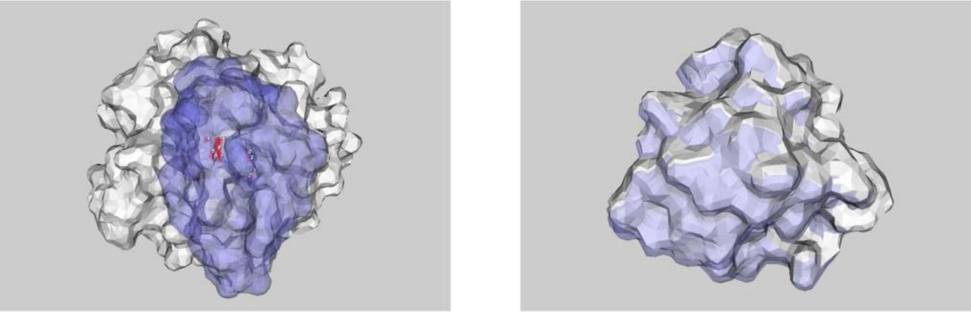


Figure 25. On the left side of this figure are presented the results of the alignment for the protein 2TEC E. and its inhibitor 2TEC I, with the vertices found as similar in evidence. On the right side are presented the aligned surface of the inhibitor (in blue) and the crystal surface of the inhibitor (in white)

5 THE HOMOLOGY IDENTIFICATION

Identifying similarities between protein surfaces can provide important insight into the functions of unknown proteins. In many cases, similarity in protein surface patterns can identify novel relationships in similar binding or catalysis that would go undetected when using traditional sequence or backbone structure comparisons. Similarity searches of this nature are usually more demanding in terms of computational time than matching to three-dimensional templates, but identifying similar surfaces between proteins can be useful for understanding protein function and annotating proteins with unknown biological roles.

Since the Protein Data Bank [74] contains more than 66,000 deposited structures, a comparison between one query protein and the whole database would be unthinkable, since comparing a target protein surface with all the known protein surfaces is computationally too expensive. We also have to consider that the number of structures will grow as technologies for resolving proteins structures are in continuous improvement. Moreover the Protein Data Bank contains a considerable redundancy of both sequence and structure: if we cluster sequences relying on their sequence similarity, we will find about 38,000 different groups. The reason is that in previous years the same structures have been determined several times by different research groups by varying experimental conditions and achieving structures with different resolution and quality. Furthermore, the same structures may be present in complexes, or proteins from different species may show a high degree of similarity, so we can consider only one of these structures.

In this way, it‘s better to reduce the cardinality of the initial dataset by identifying a subset of “representative” proteins, each one representing a cluster of proteins showing high surface similarity (and potentially even functional similarity). This approach is aimed at avoiding redundancy in comparing proteins, thus selecting as completely informative just the matches that involve the target protein and the representative proteins of each cluster. This can be achieved by clustering similar structures and identifying a representative protein for the whole cluster relying on protein properties such as the lowest percentage of residues with incomplete coordinate data or the best model resolution. The strategy needs to be reliable for collecting the representative proteins, since an improper choice would irreparably lead to the loss of part of the PDB elements. Indeed the choice of a bad representative may affect the similarity detection: for example, when the representative is an artifact, when a ligand crystallizes with a different orientation from that found in nature or when inorganic molecules in the crystallization solution are mistaken for cofactors.

Ideally, the correct approach is to obtain the molecular surfaces of all proteins and locate their functional domains. Then a surface capable of representing the same functional domains on all the other proteins has to be properly chosen. However, this solution seems to be impracticable because it would be too expensive in computational terms. This chapter describes two alternative methods by which it is possible to obtain a representative dataset of proteins. The first one is easy and is based on sequence similarity. The second one is more complex because it is based on structural similarity. At the end of this chapter the application of our algorithm with respect to the reference database will be discussed, showing that our software provides non trivial similarities relying on molecular surfaces.

5.1 The non-redundant PDB chain set of protein structures

A common choice in building a set of representatives can be to manually construct a limited set of proteins to be analyzed. By using a more systematic approach groups of proteins can be identified relying on sequence similarity and then a representative protein can be selected by means of structure quality. This solution has advantages and disadvantages: the division into groups is facilitated by the simplicity of sequence analysis because this approach avoids representing molecular surfaces or assessing fold similarity. However no one guarantees that a similar group of proteins have the same functional site.

Once we have grouped structures relying on sequence similarity, a representative of each group has to be chosen. In this process it is better to consider the quality of the three-dimensional structure of the protein. If the structure is not completely resolved or the degree of resolution is not sufficiently high, the risk is to obtain a surface that doesn't represent the group correctly.

To avoid manually creating groups relying on sequence similarity and identifying a representative for each group, we chose all the representative proteins suggested by the non redundant PDB-chain set of the NCBI structure [93], which divides all the protein structures available into the PDB in their constituent chains and then clusters all chains into groups relying on sequence similarity identified by Blast [12]. With this solution, the sequences that have similarity scores lower than a certain threshold are included in the same group. In this case, we chose to use the highest threshold in order to have a less redundant division. The sequences within a group are automatically sorted according to their quality and the integrity of the structure, rewarding structures with less unknown residues or with better resolution. The best structures are automatically identified as representative. Then a manual intervention decides whether to choose another representative of comparable quality, for example if the previous representative is a mutated form of another protein. All the proteins selected as representatives generate the non-redundant database, and all the other proteins belonging to the same groups are related to their representative.

Finding a representative for each group is not always possible: if we consider that this classification produces 11,272 groups, and only 10,932 groups have a representative because low quality structures produce groups without a representative. Moreover groups are populated with different densities. Most of the groups are composed of few members, while a few are heavily populated (Figure 26)

For each representative protein molecular surface has been calculated. All the data is recorded in a database that can be used in the search of similarity. In this way, we suppose that the surface of the representative protein can represent all the other proteins. This is not always true: in the case of multi domain proteins, the similarity between sequences doesn't cover the entire sequences, providing a representative surface which doesn't entirely represent the surface of its group. In Figure 27, we can see the surfaces of protein 1R4M chain B in blue with its representative 1JWB chain B in white. In Figure 28 we can see the sequence alignment between these two structures. For such cases, an approach based on fold similarity could be better.

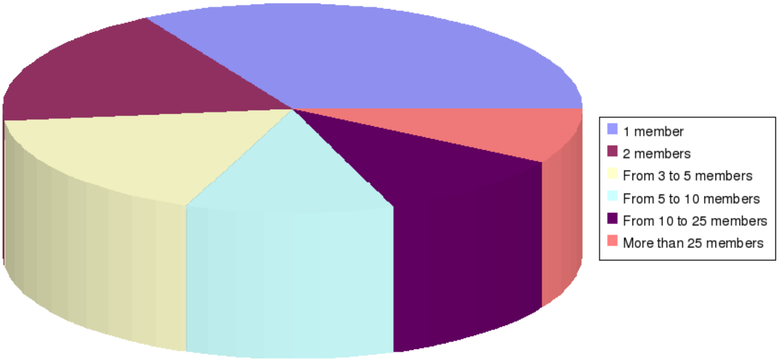


Figure 26. Groups are not populated equally; most of them are composed by few members

5.2 Functional domain identification

For all the representative chains suggested by NCBI, we performed functional domain analysis by running *InterProScan* [92] on the sequences identifying functional domains. This tool allows extending the search for domains on multiple databases simultaneously. In particular we focused our attention on the presence of sequences of specific patterns using *prosite* [17] and *PRINTS* databases [94] and on the presence of particular profiles using *Pfam* [15], *ProDom* [95] and *profile* [17] databases. With the help of InterProScan tool, the sequences are organized into families and for each family the presence of functional motifs or profiles in other database is reported. The biological function is also reported with its associated ontological terms. Results of the analysis can be obtained in various formats, for example the XML format, which is particularly useful since it is easily interpretable by computer programs.

The database information has been integrated by molecular cavities identification on each representative protein. Using the tool CASTp [96] it was possible to identify molecular cavities starting from the atomic coordinates of representative proteins. Particular attention was paid to the NMR structures in order to avoid subjecting the calculation of all models in the PDB file. The calculations of these types of structures were made only for the first model identified in the coordinate file. This choice was dictated by the fact that all models present in a NMR coordinates file are equally probable. The cavities computations were made using a probe sphere with a radius of 1.4 A°, the same radius used to calculate molecular surfaces.

In addition to information regarding the presence of domains and pockets, information concerning the presence of catalytic sites annotated in the database CSA [25] was added to the surface database. More precisely, the CSA database describes the amino-acids that take part in catalytic reaction. The annotation of these sites can be of different types, for example it can be based on literature data or experimental method, or otherwise it can be based on sequence analysis. In the last case, the alignment is manually inspected and is classified as catalytic site it there is a similarity with a known catalytic site in literature.

Once all the information concerning functional domains and catalytic sites is collected, we create a dataset of surface patches in which we correspond the surface of functional domains to the evidences collected. In this way we can correlate molecular surface regions with a precise functional annotation. This information can be used in functional annotation. In Figure 29 we can see the distribution of functional domains identified on representative proteins.

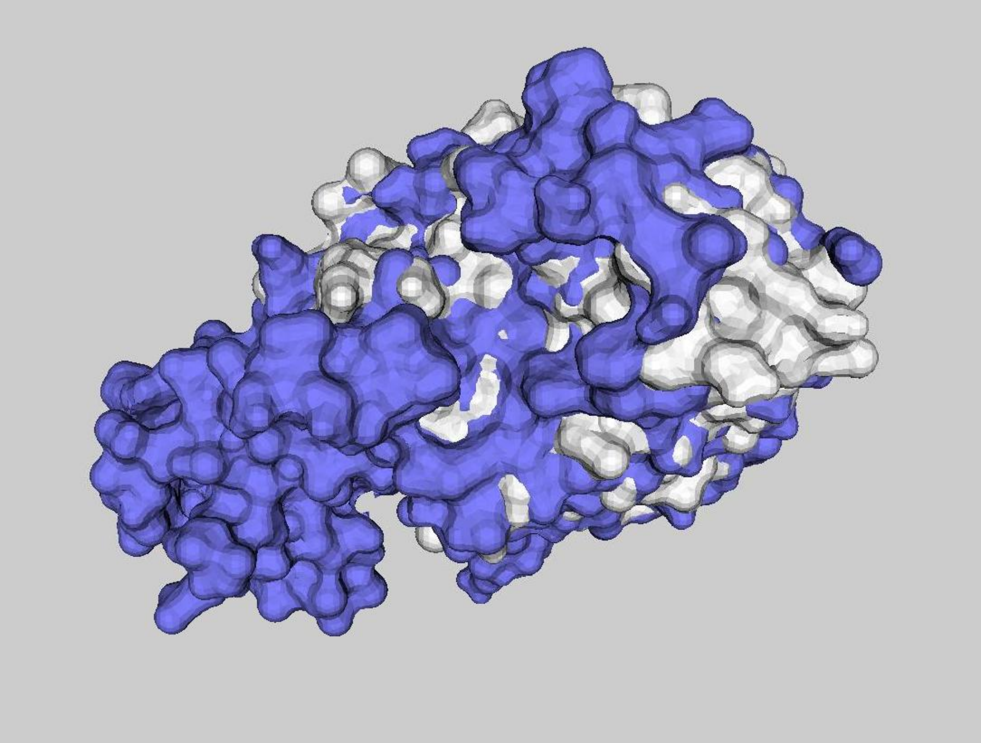
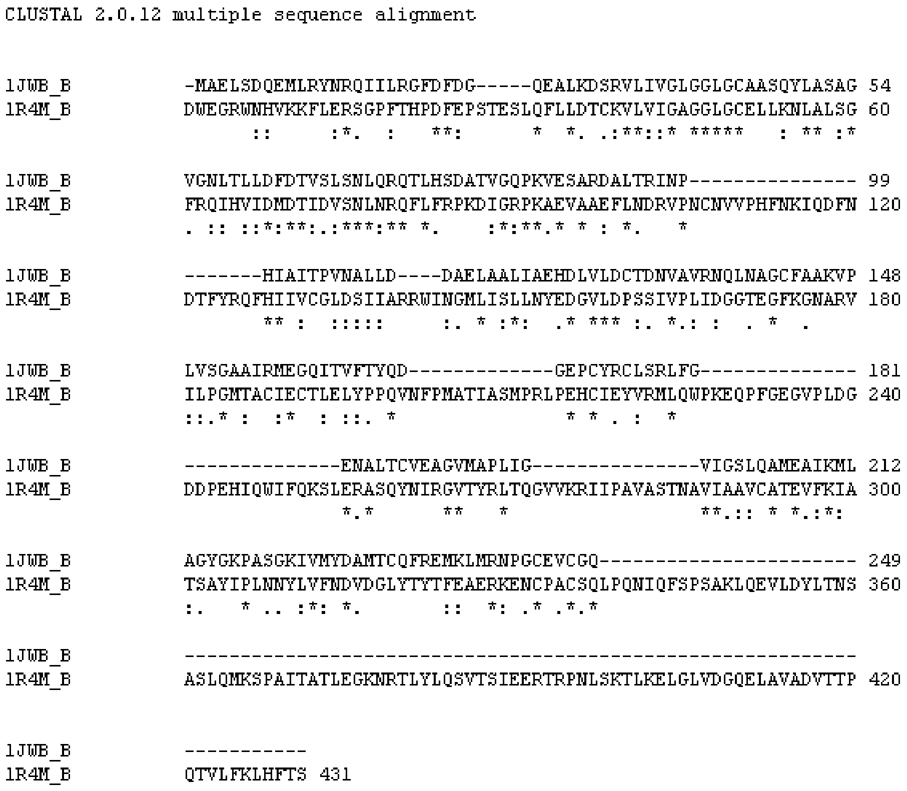


Figure 27. The surface of protein 1R4M chain B and its representative protein 1JWB chain B are presented respectively in blue and in white.

