UNIVERSITY OF THESSALY
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Dissertation

3D Shape Analysis Approaches for Protein Docking and Similarity Search

by

Apostolos V. Axenopoulos
Diploma in Electrical and Computer Engineering, 2003
M.Sc. in Advanced Computing Systems, 2008

Submitted in partial satisfaction of the requirements for
the degree of Doctor of Philosophy
October 2013

The Dissertation of Apostolos Axenopoulos is approved by:

Committee Chairperson

University of Thessaly

Dedicated to my daughter

Acknowledgements

I would like to acknowledge all those who have helped me make this dissertation possible. First of all, I would like to thank my advisor, Prof. Elias Houstis, who offered me the opportunity to purse this dissertation and provided his support during its completion. I would also like to express my sincere thanks to my colleague and friend Dr. Petros Daras for helping me establish the overall direction of the research and remain focused on achieving my goal, as well as for sharing his stimulating ideas and expertise throughout the duration of my thesis. Special thanks to Dr. Georgios Papadopoulos for the fruitful collaboration. His observations and comments helped me to significantly improve my work. I am also very grateful to the remaining members of my dissertation committee for generously sharing their expertise and time in reviewing this thesis.

I feel exceptionally lucky to have worked my PhD in an excellent working environment at the Information Technologies Institute. I would like to particularly thank my colleagues and friends Dimitris Zarpalas, George Litos, Michalis Lazaridis, Thodoris Semertzidis, Lazaros Gymnopoulos and Dimitris Rafailidis, who not only assisted me in both scientific and implementation issues but also turned this obviously laborious task into an enjoyable experience.

I am also grateful to my parents for their outstanding efforts and support. There are no words that can express my gratitude and appreciation for all they have done for me. Last but not least, I would like to thank my wife Nansi, for without her love and understanding this effort could not have been accomplished.

Abstract of the Dissertation

3D Shape Analysis Approaches for Protein Docking and Similarity Search

Apostolos Axenopoulos

Doctor of Philosophy, Graduate Program in Electrical and Computer Engineering

University of Thessaly, Volos, Greece

Protein functions are carried out through their interactions with other biological molecules. Research in protein interactions has attracted special interest from the scientific community for decades and still remains a hot research topic. Among others, there is an increasing interest to develop computational methods that automatically predict the 3D structure of protein-protein complexes, such as protein-protein docking. In this thesis, we propose novel approaches to assist protein-protein docking and protein similarity search. The proposed methods are based on the fact that when two proteins interact, their surfaces at the binding site demonstrate geometric complementarity (apart from physicochemical complementarity).

The first docking method is based on geometric complementarity matching of the molecular surfaces. The basic idea of this algorithm is to use a descriptor that measures 3D shape similarity for computing shape complementarity, since equally-sized complementary surface patches tend to have similar shapes. A basic property of the proposed shape descriptor is its invariance to rotations of the surface patches. Complementarity matching is achieved through pairwise comparison of the local shape descriptors, thus, providing a fast geometric filtering and avoiding the exhaustive translational and rotational search of existing docking techniques.

The second docking method extends this work in order to produce a docking algorithm that is robust to relatively small conformational changes of the interacting proteins. Since accurate surface complementarity has been proven to be inappropriate in unbound docking, the new method allows binding of surfaces with approximate surface complementarity. Experiments proved that the improved algorithm is more robust to conformational changes. Furthermore, a new scoring function is presented, which combines geometric complementarity with physicochemical factors, such as Coulomb potentials, van der Waals forces, hydrophobicity. The scoring function takes as input the list of candidate poses and a new rank list is produced having more near-native poses in the first positions.

The last method that is proposed in this dissertation is an approach for molecular shape comparison. It aims to assist the problem of virtual screening, a process that is commonly used in rational drug design. More specifically, a new shape descriptor is proposed that combines local, global and hybrid local-global shape features. It is experimentally proven that the proposed compound descriptor is appropriate for similarity search of flexible ligands, while at the same time is robust in shape comparison of rigid proteins. Due to its compactness, the new descriptor enables fast screening of similar ligands to a target molecule from very large compound databases.

Περίληψη της Διατριβής

Μέθοδοι Ανάλυσης Τρισδιάστατων Σχημάτων για Αναζήτηση Ομοιότητας Πρωτεϊνών και Protein Docking

Απόστολος Αξενόπουλος

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών υπολογιστών

Πανεπιστήμιο Θεσσαλίας, Βόλος

Οι λειτουργίες των πρωτεϊνών πραγματοποιούνται κυρίως μέσω των αλληλεπιδράσεών τους με άλλα μόρια. Η έρευνα στο πεδίο των αλλλεπιδράσεων των πρωτεϊνών έχει προσελκύσει το ενδιαφέρον των επιστημόνων εδώ και δεκαετίες και εξακολουθεί να παραμένει ένα από τα σημαντικότερα προβλήματα στο χώρο της βιολογίας. Μεταξύ άλλων, το ενδιαφέρον στρέφεται στην ανάπτυξη υπολογιστικών μεθόδων ικανών να προβλέψουν με αυτόματο τρόπο την τρισδιάστατη δομή των συμπλόκων μορίων πρωτεϊνών, όπως το “Protein Docking”. Στην παρούσα διατριβή, προτείνονται καινοτόμες μέθοδοι για την υποβοήθηση των διαδικασιών του protein docking και της αναζήτησης πρωτεϊνικών δομών. Οι προτεινόμενες μέθοδοι βασίζονται στην ιδέα ότι όταν δυο πρωτεΐνες αλληλεπιδρούν οι επιφάνειές τους στην περιοχή σύνδεσης πρέπει να παρουσιάζουν γεωμετρική συμπληρωματικότητα (εκτός από τη φυσικοχημική συμπληρωματικότητα).