5.3 Structural classification of proteins

Although the sequence based approach is easier, an approach based on fold similarity is probably a better choice, since similar surfaces can be associated to similar folding structures. For this reason we propose an updated version of the developed database which relies on a semantic layer. This is a crucial innovation because it not only allows one to annotate proteins more accurately, presenting a more complete and integrated information set, but also to provide a knowledge-based layer on which the database entries can be suitably queried.



```
CLUSTAL 2.0.12 multiple sequence alignment


1JWB_B          -MAELSDQEMLRYNRQIILRGFDFDG-----QEALKDSRVLIVGLGGLGCAASQYLASAG 54
1R4M_B          DWEGRWNHVKKFLERSGPFTHPDFEPSTESLQFLLDTCKVLVIGAGGLGCELLKNLALSG 60
                      ::     :*.  :   **:      *  *. .:**::* *****   : ** :*

1JWB_B          VGNLTLLDFDTVSLSNLQRQTLHSDATVGQPKVESARDALTRINP--------------- 99
1R4M_B          FRQIHVIDMDTIDVSNLNRQFLFRPKDIGRPKAEVAAEFLNDRVPNCNVVPHFNKIQDFN 120
                . :: ::*:**:.:***:** *.    :*:**.* * : *.   *

1JWB_B          -------HIAITPVNALLD----DAELAALIAEHDLVLDCTDNVAVRNQLNAGCFAAKVP 148
1R4M_B          DTFYRQFHIIVCGLDSIIARRWINGMLISLLNYEDGVLDPSSIVPLIDGGTEGFKGNARV 180
                       ** :  :::::     :. * :*:  .* *** :. *.: :  . *  .

1JWB_B          LVSGAAIRMEGQITVFTYQD-------------GEPCYRCLSRLFG-------------- 181
1R4M_B          ILPGMTACIECTLELYPPQVNFPMATIASMPRLPEHCIEYVRMLQWPKEQPFGEGVPLDG 240
                ::.* :  :*  : ::. *               * * . :  *

1JWB_B          --------------ENALTCVEAGVMAPLIG---------------VIGSLQAMEAIKML 212
1R4M_B          DDPEHIQWIFQKSLERASQYNIRGVTYRLTQGVVKRIIPAVASTNAVIAAVCATEVFKIA 300
                              *.*      **   *                 **.:: * *.:*:

1JWB_B          AGYGKPASGKIVMYDAMTCQFREMKLMRNPGCEVCGQ----------------------- 249
1R4M_B          TSAYIPLNNYLVFNDVDGLYTYTFEAERKENCPACSQLPQNIQFSPSAKLQEVLDYLTNS 360
                :.   * .. :*: *.       ::  *: .* .*.*

1JWB_B          ------------------------------------------------------------
1R4M_B          ASLQMKSPAITATLEGKNRTLYLQSVTSIEERTRPNLSKTLKELGLVDGQELAVADVTTP 420


1JWB_B          -----------
1R4M_B          QTVLFKLHFTS 431
```

Figure 28. Sequence alignment of protein 1R4M chain B and its representative protein 1JWB chain B.

In a fold-based approach, protein structures are aligned together and clustered by fold similarity. This may guarantee that surfaces are more similar than with the sequence-based approach and that the cluster is homogeneous in structures and functions. By this type of classifications, structures are grouped together in a hierarchical way and a precise term is associated on each node of the classification tree. Done like this, all the proteins can be characterized by terms used for classification. So we propose an alternative strategy of representative selection by utilizing the semantic-knowledge associated on fold classifications. Such terms can be considered ontologies, since they are defined by a controlled vocabulary organized hierarchically: they represent a source of standardized and recognized descriptive terms associated with a set of relations among terms, crucial to infer links among objects that are annotated through ontologies. This solution permits the performance of knowledge-based queries and can be useful for annotations.

In the context of our system, both the hinted targets (the use of a shared vocabulary and the identification of relations between protein structures) are considered: the first was reached by means of the three Gene Ontology (GO) classifications (Cellular Component, Molecular Functions, Biological Process) [4], the second by using a protein-domain specific ontology, CATH [20]. The *Gene*

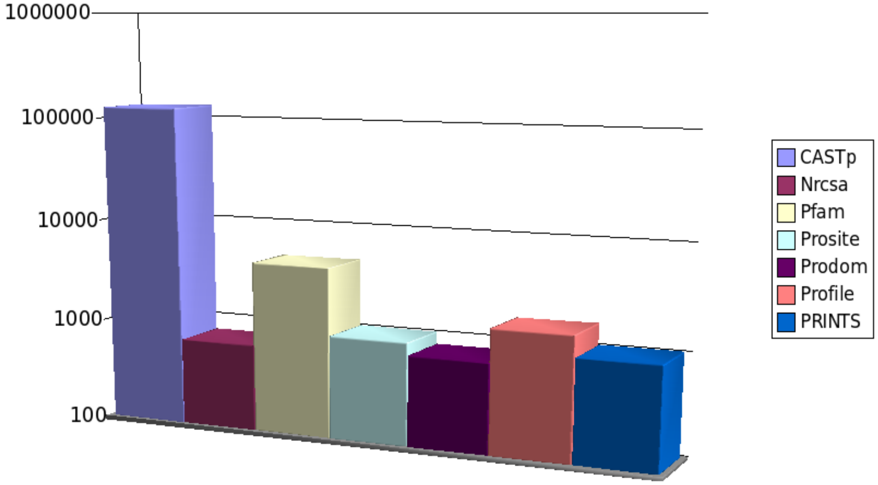


Figure 29. The distribution of functional domains on representative proteins

Ontology (GO) is the most widely used and multilevel ontology in the biomolecular domain. It collects genome and proteome related information in a graph-based hierarchical structure suitable for annotating and characterizing genes and proteins in respect to the molecular function (i.e. GO:0070402 : NADPH binding) and biological process they are involved in (i.e. GO:0055114 : oxidation reduction), and the spatial localization they present within a cell (i.e. GO:0043226 : organelle).

The identification of relations between proteins structures and the classification based on fold similarity is obtained from CATH classification. CATH is a structured vocabulary used for the classification of protein structures and aims to organize all the known protein domains recorded into PDB hierarchically (just those obtained by chopping proteins that present crystal structures produced to a resolution better than 4.0 angstroms or NMR structures). Multidomain protein structures are divided into their constituent domains. Classification hierarchy relies on four major levels: Class (C), Architecture (A), Topology (T) and Homologous superfamily (H). C-level is determined according to the secondary structure composition (alpha-helices and beta-sheets) and packing within the structure; A-level describes the overall shape of the domain structure as determined by the orientations of the secondary structures, regardless of the connectivity between the secondary structures; T-level is given according to whether they share the same topology or fold in the core of the domain (the same overall shape and connectivity of the secondary structures); H-level groups together protein domains which are thought to share a common ancestor and can therefore be described as homologous (determined by high sequence identity, more than 30%, or by structure comparison).

The identification of representative proteins can be performed through the semantic layer provided by the CATH ontology. Most of the PDB proteins are associated with domain ontology terms: thus, PDB proteins clusters can be obtained through the semantic level provided by CATH structured vocabulary. Multi-domain proteins are divided into their constituent domain, so the whole classification is performed on individual protein domains. The creation of the CATH domain is performed by aligning all protein domains with each other through the Statistical Signal and Array Processing (SSAP) program [97] and by classifying them into groups by fold similarity and evidence of homology. Protein structures are then classified using a combination of automated and manual procedures: if a given protein chain has a sufficiently high sequence identity and structural similarity (ie. 80% sequence identity, SSAP score >= 80) with a chain that has previously been chopped, the domain boundary assignment is performed automatically by inheriting the boundaries from the other chain, otherwise the domain boundaries are assigned manually.

Each level of annotation in CATH is identified by a number, which corresponds to a precise ontological term. In this way the ontological classification can be considered as a specific sequence of numbers (a code). Even in this case a unique representative protein for each group is chosen, on the basis of the parameters chosen for the identification of representative proteins indentified with sequence based methods. Considering our need to create protein clusters based on the topological aspect and the evolutionary relationship between proteins in the same group, we chose the representatives of the clusters belonging to the “H” ontology level. According to this type of classification, we retrieved about 2,400 representatives. From the computational biology point of view this approach, which relies on the semantic-driven folding-based clustering, is more biologically compliant compared to the sequence based approach. In fact, similar protein surfaces are associated to similar folding structures more than to similar primary structures. Therefore it is obvious that to obtain homogeneous clusters from protein surface point of view a folding-based approach can lead to better results. According to CATH classification, all the proteins belonging to the same group share a common fold: Figure 30 shows the backbone of the 51 proteins included in group 1.10.150.50.

Especially in the context of multi-domain proteins, where different folding structures coexist in the same protein chain, clustering by sequence matching can generate groups with low structure similarity, providing the identification of acceptable representatives just for few structures and not for the whole group. An example of this not uncommon condition is shown in Figure 31, where the representative protein (green) is largely different from the red protein, even though they belong to the same sequence-based group. Multi-domain proteins in CATH are treated using a performing strategy, by chomping all protein domains and considering them separately.

Even when using a folding-based approach, the obtained clusters might include some inhomogeneous protein structures and a CATH representative will probably not share its features with all the members of its group. But a marginal error is intrinsically involved in the approximation strategy aimed at reducing the comparison dataset during surface matching, in order to lower the computational load. Better similarities within group elements can be reached by going deeper in the CATH tree, but this will decrease the pruning effect that is necessary to improve system efficiency. Despite this intrinsic limit, the method of classification suggested by CATH implies that fold is conserved within the group, and at ’H’ level all the structures share a common ancestor and can be described as homologous. Considering CATH group “1.10.150.50”, in Table 2 we present a subset of proteins, which are selected as representatives of different clusters in the non-redundant PDB chain set. The similarity in their structures is evident in Figure 32.

The semantic-driven approach is more efficient when concerning the surface matching. Let us consider, for example, the group 1055 of the sequence similarity based dataset presented in Table 3. In this small group of three elements, 1D0Q chain A is the representative protein and the other elements are supposed to be sequence similar. According to the CATH classification, only two of these domains have the 3.90.580.10 classification-code (Alpha Beta. Alpha-Beta Complex. DNA Primase; Chain A. DNA Primase; Chain A) and must have similar fold. Assuming that similar fold involves similar surfaces, we can measure surface similarity by our

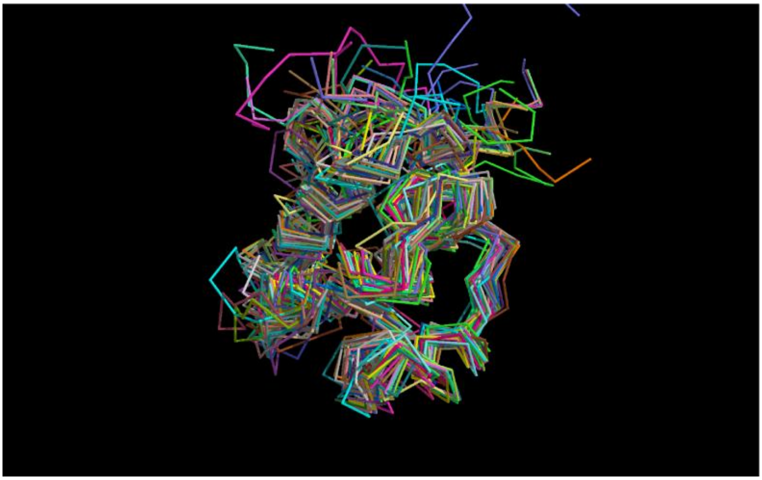


Figure 30. All the 51 structures sharing the 1.10.150.50 CATH domain.

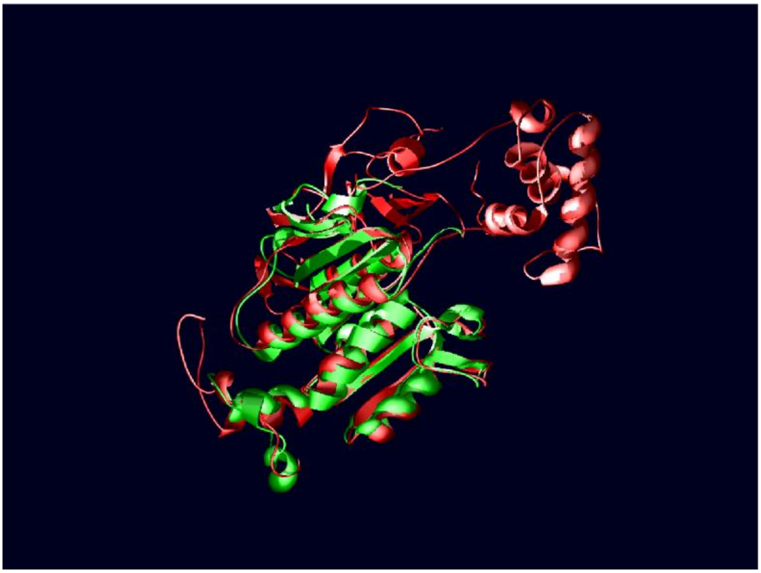
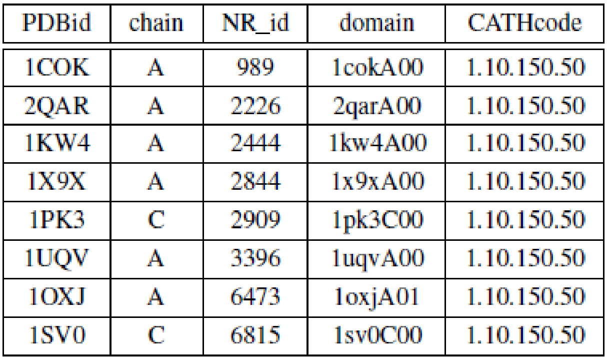


Figure 31. The representative protein 1JWB chain B (green) and the protein 1R4M chain B (red) belonging to the same cluster



PDBid	chain	NR_id	domain	CATHcode
1COK	A	989	1cokA00	1.10.150.50
2QAR	A	2226	2qarA00	1.10.150.50
1KW4	A	2444	1kw4A00	1.10.150.50
1X9X	A	2844	1x9xA00	1.10.150.50
1PK3	C	2909	1pk3C00	1.10.150.50
1UQV	A	3396	1uqvA00	1.10.150.50
1OXJ	A	6473	1oxjA01	1.10.150.50
1SV0	C	6815	1sv0C00	1.10.150.50

Table 2. Different sequence based representative sharing the same CATH domain

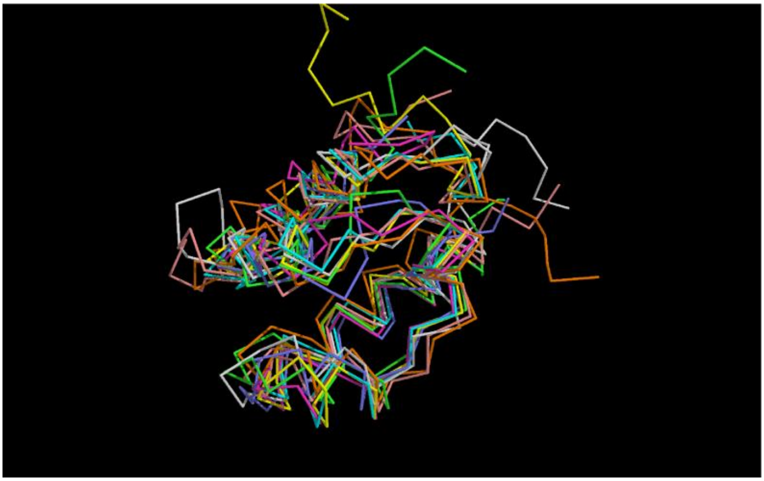
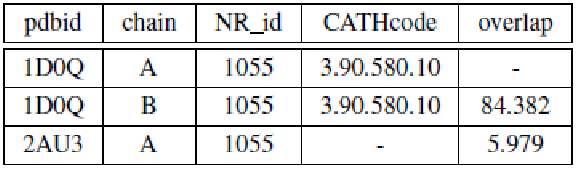


Figure 32. Different sequence based representative sharing the same CATH domain

system and evaluate how much the surface calculated from 1D0Q chain A is representative of the other surfaces. This can be evaluated by calculating the overlap index between the compared surfaces. The overlap expresses the percentage of surface points that are at a distance less than 1 °A from the other surface. By measuring this parameter, a great similarity arises between 1D0Q chain B and 1D0Q chain A (overlap about 80%), while alignment between 2AU3 chain A and 1D0Q chain A retrieves a very small score (about 5%). We can see how these surfaces are aligned by our system in Figure 33. Analyzing these results, we can see how the folding-base approach is more efficient in the selection of the representative.

5.4 Applications

In order to demonstrate the reliability of the proposed method, we have to validate our surface based identification approach. This could be achieved by aligning surfaces belonging to the same group with their representative surface: however, it



pdbid	chain	NR_id	CATHcode	overlap
1D0Q	A	1055	3.90.580.10	-
1D0Q	B	1055	3.90.580.10	84.382
2AU3	A	1055	-	5.979

Table 3. Three proteins belonging to the 1055 sequence-based group. Proteins which share the same CATH domain are more similar.

is assumed that this similarity is highly due to the criteria chosen in the identification of representative proteins, and this evidence is not enough to demonstrate the possibility of aligning surfaces with similar function even in proteins that are distant in the evolutionary tree.

Another solution in order to demonstrate the validity of the algorithm is to isolate surface patches in correspondence of functional domain on representative protein, in order to search these patches in proteins that are known to have the same domain. The identification of a strong alignment between surfaces will be the evidence that homologous proteins are correctly identified and that this method can be used for functional annotation. This idea in fact motivated the construction of the dataset of representative proteins and the extraction of surface patches in correspondence of their functional domains. However, before trying to annotate unknown proteins, it is necessary to demonstrate that known similarities are correctly identified.

To accomplish this task, we studied the similarity between 20 sample proteins randomly chosen from the Protein Data Bank not included in the representative dataset of protein structures and the dataset of surface patches created employing the prosite analysis on representative proteins. In our test we identified a total of 1,009 surface patches, which were collected in a relational database with all the information of the corresponding representative protein and the prosite domains detected. The proteins used for this test were chosen randomly preliminary discarding whether they have a prosite domain or not. More precisely, 7 proteins of

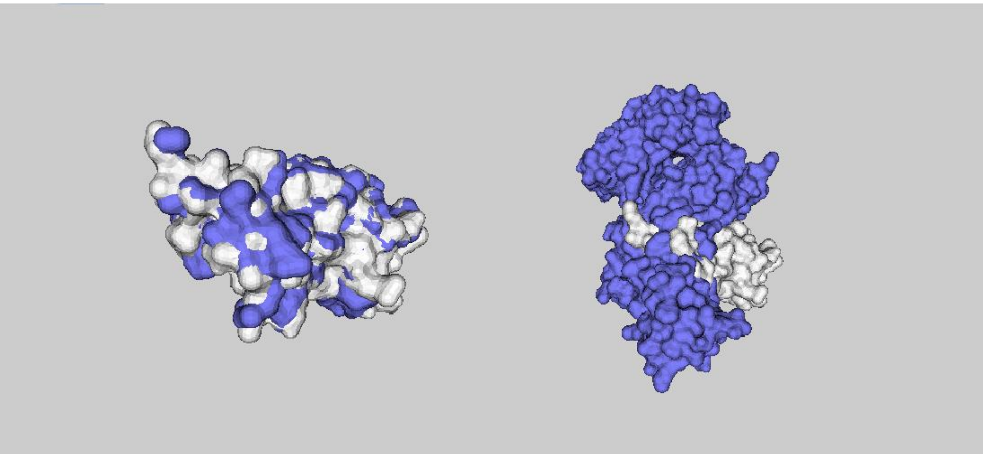


Figure 33. Protein 1D0Q chain B on the left and protein 2AU3 chain A on the right are presented in blue, while the representative protein 1D0Q chain A is presented on each side in white.

our dataset lacked of the prosite domain, while 13 proteins had a domain recorded in the patches database. The key idea was to investigate the correctness of the alignment identified and to study the results of the alignment also for the proteins without a prosite domain.

For each test case, we performed a simulation by searching for its surface against each of the 1,009 patches identified on all the representatives. The whole challenge was divided in 400 grid jobs in which we evaluated the surface similarity among one sample protein against 50 prosite patches. Each job was submitted to the grid infrastructure through the system described in paragraph 3.4.1. During 74 hours, the time between the submission of the first calculation and the retrieve of the last results, it was possible to perform about 931 hours of calculation on a single CPU. The results that came up were ranked by percentage of vertices between the matching surfaces that are closer to 1 A°, which can be considered a good measure of surface similarity.

For each simulation, we evaluated the linear correlation coefficient of the electrostatic potential between neighboring vertices of both surfaces once the alignment was established. This index can support the attribution of surface similarity when a strong correlation is found. The three best solutions for each simulation are reported in Table 4, while the list of correct functional domains present in proteins used for the whole test is reported in Table 5.

From a first preliminary analysis, we can see that for 10 of the 13 proteins with a prosite domain, it was possible to correctly identify the functional domain as the patch with better overlap than all the others found during this simulation. It is reasonable to assume that a functional domain must have certain geometrical characteristic, due to the precise spatial arrangement of functional residues that compose it. Being able to identify a good alignment between a patch which represent a functional domain of a protein and a protein with the same functional domain is a good indicator that this algorithm can be used to perform functional annotation.

Paying attention to overlap values, we realized that in some cases they tend to be very high, thus indicating a strong similarity in particular when there are a lot of vertices indentified as similar and when the surface of the patch is large. For example, in Figure 34 it is possible to observe a surface alignment between prosite domain PS01235 and 1EBA chain A where the overlap value is about 81%. This evidence suggests us that surfaces are very similar. Nevertheless, the overlap index doesn’t assume lower values in incorrect cases, and for such a reason it will be difficult to accurately assign an annotation relying only on this index. Regarding this consideration, we have to make two important assumptions. The first concerns the size of the patch: if a patch is small, it will be easily aligned with a good score. For example, in Figure 35 a prosite functional domain is aligned with a protein with no functional domains. Relying on this image, we have to say that the alignment is correct, but from a biological point of view it is reasonable to think that this match is negligible. Having the possibility to inspect the results visually, it becomes easier to judge the alignment, especially in cases where the similarity isn’t strong.

For what concerns the analysis of the linear correlation coefficient of electrostatic potential in these results, it seems that this parameter varies little between correct and incorrect cases. Indeed, patches correctly aligned haven’t a strong correlation, but rather tend to maintain lower values near zero. However, p-value tends to assume values near to 0 in the case of strong surface similarity. By observing the

PDB	Pros_Acc	Pros_Desc	RMSD	Vertices	overlap	Corr.	P-value	Note
1EBA_A	PS01352	HEMATOPO_REC_L_F1	0.575	522	81.44	0.3537	0.00E+000	Ok
1EBA_A	PS01352	HEMATOPO_REC_L_F1	0.584	86	81.17	0.3611	0.00E+000	Ok
1EBA_A	PS01352	HEMATOPO_REC_L_F1	0.496	475	81.077	0.3581	0.00E+000	Ok
1ESL	PS00022	EGF_1	0.988	131	58.245	0.1488	8.39E-012	Ok
1ESL	PS01186	EGF_2	0.988	131	58.245	0.1488	8.39E-012	Ok
1ESL	PS01248	EGF_LAM_1	0.24	63	51.41	0.2041	1.32E-025	Not Ok
1EU3_A	PS01322	PHOSPHOTRIESTERASE_1	1.092	46	50.662	0.2182	1.16E-010	Not Ok
1EU3_A	PS011295	ISPD	1.395	31	45.16	0.1849	1.04E-008	Not Ok
1EU3_A	PS00289	PENTAXIN	1.282	30	44.908	0.2310	1.01E-013	Not Ok
1F9Q_D	PS00200	RIESKE_2	1.74	42	57.413	0.1406	9.91E-005	Not Ok
1F9Q_D	PS00999	SSI	1.362	27	51.259	0.1360	9.05E-009	Not Ok
1F9Q_D	PS00471	SMALL_CYTOKINES_CXC	1.139	48	50.543	0.2650	0.00E+000	Ok
1GHV_H	PS00726	AP_NUCLEASE_F1_1	0.821	41	55.759	0.1305	4.32E-006	Not Ok
1GHV_H	PS00593	HEME_OXYGENASE	1.662	31	52.576	0.1531	3.64E-006	Not Ok
1GHV_H	PS00112	GUANIDO_KINASE	1.242	32	49.623	0.2460	1.04E-009	Not Ok
1H2D_B	PS00917	ASN_GLN_ASE_2	1.025	88	50.303	0.2222	3.99E-008	Not Ok
1H2D_B	PS00917	ASN_GLN_ASE_2	1.032	33	49.092	0.4027	1.99E-022	Not Ok
1H2D_B	PS00198	4FE4S_FERREDOXIN	0.896	27	47.945	0.1380	3.69E-006	Not Ok
1HBX_E	PS00350	MADS_BOX_1	0.668	272	88.366	0.4905	0.00E+000	Ok
1HBX_E	PS00350	MADS_BOX_1	0.681	359	81.063	0.4570	0.00E+000	Ok
1HBX_E	PS00350	MADS_BOX_1	0.824	49	75.481	0.4442	0.00E+000	Ok
1JSR_B	PS00917	ASN_GLN_ASE_2	0.494	165	95.073	0.5377	0.00E+000	Ok
1JSR_B	PS00917	ASN_GLN_ASE_2	0.606	197	94.728	0.5434	0.00E+000	Ok
1JSR_B	PS01182	GLYCOSYL_HYDROL_F35	1.084	30	56.941	0.2808	5.41E-022	Not Ok
1LUG_A	PS00162	ALPHA_CA_1	0.596	424	98.875	0.6643	0.00E+000	Ok
1LUG_A	PS00162	ALPHA_CA_1	0.563	574	98.33	0.6377	0.00E+000	Ok
1LUG_A	PS00162	ALPHA_CA_1	0.608	139	94.683	0.6517	0.00E+000	Ok
1NIB_A	PS00659	GLYCOSYL_HYDROL_F5	1.15	66	57.776	0.1935	2.32E-007	Not Ok
1NIB_A	PS01322	PHOSPHOTRIESTERASE_1	1.736	18	53.51	0.3094	3.21E-020	Not Ok
1NIB_A	PS00112	GUANIDO_KINASE	1.038	22	52.637	0.1564	3.77E-005	Not Ok
1RJC_B	PS00128	LACTALBUMIN_LYSOZYME	0.835	200	94.624	0.3027	0.00E+000	Ok
1RJC_B	PS00128	LACTALBUMIN_LYSOZYME	0.402	26	94.393	0.3524	0.00E+000	Ok
1RJC_B	PS00128	LACTALBUMIN_LYSOZYME	0.597	672	93.832	0.3592	0.00E+000	Ok
1T2B_A	PS00777	GLYCOSYL_HYDROL_F11_2	1.261	32	53.896	0.1570	1.38E-009	Not Ok
1T2B_A	PS00123	ALKALINE_PHOSPHATASE	1.138	50	53.783	0.3953	2.08E-037	Not Ok
1T2B_A	PS00777	GLYCOSYL_HYDROL_F11_2	0.832	75	51.258	0.2082	1.23E-014	Not Ok
1TED_D	PS00125	SER_THR_PHOSPHATASE	1.501	683	51.114	0.1744	3.34E-007	Not Ok
1TED_D	PS00659	GLYCOSYL_HYDROL_F5	0.598	59	51.008	0.3213	1.03E-016	Not Ok
1TED_D	PS00551	MOLYBDOPTERIN_PROK_1	0.991	38	47.643	0.1614	2.11E-007	Not Ok
1UIB	PS00128	LACTALBUMIN_LYSOZYME	0.594	165	73.087	0.2476	5.04E-035	Ok
1UIB	PS00128	LACTALBUMIN_LYSOZYME	0.676	269	71.867	0.2276	9.71E-30	Ok
1UIB	PS00128	LACTALBUMIN_LYSOZYME	0.687	40	43.633	0.1927	5.96E-017	Ok
1UKW_A	PS00072	ACYL_COA_DH_1	0.559	231	85.028	0.2099	4.65E-028	Ok
1UKW_A	PS00073	ACYL_COA_DH_2	0.635	215	79.947	0.1290	2.97E-016	Ok
1UKW_A	PS00073	ACYL_COA_DH_2	0.746	1275	79.727	0.1162	1.66E-013	Ok
1WVL_A	PS00205	TRANSFERRIN_1	0.578	16	47.556	0.2456	2.79E-021	Not Ok
1WVL_A	PS01121	CASPASE_HIS	0.953	36	45.049	0.1650	1.93E-013	Not Ok
1WVL_A	PS00374	MGMT	0.966	39	44.354	0.2747	3.11E-013	Not Ok
1XK4_F	PS00018	EF_HAND_1	0.781	174	61.783	0.4265	0.00E+000	Ok
1XK4_F	PS01027	GLYCOSYL_HYDROL_F39	1.452	66	59.661	0.2615	5.40E-014	Not Ok
1XK4_F	PS00018	EF_HAND_1	0.693	104	58.226	0.4013	0.00E+000	Ok
2BE5_C	PS00438	CATALASE_2	1.139	105	40.17	0.1945	4.30E-022	Not Ok
2BE5_C	PS01121	CASPASE_HIS	1.184	53	36.87	0.1571	5.93E-011	Not Ok
2BE5_C	PS01121	CASPASE_HIS	0.744	17	36.288	0.1608	8.88E-011	Not Ok
2MTA_H	PS00125	SER_THR_PHOSPHATASE	1.169	73	53.969	0.1413	1.87E-005	Not Ok
2MTA_H	PS00149	SULFATASE_2	1.596	92	52.987	0.4086	3.89E-039	Not Ok
2MTA_H	PS00810	ADP_GLC_PYROPHOSPH_3	1.743	23	50.921	0.1431	1.70E-007	Not Ok
2VHB_B	PS00142	ZINC_PROTEASE	1.373	86	60.554	0.2568	1.83E-014	Not Ok
2VHB_B	PS00153	ATPASE_GAMMA	0.866	135	59.471	0.1658	5.19E-014	Not Ok
2VHB_B	PS00142	ZINC_PROTEASE	1.259	201	57.958	0.1458	2.41E-005	Not Ok

Table 4. This table shows the three best solutions for each protein analyzed. For each solution is reported the prosite name and the prosite description of the patch identified as similar, the RMSD of the similar vertices after the alignment, the number of correspondences identified, the overlap between the two aligned surfaces, the correlation of the electrostatic potential and its relative P-value.

cases in which the overlap tends to higher values, we can notice how the linear correlation coefficient of the electrostatic potential tends to assume higher values while p-values tend to 0. For example, for the protein 1EBA chain A, where we can observe an overlap value close to 81%, we can see that the P-value of the linear correlation coefficient has a value of 0. In such cases, where the overlap is very strong, those values of P-values support the hypothesis of functional assignment by homology identification. Most likely when the electrostatic potential is identical, the molecular surface of the representative protein is very similar to that of other proteins belonging to the same group, so it is possible to indentify a high geometric similarity and good potential score. Conversely, lower potential scores do not allow discarding solutions unambiguously. For example, the correct alignment between PS00073 domain and protein 1UKW chain A has a lower correlation score than the wrong alignment between PS00123 domain and protein 1T2B chain A.

For cases in which the functional domain was not identified, we have to make some considerations. Concerning the number of functional domains, for this simulation we analyzed about 1,000 different domains whereas representative proteins are about 10,000. This means that it is not always possible to identify a functional prosite domain for each representative protein, and only a part of all possible functional domains are represented in this simulation: we cannot find a prosite domain if it isn't described in our dataset.

Another consideration concerns the division into groups and the choice of representative protein. In general terms, within a group there should be enough homogeneity determined by the criteria chosen to create the representative dataset. Indeed we chose to obtain the minimum number of representative proteins in order to contain the dimensionality of the dataset. This choice has the effect to limit the homogeneity within a group, since the number of representative proteins is very low, and this means that proteins within a group are different and are not very similar to their representative. For this reason the generation of the representative dataset can determine a bias in which patches were generated only for certain classes of proteins.

By inspecting the patches aligned with proteins with no prosite domains, we found that the aligned surface is small, while the rest of the surface is not aligned in the same way of the corrected aligned patches. The possibility of observing the results of these alignments allows us to discriminate the correct results from incorrect ones, while it is difficult to make the same consideration relying only on the values of certain parameters in each of these simulations. These values are not high as

PBD	Prosite_Acc	Prosite_Desc
1EBA_A	PS01352	HEMATOPO_REC_L_F1
1ESL	PS00022, PS00615, PS01186	EGF_1, C_TYPE_LECTIN_1, EGF_2
1EU3_A	PS00278	STAPH_STREP_TOXIN_2
1F9Q_D	PS00471	SMALL_CYTOKINES_CXC
1GHV_H	PS00134, PS00135	TRYPSIN_HIS, TRYPSIN_SER
1H2D_B		
1HBX_E	PS00350	MADS_BOX_1
1JSR_B	PS00144, PS00917	ASN_GLN_ASE_1, ASN_GLN_ASE_2
1LUG_A	PS00162	ALPHA_CA_1
1NIB_A		
1RJC_B	PS00128	LACTALBUMIN_LYSOZYME
1T2B_A		
1TED_D		
1UIB	PS00128	LACTALBUMIN_LYSOZYME
1UKW_A	PS00072, PS00073	ACYL_COA_DH_1, ACYL_COA_DH_2
1WVL_A		
1XK4_F	PS00018, PS00303	EF_HAND_1, S100_CABP
2BE5_C	PS01166	RNA_POL_BETA
2MTA_H		
2VHB_B		

Table 5. Prosite domains present in proteins used for the similarity test

the values assumed in correct cases, however the boundary between the patches properly aligned and the other is not as clear as we expected.

Regarding the sensibility and the specificity of the method, we can say that the method is sensible, since we can almost always find a correct alignment, but it seems to be not specific enough since we have some incorrect alignment even for proteins with no prosite domains. We can try to eliminate some incorrect alignments by filtering results on certain parameters. For example, the overlap is quite different in the case of correct alignments compared to incorrect alignments. Even the number of vertices found as similar is quite different between correct and incorrect cases. Nevertheless choosing fairly stringent criteria implies the loss of correct solutions in cases where evidence is not strong: it is always better to visually inspect the result of alignment in order to avoid making wrong decisions.

An important consideration is that we are able to align patches built on representative protein belonging to the same group of the protein of interest, but we are also able to align patches on their corresponding functional domain when the representative protein is different from the protein of interest. If we consider, for example the protein 1XK4 chain F and its representative protein 1HKA we can see that they have a very low percentage of sequence identity, due to their substantial differences. In this case, grouping structures by sequence similarity did not provide a good representative. Even the functional domains that can be identified on both sequences are different. Nevertheless, the functional domain PS00018 is properly aligned to the surface of 1XK4 chain F. This means that potentially patches built on

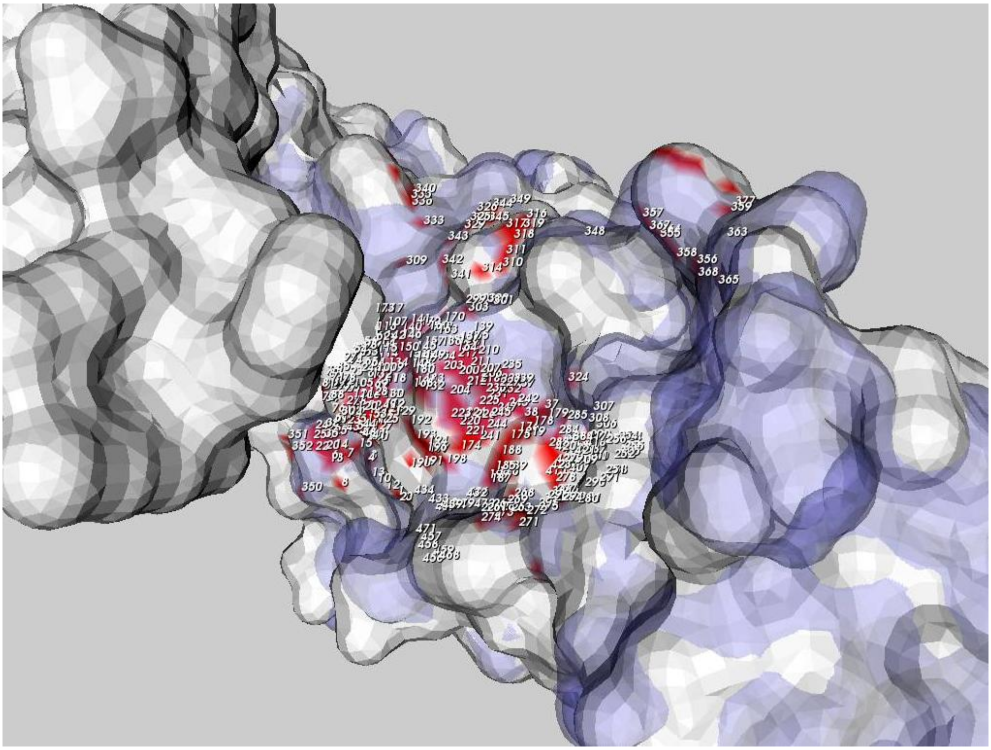


Figure 34. The domain PS01355 aligned correctly with protein 1EBA chain A.

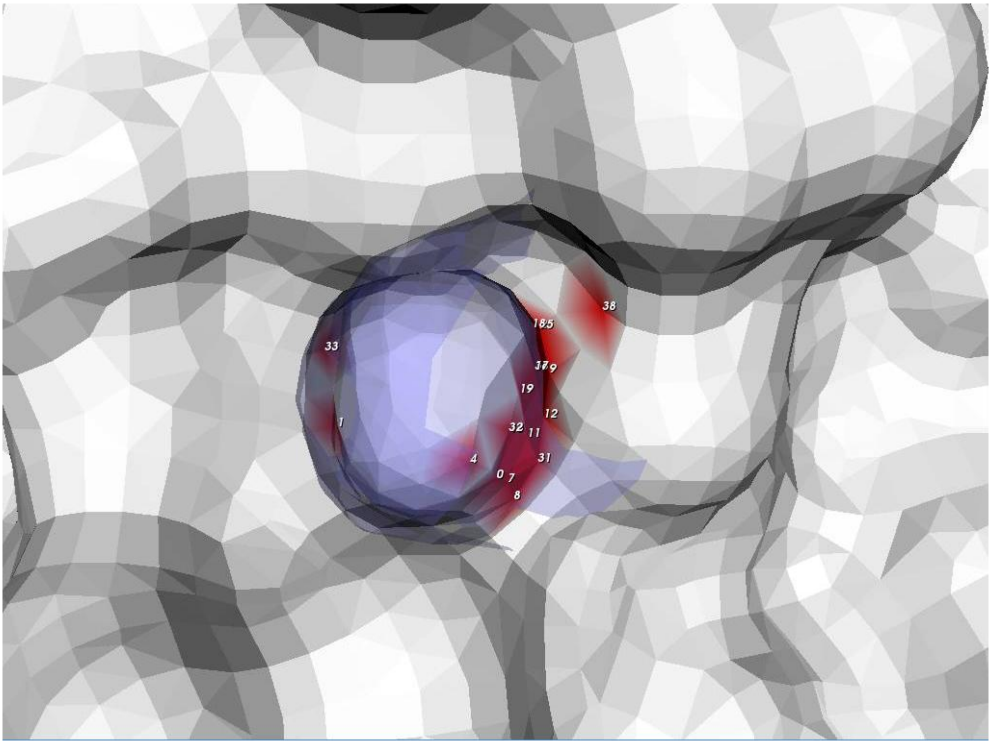


Figure 35. The domain PS00659 aligned with protein 1EBA chain A. This incorrect solution is detected through results visualization

representative proteins can effectively represent domains on protein surface of different groups. This consideration is important for two reasons: the first is that regardless of the representative protein we can identify the correct functional domains within a group. The second is that the surface obtained describes the surface of molecules similar to their representative but even different protein form a structural point of view. This consideration in particular supports the idea that the use a dataset of well characterized molecular surface allows us to annotate the other through the identification of similarities.

6 THE DOCKING PROBLEM

Protein-protein interactions are the most ubiquitous types of interactions in biological systems, and play a key role in all cellular processes, such as protein regulation and signal transduction. Determining the interaction network of whole organisms has therefore become a major theme of functional genomics and proteomics efforts. A common strategy to address the problem of protein docking is to minimize the global energy of the system, using different methods of optimization. These approaches are very time consuming because the calculation of the free energy is computationally intense and a huge number of spatial conformations must be considered.

As previously discussed, Camacho hypothesizes that protein interactions are driven at first by macromolecular surfaces and only in a second phase proteins physicochemical properties are involved through the formation of high affinity interactions by modifications of side-chains and backbone conformations [66].

In this context the analysis of molecular surfaces is interesting since they are the interface of interaction between proteins. They can be used for a first analysis of protein interactions in terms of surface complementariness, in particular to screen systems with low a priori knowledge of the involved components. This approach can be extremely useful to mime the first stage of the physical phenomenon, in particular to predict non-trivial interactions, and then there should be a phase of physicochemical analysis of the interacting system, but only on a small range of spatial conformations.

However it is clear that surface complementarity is very sensitive to small structural perturbations and hence such rigid methods may not work well for the realistic problem of docking crystallized proteins or protein models separately. In this chapter we'll take a look at the application of the image based algorithm for the analysis of surface complementarity.

6.1 Shape complementarity and smooth surfaces

Shape complementarity is an important feature of the interfaces of biological assemblies, such as protein–protein complexes. Over the years, the notion of shape complementarity has been confirmed by inspection of a large number of protein complexes in the Protein Data Bank. Although it does not represent a physical interaction, it is highly correlated with certain interaction energies, such as Van der Waals and non-polar desolvation. Thus, it has been widely used in protein–protein docking for searching and evaluating possible binding modes between two proteins. More precisely, it can be used to eliminate decoys with severe overlaps or few contacts at the interfaces.

Sanner and co-workers have analyzed surface complementarity in protein complexes and noticed that enzyme-inhibitor complexes have the highest shape complementarity while antibody-antigen complexes have the lowest [98]. This indicates that shape complementarity alone may not be sufficient in docking to identify the native binding modes of the complexes with poor shape complementarity. More comprehensive searching schemes and scoring functions considering other interactions, such as electrostatics, are definitely needed.

Moreover, the more realistic question is how well does shape complementarity perform when the structures of the two molecules to be docked have been determined separately, and consequently display surface variability. Surface side chains move, and inevitably, the extent of molecular surface complementarity is affected. This raises problems in docking proteins using their unbounded structures. However, rigid body docking avoids the use of flexible approaches that are inherently slower and generally examine far fewer docked conformations. Moreover, even considering the variability due to the formation of interactions, Norel and co-workers demonstrate that shape complementarity is determinant in binding even when the structures are in their unbound conformation [48].

When we talk about shape complementarity we should not imagine that surfaces get stuck together like puzzle piece, since they represent a model and the concept of complementarity can be considered more as a qualitative concept than something measurable, like we can see in Figure 36 on the left.

In order to alleviate the differences while searching for complementarities between bounded and unbounded proteins and local rearrangements of side chains, Sanner and co-workers adopted a strategy in which molecular surfaces are smoothed with a Gaussian Blur approach [98]. In their opinion smoother surfaces are more tolerant to small conformational changes. They demonstrated this hypothesis by studying surface complementarity of 66 protein-protein complexes using their Gaussian blur function.

For this reason, we adopted the strategy of smoothing surfaces in our docking studies in order to simplify the problem and to maximize the interaction area between complementary surfaces. On the right side of Figure 36 we present an example of the smoothed surface of the same complex reported on the left. Our algorithm is able to identify surface complementarities even with smoothed surfaces.

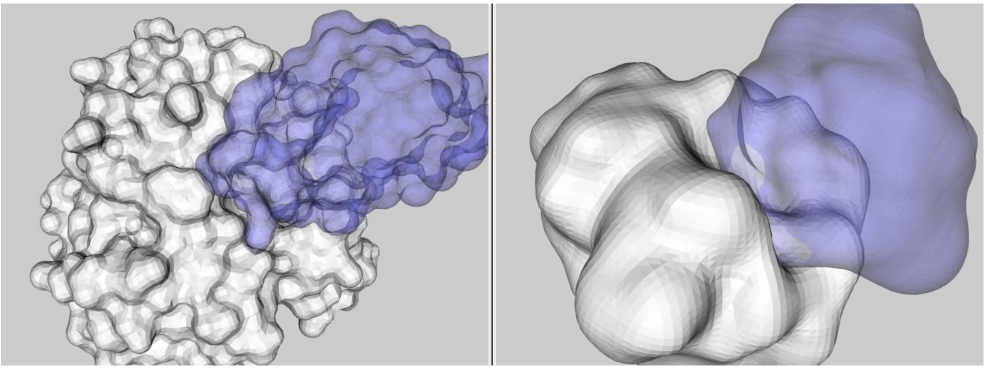


Figure 36. The complex between 2TEC chain I presented in blue and 2TEC chain A presented in white are presented in both sides of this figure. In particoular on the left side we see the "standard" surfaces, while on the right side we see the "smoothed" surfaces

6.2 The electrostatic potential evaluation

According to Gruber and co-workers, protein-protein interactions are determined by the burial of the hydrophobic surfaces, where electrostatics have been shown to play a key role in determining specificity and, in some cases, the thermodynamics and kinetics of macromolecular association [39]. McCoy and co-workers also demonstrate that at the protein interface there is a sort of electrostatic and charge complementarity between interacting proteins [44]. This behavior can be justified by the fact that a complex should be energetically stable, and for this reason the electrostatic complementarity, combined with hydrophobic burial, allows the establishment of favorable interactions and the stability of the complex.

According to this consideration, in order to simplify the energy evaluation of a complex obtained by the aligning two surfaces through the identification of a geometric complementarity computed using our algorithm, we have adopted the same approach adopted by Gruber in order to evaluate the electrostatic complementarity. More precisely, for a complex of known structures containing proteins A and B, two potential maps are calculated using the APBS program [80]. The first corresponds to the potential that would prevail with a solvent envelope defined by the A+B complex, but only with the atoms of protein A charged (Figure 37 on the left), while the second corresponds to the potential that would prevail with a solvent envelope defined by the complex but with only atoms of protein B charged (Figure 37 on the right). Then we determine the interface vertex on surface A, which are the vertices closer than 1 A° to the vertices of surface B through an approach similar to the evaluation of overlap values (see paragraph 2.4). Subsequently, two potentials are associated with each vertex interface vertices of protein A: one derived from interpolating surface vertex positions into the first potential map and the other from interpolating into the second potential map. The electrostatic complementarity of the interface can then be evaluated by calculating the linear correlation coefficient of the two different potentials over the whole set of surface vertices that define the interface. This approach in particular allows us to evaluate a couple of potential values on each interface vertices previously determined.

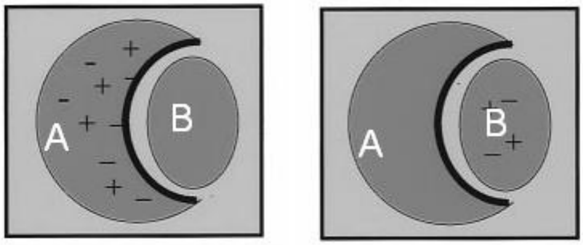


Figure 37. The potential evaluation for a protein complex. The first potential is evaluated for the charged protein A buried by the un-charged protein B. The second is evaluated for the charged protein B buried by the uncharged protein A.

6.3 The Chen dataset

To test our system in relation to the search for surface complementarities for predicting possible interactions, we accomplished a test using a well-known non-redundant data set of bounded and unbounded protein complexes. In particular, we performed a wide range of tests on the Chen [81] data set of protein-protein interactions, consisting of co-crystallized structures which are widely used to test software for interaction analysis. The objective was to verify if the system was able to identify complementarities in fully described protein-protein interactions, whilst completely discarding any a priori knowledge.

In details, the Chen dataset contain 59 test cases: 22 enzyme-inhibitor complexes, 19 antibody-antigen complexes, 11 other complexes, and 7 difficult test cases. Among them, there are 31 unbound-unbound and 28 unbound-bound test cases. Among the unbound-unbound test cases, 16 are enzyme-inhibitor, 5 antibody-antigen, 5 others, and 5 difficult. The seven difficult test cases have significant conformational change for more than half of the interface backbone residues; therefore, they are suitable for testing docking algorithms that explicitly perform backbone conformational search. The remaining test cases should be amenable to rigid body docking algorithms with some consideration of flexibility.

The test was performed by splitting the co-crystallized structures and trying to re-identify the interaction surfaces using our algorithm, working without using any a priori biological knowledge of the system. Usually in rigid docking tests unbounded structures are aligned, in order to assess the behavior of the programs in handling conformational change due to interactions between proteins. In our case, however, bounded structures were used because our aim was to study our algorithm's ability to detect geometric complementarity. For each structure the smoothed surface was calculated in order to minimize the differences in surface complementarity between structures and the electrostatic potential has been evaluated for each hypothetical complex obtained by our analysis. Molecular surfaces were randomly rototraslated in space to ensure that the complexes were really produced by the algorithm for the search of similarities.

6.4 Docking simulations

For each simulation different matches were identified and among them the best match was selected according to the effective conformation of the co-crystallized, experimentally validated, structures. The quality of results was assessed by the root mean square distance (RMSD) of C atoms after superposition of the best prediction onto the co-crystallized protein complex.

The graph in Figure 38 shows the distribution of the data set of the RMSD between the structures as conformed in the co-crystallized experimental data and those provided by our algorithm. The data is ranked from the lowest to the highest RMSD and a comparison with ZDOCK [70] and Rosetta [71] predictions are provided (for a short description of these two programs, see paragraph 1.6.1). According to the evaluation criteria proposed in the CAPRI experience [82], which identifies a threshold of 10 °A for suitable results, in half of our simulations we were able to find good solutions. A detailed output of our simulations is reported in Table 6.

The matching algorithm relies on a pipeline of filtering and clustering procedures that are inevitably characterized by parameters that should be carefully tuned to obtain good results. In paragraph 3.1 I described how the different stages of the algorithm require increasing calculation time and how parameters have to be set accordingly to the desired compromise between performance and quality of results. Simply stated, the possibility of achieving good results, in particular while analyzing difficult systems is inversely proportional to the computational time, a suitable trade off should be found for each case study. On the other hand, there are other parameters that have great impact on the quality of results, because they influence the image based representation according to the topology and the size of the surface, but have little influence on the computational time. The problem is clear: the reliability of this approach relies in a correct identification of parameters which affect the quality of results and which permits to find a solution in an acceptable time.

In this test we used a parameter configuration detected using the genetic algorithm by performing the alignment on a limited set of complexes (see paragraph 3.3). However in this phase of evaluation, it is difficult to obtain a parameter configuration capable of optimizing the alignment on all the structures. The reason for this can be the fact that the geometric complementarity depends on the type of the complex: enzyme-inhibitor complexes have the highest shape complementarity while antibody-antigen complexes have the lowest [98]. Regarding this consideration, there may be parameter configurations specific for the class type of the complex.

Ideally, a more sophisticated method than Genetic Algorithm could achieve better results, which could be a method that is able to constrain the problem of optimization relying on the achievement of the optimum solution, since genetic algorithm can explore different conformation without moving towards the solution of the problem. For now, we decided to use genetic algorithms since they provide a solution easier to implement.

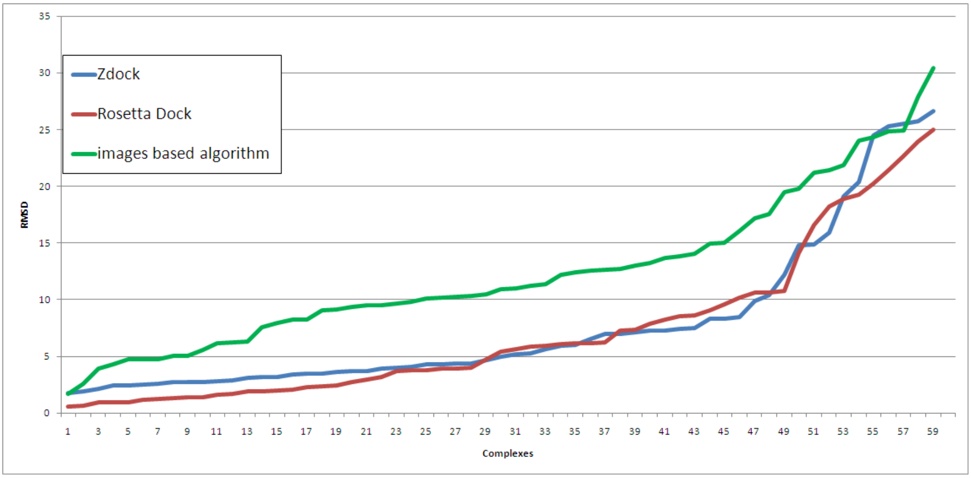


Figure 38. The graph shows the distribution on the data set of the RMSD, calculated using the $C\alpha$ atoms positions, between the real experimental conformation and the best prediction provided by our software, in comparison with two of its competitors

simu ation	rmsd	correlation	p-value	compenetration	type
05-1CGI	1.716493	-0.109352659	0.020753455	4 287652493	Enzyme-inhibitor
17-1PPE	2.563398	-0.132283271	0.010754516	7.542067051	Enzyme-inhibitor
16-2SNI	3.914753	-0.284262174	1.3128E-07	4.450510502	Enzyme-inhibitor
20-1UDI	4.312239	-0.031154927	0.480517115	9.115504265	Enzyme-inhibitor
21-2TEC	4.722694	-0.133594332	0.015472698	5.601659775	Enzyme-inhibitor
19-1TAB	4.736288	-0.06975787	0.136491798	7.498036861	Enzyme-inhibitor
13-2KAI	4.769309	-0.434670708	4.21307E-14	3.471849918	Enzyme-inhibitor
28-1BQL	5.010163	0.117634803	0.083832729	2.126300335	Antibody-antigen
40-2JEL	5.072633	0.052765637	0.397740016	3.677478313	Antibody-antigen
42-1AVZ	5.534021	0.031979427	0.552128147	8.594730377	Others
08-1DFJ	6.133186	-0.070038679	0.145206467	7.439703465	Enzyme-inhibitor
15-2SIC	6.215697	-0.237287373	4.74934E-05	2.065095425	Enzyme-inhibitor
03-1BRC	6.274498	-0.240351067	3.116E-05	7.514063835	Enzyme-inhibitor
53-1BTH	7.583223	-0.003824575	0.940529765	17.02717018	Difficult test case
11-1TGS	7.940306	0.139981692	0.00089523	12.37454128	Enzyme-inhibitor
04-1BRS	8.209125	-0.16412066	0.000555741	4.126440525	Enzyme-inhibitor
23-1AHW	8.242391	0.005484388	0.928926262	3.467393398	Antibody-antigen
18-1STF	9.063947	0.009960213	0.841225087	11.51950264	Enzyme-inhibitor
44-1WQ1	9.122843	0.037806761	0.534681507	5.00927639	Others
01-1ACB	9.346735	-0.074630454	0.124054802	12.24332047	Enzyme-inhibitor
02-1AVW	9.48625	-0.260252488	7.93513E-06	2.016129017	Enzyme-inhibitor
22-4HTC	9.52024	-0.092922647	0.02317185	15.60871601	Enzyme-inhibitor
09-1FSS	9.661633	0.011447731	0.878099174	11.94308376	Enzyme-inhibitor
14-2PTC	9.778445	-0.007243589	0.906220236	3.502415419	Enzyme-inhibitor
52-2BTF	10.12305	-0.107083132	0.0134666	13.507864	Others
33-1KXQ	10.14528	0.163062923	0.000162992	15.0626297	Antibody-antigen
47-1A0O	10.25883	0.022490688	0.682594696	5.075493336	Others
07-1CSE	10.32123	-0.083292465	0.177251487	8.12709713	Enzyme-inhibitor
10-1MAH	10.49144	-0.249112516	3.97847E-05	16.34458542	Enzyme-inhibitor
36-1MEL	10.93443	0.036326345	0.357003834	6.404129505	Antibody-antigen
46-2PCC	11.03179	-0.162597152	0.003752554	3.122704029	Others
12-1UGH	11.2374	-0.056456203	0.197381792	15.70082188	Enzyme-inhibitor
41-2VIR	11.36129	-0.03207	0.275024	2.52863431	Antibody-antigen
27-1WEJ	12.15638	-0.205634097	0.000140493	4.031299591	Antibody-antigen
45-2MTA	12.38553	0.065601366	0.168100492	12.67101002	Others
35-1KXV	12.56591	-0.019149901	0.728901616	15.3504343	Antibody-antigen
24-1BVK	12.66525	-0.055563791	0.520551719	2.488601685	Antibody-antigen
26-1MLC	12.67804	0.019011171	0.801692074	8.117909431	Antibody-antigen
57-1KKL	13.02322	-0.298499401	2.24964E-05	13.72718334	Difficult test case
50-1IGC	13.22686	-0.044834836	0.499611622	21.29512787	Others
34-1KXT	13.70471	0.308579306	1.14885E-08	14.69654846	Antibody-antigen
32-1JHL	13.81978	0.166306888	0.002553984	12.2737999	Antibody-antigen
55-1FQ1	14.06578	0.097099609	0.078175274	4.885774612	Difficult test case
59-3HHR	14.904	0.038203559	0.440420665	11.30118752	Difficult test case
30-1FBI	15.03494	0.112093372	0.034243654	26.24033165	Antibody-antigen
38-1NMB	16.07531	0.149626198	0.021765987	1.372578979	Antibody-antigen
51-1SPB	17.1505	0.075637265	0.174414479	8.704061508	Others
29-1EO8	17.55028	-0.082029408	0.218266536	5.239677429	Antibody-antigen
31-1IAI	19.47273	-0.027647092	0.643273966	2.956038237	Antibody-antigen
49-1GLA	19.81571	0.04395087	0.383062166	16.24264145	Others
54-1FIN	21.20693	0.058537053	0.179672337	11.10006619	Difficult test case
39-1QFU	21.43527	0.209291249	0.000226884	5.284872055	Antibody-antigen
25-1DQJ	21.86836	-0.071322122	0.128736545	29.10539246	Antibody-antigen
48-1ATN	24.00849	-0.062497791	0.086982643	14.24047947	Others
58-1EFU	24.35507	0.031150347	0.520899795	8.913427353	Difficult test case
43-1L0Y	24.81022	-0.008467666	0.881386666	6.389776707	Others
56-1GOT	24.90108	0.003036348	0.951788585	5.273097038	Difficult test case
37-1NCA	27.92505	0.10312822	0.030743364	11.43241978	Antibody-antigen
06-1CHO	30.44303	-0.090153097	0.0027897	5.813079357	Enzyme-inhibitor

Table 6. This table reports the tests performed by our algorithm using the Chen dataset . RMSD. Compenetration values. correlation and the class of the complex are reported.

By analyzing the graph trend for each of the three programs used for detecting the best complex conformation, we have to notice how it is generally difficult to identify good solutions in all cases used for each simulation. Our algorithm has a comparable performance with the other two programs; although the other presented algorithms consider more aspects than the mere geometrical similarity, such as physic-chemical characteristic, and so they have a more effective function, able to assess different conformations during the calculation phase. In our case we evaluated the electrostatic complementarity and the compenetration only when complexes are produced: taking these aspects in consideration during the calculation phase can lead to better results.

A critical consideration that we can make is that the system is more effective in evaluating patch similarity than surface complementarity. This can be determined by the fact that two complementary surfaces are not completely similar.

One consequence of such low similarity is that evidences are weaker: fewer matches are identified in search for complementarity than in search for similarity and for such reason alignment may be incorrect. In fact, in the case of a large group of correspondences, an incorrect match will not significantly affect the alignment. Vice versa, in case of a small group an incorrect match will affect the alignment and for such reason the complex can be oriented differently compared to a biological solution.

In conclusion, we adapted the algorithm of search of surface similarities in order to identify the complementarity between different proteins, and on the basis of this evidence we tried to rebuild the complex. Certainly, results are not significantly better than others produced by other programs that are currently available, but considering that the algorithm doesn’t consider energy but only geometry, we can conclude that the behavior is comparable to other programs. Therefore considering that our algorithm works at this point without any energetic minimization relying in chemical-physical considerations, we can conclude that we implemented a pretty good solution for a first fast screening of the possible interactions. The objective of this program is to find a favorable solution that can be a starting point for further analysis in which the orientation of side chains can vary in order to stabilize the complex formation.

7 CONCLUSIONS

During this thesis an algorithm for functional annotation and protein-protein docking problem has been presented. The algorithm has been studied in detail and optimized in order to improve efficiency and reduce calculation times, especially through the use of heuristics neighbor during the clustering of the correspondences which support the hypothesis of similarity on two molecular surfaces.

The problem of parameter estimation and how it was possible to address this problem using genetic algorithms was tackled. Indeed the problem was represented through a population of individuals obtained from different configurations of parameters. A fitness function has been applied in order to indentify parameter configurations which maximize the identification of similarities. Through the implementation of a parallel version of this algorithm, it was possible to evaluate simultaneously more than one parameter configuration, thereby reducing the time required to identify an optimal conformation. A platform was developed that is able to perform all the calculations required using the resources provided by the grid platform, in order to do all the calculations needed, without overwhelming local resources.

The implementation of a system for result displaying has been discussed, which provides an easy tool to visually inspect the alignment quality, allowing the user to interact with the represented structures and have a more concrete idea of the similarity between different surfaces.

Two different methods for reducing the dimensionality of three-dimensional structures through the identification of representative protein relying on sequence or structure similarity were proposed. Despite the fact that the first methodology is easier to implement, the dataset derived from fold similarity seem to have better effects in the identification of representative proteins whose surfaces may resemble other surface within the same group. Therefore we demonstrated how the method is effective in identifying surface similarities that can be used in assigning function through homology. To accomplish this task, surfaces of representative proteins were determined and functional patches were extracted in correspondence of functional domains. These patches were used to build a dataset of functional surfaces in order to annotate a sample of known proteins. The results demonstrate that if the evidence is good, then surfaces are similar. On the other hand, to exclude cases where similarities are low, it is necessary to visually inspect the results. Such patch analysis has demonstrated how the method is valid in identifying similarities between homologous proteins. Another interesting application is to describe the function of molecular cavities, and to derive from them a database that can be used for functional annotation rather than drug design studies. This particular aspect will be very interesting since this kind of considerations made on molecular surfaces is unique and not deductible by current methodologies.

Analyses were made in order to verify if the method is able to identify complementarities between molecular surfaces. This particular aspect can be useful in protein-protein interaction studies, since traditional methods are computationally expensive. In this context, the analysis of molecular surfaces may be useful as a screening to identify molecular complexes which can be successively evaluated by more sophisticated methods. Although the methods have behaviors similar to other programs currently available, the algorithm is not able to identify strong evidence which as can be found in the case of similarity detection. On one hand this behavior may arise from the fact that two molecular surfaces are not necessarily similar as two homologous surfaces. Second, the algorithm suffers when aligning few matches, and so a wrong correspondence has a negative effect on group formation and therefore on surface alignment. However, the aim of this program is not to provide an optimal conformation but rather to identify a limited set of probable conformations on which to perform further analysis.

In conclusion, in this work we proposed an algorithm to match macromolecular surfaces based on images of local description. The obtained results demonstrate that the proposed system is effective for matching surfaces in the context of functional annotation and can suggest useful information in interaction studies.

From a computational point of view, future developments will regard the parallelization of the algorithm, to allow efficient analysis at a higher level of detail and on wider data sets of proteins. We shall consider that not only the full Protein Data Bank can be screened to verify non trivial matches against surfaces, but in theory also a huge number of theoretical structural models can be analyzed using our algorithm. The implementation of a parallel solution is encouraged by the intrinsic parallelism of many steps of this algorithm and it will be extremely useful in order to accomplish large scale screening.

Regarding the matching algorithm, future work will be represented by a greater integration of the information related to amino acids which profile the surfaces. This idea is not in contrast with the aim of the algorithm to work, at first, without physicochemical information, but clearly the availability of such data should be considered, by providing the possibility to insert this information in the context of specific analysis to improve the quality of the results. Moreover, biological data integration can be very useful later, when matches have been identified and information about functionality can be transferred from one surface to another.

Finally this system combines multiple applications which have to be executed by the user in a series of steps. Our future plans are to extend and integrate the code in order to develop a more generic and user friendly standalone application, where a user can interact with windows and toolbars for visualizing the results, or where a user can execute queries through a web browser, thus hiding the complexity of the system.

8 BIBLIOGRAPHY

1. Friedberg, I. (2006). Automated protein function prediction--the genomic challenge. *Brief. Bioinformatics 7*, 225-242.
2. Rost, B., Liu, J., Nair, R., Wrzeszczynski, K.O. and Ofran, Y. (2003). Automatic prediction of protein function. *Cell. Mol. Life Sci. 60*, 2637-2650.
3. Shrager, J. (2003). The fiction of function. *Bioinformatics 19*, 1934-1936.
4. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet. 25*, 25-29.
5. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Res. 28*, 235-242.
6. Kinoshita, K. and Nakamura, H. (2003). Protein informatics towards function identification. *Curr. Opin. Struct. Biol. 13*, 396-400.
7. Whisstock, J.C. and Lesk, A.M. (2003). Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys. 36*, 307-340.
8. Watson, J.D., Laskowski, R.A. and Thornton, J.M. (2005). Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol. 15*, 275-284.
9. Brenner, S.E. (2001). A tour of structural genomics. *Nat. Rev. Genet. 2*, 801-809.
10. Nagano, N., Orengo, C.A. and Thornton, J.M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol. 321*, 741-765.
11. Campbell, S.J., Gold, N.D., Jackson, R.M. and Westhead, D.R. (2003). Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol. 13*, 389-395.
12. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol. 215*, 403-410.
13. Rost, B. (2002). Enzyme function less conserved than anticipated. *J. Mol. Biol. 318*, 595-608.
14. Doolittle, R.F. (1995). The multiplicity of domains in proteins. *Annu. Rev. Biochem. 64*, 287-314.
15. Finn, R.D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S.R., Sonnhammer, E.L.L. and Bateman, A. (2006). Pfam: clans, web tools and services. *Nucleic Acids Res. 34*, D247-51.
16. Karp, P.D. (1998). What we do not know about sequence analysis and sequence databases. *Bioinformatics 14*, 753-754.
17. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Pagni, M. and Sigrist, C.J.A. (2006). The PROSITE database. *Nucleic Acids Res. 34*, D227-30.
18. Rost, B. (1997). Protein structures sustain evolutionary drift. *Fold Des 2*, S19-24.
19. Holm, L. and Sander, C. (1998). Touring protein fold space with Dali/FSSP. *Nucleic Acids Res. 26*, 316-319.

20. Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., Akpor, A., Maibaum, M., Harrison, A., Dallman, T., Reeves, G., Diboun, I., Addou, S., Lise, S., Johnston, C., Sillero, A., Thornton, J. and Orengo, C. (2005). The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res. 33*, D247-51.

21. Orengo, C.A., Todd, A.E. and Thornton, J.M. (1999). From protein structure to function. *Curr. Opin. Struct. Biol. 9*, 374-382.

22. http://en.wikipedia.org/wiki/Serine_protease

23. Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C., Pogson, C.I., Wilson, I.A., Corran, P.H., Furth, A.J., Milman, J.D., Offord, R.E., Priddle, J.D. and Waley, S.G. (1975). Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 angstrom resolution using amino acid sequence data. *Nature 255*, 609-614.

24. Eswaramoorthy, S., Gerchman, S., Graziano, V., Kycia, H., Studier, F.W. and Swaminathan, S. (2003). Structure of a yeast hypothetical protein selected by a structural genomics approach. *Acta Crystallogr. D Biol. Crystallogr. 59*, 127-135.

25. Porter, C.T., Bartlett, G.J. and Thornton, J.M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res. 32*, D129-33.

26. Nadassy, K., Wodak, S.J. and Janin, J. (1999). Structural features of protein-nucleic acid recognition sites. *Biochemistry 38*, 1999-2017.

27. Ivanisenko, V.A., Pintus, S.S., Grigorovich, D.A. and Kolchanov, N.A. (2004). PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucleic Acids Res. 32*, W549-54.

28. Jambon, M., Imberty, A., Deléage, G. and Geourjon, C. (2003). A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins 52*, 137-145.

29. Moodie, S.L., Mitchell, J.B. and Thornton, J.M. (1996). Protein recognition of adenylate: an example of a fuzzy recognition template. *J. Mol. Biol. 263*, 486-500.

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

31. Tseng, Y.Y., Dundas, J. and Liang, J. (2009). Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. *J. Mol. Biol. 387*, 451-464.

32. Ren, J., Xie, L., Li, W.W. and Bourne, P.E. (2010). SMAP-WS: a parallel web service for structural proteome-wide ligand-binding site comparison. *Nucleic Acids Res. 38 Suppl*, W441-4.

33. Shulman-Peleg, A., Nussinov, R. and Wolfson, H.J. (2005). SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res. 33*, W337-41.

34. Jones, S. and Thornton, J.M. (1996). Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U.S.A. 93*, 13-20.

35. Lo Conte, L., Chothia, C. and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol. 285*, 2177-2198.

36. Ma, B., Elkayam, T., Wolfson, H. and Nussinov, R. (2003). Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl. Acad. Sci. U.S.A. 100*, 5772-5777.
37. Jones, S. and Thornton, J.M. (1997). Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol. 272*, 121-132.
38. Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J. and Huang, E.S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?. *Protein Sci. 13*, 190-202.
39. Gruber, J., Zawaira, A., Saunders, R., Barrett, C.P. and Noble, M.E.M. (2007). Computational analyses of the surface properties of protein-protein interfaces. *Acta Crystallogr. D Biol. Crystallogr. 63*, 50-57.
40. Lichtarge, O. and Sowa, M.E. (2002). Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol. 12*, 21-27.
41. Elcock, A.H. (2001). Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol. 312*, 885-896.
42. Ota, M., Kinoshita, K. and Nishikawa, K. (2003). Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J. Mol. Biol. 327*, 1053-1064.
43. Burgoyne, N.J. and Jackson, R.M. (2006). Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics 22*, 1335-1342.
44. McCoy, A.J., Chandana Epa, V. and Colman, P.M. (1997). Electrostatic complementarity at protein/protein interfaces. *J. Mol. Biol. 268*, 570-584.
45. Pazos, F., Helmer-Citterich, M., Ausiello, G. and Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol. 271*, 511-523.
46. Wodak, S.J. and Méndez, R. (2004). Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications. *Curr. Opin. Struct. Biol. 14*, 242-249.
47. Pettit, F.K. and Bowie, J.U. (1999). Protein surface roughness and small molecular binding sites. *J. Mol. Biol. 285*, 1377-1382.
48. Norel, R., Petrey, D., Wolfson, H.J. and Nussinov, R. (1999). Examination of shape complementarity in docking of unbound proteins. *Proteins 36*, 307-317.
49. Camacho, C.J., Gatchell, D.W., Kimura, S.R. and Vajda, S. (2000). Scoring docked conformations generated by rigid-body protein-protein docking. *Proteins 40*, 525-537.
50. Binkowski, T.A., Joachimiak, A. and Liang, J. (2005). Protein surface analysis for function annotation in high-throughput structural genomics pipeline. *Protein Sci. 14*, 2972-2981.
51. Lee, B. and Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol. 55*, 379-400.
52. Connolly, M.L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science 221*, 709-713.
53. http://www.netsci.org/Science/Compchem/feature15.html
54. Wang, X. (2005). Finding Patterns on Protein Surfaces: Algorithms and Applications to Protein Classification. *IEEE Transactions on Knowledge and Data Engineering 17*, 1065-1078.

55. Binkowski, T.A., Adamian, L. and Liang, J. (2003). Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J. Mol. Biol. 332*, 505-526.
56. de Rinaldis, M., Ausiello, G., Cesareni, G. and Helmer-Citterich, M. (1998). Three-dimensional profiles: a new tool to identify protein surface similarities. *J. Mol. Biol. 284*, 1211-1221.
57. Fischer, D., Wolfson, H., Lin, S.L. and Nussinov, R. (1994). Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. *Protein Sci. 3*, 769-778.
58. Tisi, L.C. and Evans, P.A. (1995). Conserved structural features on protein surfaces: small exterior hydrophobic clusters. *J. Mol. Biol. 249*, 251-258.
59. Rosen, M., Lin, S.L., Wolfson, H. and Nussinov, R. (1998). Molecular shape comparisons in searches for active sites and functional similarity. *Protein Eng. 11*, 263-277.
60. Shulman-Peleg, A., Nussinov, R. and Wolfson, H.J. (2004). Recognition of functional sites in protein structures. *J. Mol. Biol. 339*, 607-633.
61. Kinoshita, K. and Nakamura, H. (2005). Identification of the ligand binding sites on the molecular surface of proteins. *Protein Sci. 14*, 711-718.
62. Tsuchiya, Y., Kinoshita, K. and Nakamura, H. (2004). Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins 55*, 885-894.
63. Sael, L., Li, B., La, D., Fang, Y., Ramani, K., Rustamov, R. and Kihara, D. (2008). Fast protein tertiary structure retrieval based on global surface shape similarity. *Proteins 72*, 1259-1273.
64. La, D., Esquivel-Rodríguez, J., Venkatraman, V., Li, B., Sael, L., Ueng, S., Ahrendt, S. and Kihara, D. (2009). 3D-SURFER: software for high-throughput protein surface comparison and analysis. *Bioinformatics 25*, 2843-2844.
65. Bock, M.E., Garutti, C. and Guerra, C. (2007). Discovery of similar regions on protein surfaces. *J. Comput. Biol. 14*, 285-299.
66. Camacho, C.J. and Vajda, S. (2002). Protein-protein association kinetics and protein docking. *Curr. Opin. Struct. Biol. 12*, 36-40.
67. Gray, J.J. (2006). High-resolution protein-protein docking. *Curr. Opin. Struct. Biol. 16*, 183-193.
68. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C. and Vakser, I.A. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. U.S.A. 89*, 2195-2199.
69. Moont, G., Gabb, H.A. and Sternberg, M.J. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins 35*, 364-373.
70. Chen, R. and Weng, Z. (2003). A novel shape complementarity scoring function for protein-protein docking. *Proteins 51*, 397-408.
71. Gray, J.J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C.A. and Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol. 331*, 281-299.

72. Méndez, R., Leplae, R., Lensink, M.F. and Wodak, S.J. (2005). Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins 60*, 150-169.

73. Johnson, A.E. (1997). Spin-images: A representation for 3D surface matching. , .

74. Berman, H., Henrick, K., Nakamura, H. and Markley, J.L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res. 35*, D301-3.

75. http://www.wwpdb.org/documentation/format32/v3.2.html

76. Sanner, M.F., Olson, A.J. and Spehner, J.C. (1996). Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers 38*, 305-320.

77. van de Waterbeemd, H., Carter, R.E., Grassy, G., Kubinyi, H., Martin, Y.C., Tute, M.S. and Willett, P. (1997). Glossary of terms used in computational drug design (IUPAC Recommendations 1997). *Pure Appl. Chem. 69 (5)*, 1137-1152.

78. http://www.martinreddy.net/gfx/3d/OFF.spec

79. http://www.ics.forth.gr/~lourakis/levmar/

80. Baker, N.A., Sept, D., Joseph, S., Holst, M.J. and McCammon, J.A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A. 98*, 10037-10041.

81. Chen, R., Mintseris, J., Janin, J. and Weng, Z. (2003). A protein-protein docking benchmark. *Proteins 52*, 88-91.

82. Méndez, R., Leplae, R., De Maria, L. and Wodak, S.J. (2003). Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins 52*, 51-67.

83. Foster, I., Kesselman, C. and Tuecke, S. (Fall 2001). The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International Journal of High Performance Computing Applications 15*, 200-222.

84. https://edms.cern.ch/file/722398//gLite-3-UserGuide.html

85. Milanesi, L., Merelli, I., Trombetti, G., Cozzi, P. & Orro, A. (2009).Spin-images: A representation for 3D surface matching. In: Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine and Healthcare, Edition, ed. IGI Global, p. 149, 168.

86. Trombetti, G.A., Bonnal, R.J.P., Rizzi, E., De Bellis, G. and Milanesi, L. (2007). Data handling strategies for high throughput pyrosequencers. *BMC Bioinformatics 8 Suppl 1*, S22.

87. http://www.pymol.org

88. Guex, N. and Peitsch, M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis 18*, 2714-2723.

89. Humphrey, W., Dalke, A. and Schulten, K. (1996). VMD -- Visual Molecular Dynamics. *Journal of Molecular Graphics 14*, 33-38.

90. Schroeder, W.J., Avila, L.S. and Hoffman, W. (2000). Visualizing with VTK: A Tutorial. *IEEE Computer Graphics and Applications 20*, 20-27.

91. Schroeder, W.J., Martin, K. and Lorensen, B. (2004).The Visualization Toolkit - An object
Oriented Approach to 3D Graphics, Volume , Edition (Kitware).

92. Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R. (2005). InterProScan: protein domains identifier. *Nucleic Acids Res. 33*, W116-20.

93. http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html

94. Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. and Zygouri, C. (2003). PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res. 31*, 400-402.

95. Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S. and Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res. 33*, D212-5.

96. Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y. and Liang, J. (2006). CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res. 34*, W116-8.

97. Orengo, C.A. and Taylor, W.R. (1996). SSAP: sequential structure alignment program for protein structure comparison. *Meth. Enzymol. 266*, 617-635.

98. Zhang, Q., Sanner, M. and Olson, A.J. (2009). Shape complementarity of protein-protein complexes at multiple resolutions. *Proteins 75*, 453-467.

9 PUBLICATIONS

F. Viti, I. Merelli, A. Calabria, P. Cozzi, E. Mosca, R. Alfieri, L. Milanesi, “Ontology-based resources for bioinformatics analysis” in Int. Jour. Signal and Imaging Systems Engineering, *Accepted*.

A. Calabria, D. Di Pasquale, M. Gnocchi, P. Cozzi, A. Orro, G. Trombetti, L. Milanesi, “Grid Based Genome Wide Studies on Atrial Flutter”, in Journal of Grid Computing, 2010, 1-17

Torri F, Akelai A, Lupoli S, Sironi M, Amann-Zalcenstein D, Fumagalli M, Dal Fiume C, Ben-Asher E, Kanyas K, Cagliani R, Cozzi P, Trombetti G, Strik Lievers L, Salvi E, Orro A, Beckmann JS, Lancet D, Kohn Y, Milanesi L, Ebstein RB, Lerer B, Macciardi F, “Fine mapping of AHI1 as a schizophrenia susceptibility gene: from association to evolutionary evidence”, FASEB J. 2010 Aug;24(8):3066-3082.

P. Cozzi, I. Merelli, D. D’Agostino and L. Milanesi, "A parallel implementation of Genetic Algorithms for parameters estimation of Molecular Surfaces similarity analysis", in Proceedings of BITS 2010, Bari, Italy. 14-16 April, 2010.

I. Merelli, P. Cozzi, D. D’agostino, A. Clematis, L. Milanesi, “Image Based Surface Matching Algorithm Oriented to Structural Biology”, in IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume PP, Issue 99, 1

D. D’agostino, A. Clematis, I. Merelli, P. Cozzi, L. Milanesi, “A Parallel Algorithm for Molecular Surface Matching Through Image representation”, in Procedings of Parallel, Distributed and Network-Based Processing (PDP), 17-19 Feb. 2010

P. Cozzi, F.Viti, I. Merelli and L. Milanesi, "Exploitation of a 3D-oriented ontology for choosing the representative proteins in protein surfaces comparison.", in Proceedings of FOCUS K3D Conference on Semantic 3D Media and Content, INRIA Sophia Antipolis - Méditerranée, France. 11-12 February 2010.

P. Cozzi, I. Merelli, L. Milanesi, “A Visualization ToolKit based application for representing macromolecular surfaces”, in Lecture Notes in Computer Science 2009 Jun 23, Volume 5488/2009, 284-292, doi: 10.1007/978-3-642-02504-4_26*.*

P. D'Ursi, F. Chiappori, I. Merelli, P. Cozzi, E. Rovida, L. Milanesi, “Virtual screening pipeline and ligand modelling for H5N1 neuraminidase”, Biochem Biophys Res Commun. 2009 Jun 12;383(4):445-449.

D. D’Agostino, A. Clematis, I. Merelli, P. Cozzi and L. Milanesi, “Parallel Decomposition of 3D Surfaces in Images of Local Descriptors for Molecular Screening” Parallel, Distributed and Network-based Processing, pp.261-267, 2009

A. Calabria, D. Di Pasquale, G. Trombetti, P. Cozzi, A. Orro, M, Gnocchi, L. Milanesi, Genome Wide Linkage analysis and Systems Knowledge: an Integrated Web Framework for Distributed Excecution and Results Annotation., in: Proceedings of Sysbiohealth 2008, Bologna, 24-25 November 2008

P. Cozzi, I. Merelli, L. Milanesi, A Visualization ToolKit based application for representing macromolecular surfaces., in: Proceedings of Computational Intelligence Methods For Bioinformatics And Biostatistic, Vietri sul Mare (Salerno), 3-4 October 2008

I. Merelli, P. Cozzi, D. D'Agostino, A. Cleamatis and L. Milanesi, "Images Based System for Surface Matching in Macromolecular Screening", IEEE International Conference on Bioinformatics and Biomedicine, pp.397-401, 2008

L. Milanesi, I. Merelli, G. Trombetti, P. Cozzi, A. Orro, Functional Genomics Applications in Grid., in: Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine and Healtcare, Ed. Mario Catannaro. ed. IGI Global, p. 149, 168.

I. Merelli, P. Cozzi, L. Milanesi, A reference database for protein surface comparison using grid technology. in: Proceedings of NETTAB, Santa Margherita di Pula, 10-13 July 2006.

P. Cozzi, I. Merelli, L. Milanesi, Creazione di un database di supporto per analisi delle superfici proteiche tramite un approccio di calcolo ad alte prestazioni., in: Proceedings of AITIM, Torino, 11-13 December 2006.

10 AKNOLEDGMENTS

I would like to thank the ITB-CNR group for the support given during the course of this thesis. A special thanks to my colleagues for having been close during these three years and to Luciano Milanesi for having supported this project. Finally, I thank the University of Milano for the scholarship and for the training given to me during this PhD course of Molecular Medicine.