Η πρώτη μέθοδος για protein docking βασίζεται στο ταίριασμα γεωμετρικά συμπληρωματικών επιφανειών. Η βασική ιδέα του αλγορίθμου είναι η χρήση ενός περιγραφέα (descriptor) που υπολογίζει ομοιότητα τρισδιάστατων σχημάτων (3D shape similarity) για τον υπολογισμό γεωμετρικής συμπληρωματικότητας, δεδομένου ότι ισομεγέθη συμπληρωματικά τμήματα επιφανειών τείνουν να έχουν παρόμοιο σχήμα. Μια σημαντική ιδιότητα του προτεινόμενου περιγραφέα είναι η αμεταβλητότητά του κατά την περιστροφή των επιφανειακών τμημάτων. Το ταίριασμα των συμπληρωματικών τμημάτων πραγματοποιείται με σύγκριση κατά ζεύγη των αντίστοιχων περιγραφέων, παρέχοντας έτσι ένα γρήγορο γεωμετρικό

φιλτράρισμα. Με τον τρόπο αυτό, αποφεύγονται οι εξαντλητικές μετατοπίσεις και περιστροφές και επιταχύνεται η διαδικασία του docking.

Η δεύτερη μέθοδος αποτελεί επέκταση της προηγούμενης με στόχο τη δημιουργία αλγορίθμου για docking που παρέχει μεγαλύτερη σταθερότητα όταν οι πρωτεΐνες παρουσιάζουν σχετική ευκαμψία (flexibility). Η περίπτωση του “unbound docking”, όπως αλλιώς λέγεται, απαιτεί οι δυο επιφάνειες να παρουσιάζουν “κατά προσέγγιση” συμπληρωματικότητα, παρά “ακριβή” συμπληρωματικότητα όπως συμβαίνει στην περίπτωση του “bound docking”. Πειραματικά αποτελέσματα απέδειξαν ότι ο βελτιωμένος αλγόριθμος είναι πιο εύρωστος στα θέματα ευκαμψίας των πρωτεϊνών. Επιπλέον, παρουσιάζεται μια νέα συνάρτηση βαθμολόγησης (scoring function), η οποία συνδυάζει γεωμετρική συμπληρωματικότητα με φυσικοχημικούς παράγοντες, όπως δυναμικά Coulomb, δυνάμεις van der Waals, υδροφοβικότητα. Η συνάρτηση δέχεται ως είσοδο τη λίστα με τις υποψήφιες θέσεις (πιθανά σύμπλοκα) και παράγει μια νέα κατάταξη όπου τα σωστά σύμπλοκα εμφανίζονται συχνότερα στις πρώτες θέσεις.

Η τελευταία μέθοδος είναι μια προσέγγιση για σύγκριση ομοιότητας τρισδιάστατων πρωτεϊνικών δομών. Μια από τις πιθανές εφαρμογές της είναι στο πρόβλημα του “virtual screening”, της αναζήτησης όμοιων προσδετών (ligands) μέσα από βάσεις μορίων, με στόχο το σχεδιασμό φαρμάκων. Πιο συγκεκριμένα, προτείνεται ένας νέος περιγραφέας σχήματος, ο οποίος συνδυάζει τοπικά (local), καθολικά (global) και υβριδικά (local-global) χαρακτηριστικά σχήματος. Πειραματικά αποτελέσματα αποδεικνύουν ότι ο προτεινόμενος περιγραφέας είναι κατάλληλος για αναζήτηση ομοιότητας ακόμα και όταν τα μόρια-προσδέτες παρουσιάζουν σχετική ευκαμψία. Παράλληλα, είναι εξίσου αποτελεσματικός στην αναζήτηση ομοιότητας άκαμπτων (rigid) πρωτεϊνών. Λόγο του μικρού του μεγέθους (compactness), ο νέος περιγραφέας είναι κατάλληλος για γρήγορη αναζήτηση όμοιων μορίων ακόμα και από πολύ μεγάλες βάσεις.

Related Publications

Journal publications:

1. **A. Axenopoulos**, P. Daras, G. Papadopoulos, E. Houstis, "A Shape Descriptor for Fast Complementarity Matching in Molecular Docking", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8 no. 6, pp. 1441-1457 Nov.-Dec. 2011.
2. **A. Axenopoulos**, P. Daras, G. Papadopoulos, E. Houstis, "SP-Dock: Protein-Protein Docking using Shape and Physicochemical Complementarity", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 10 no. 1, pp. 135-150 Jan.-Feb. 2013.
3. **A. Axenopoulos**, D. Rafailidis, G. Papadopoulos, E. Houstis, and P. Daras, "Similarity Search of Flexible 3D Molecules combining Local and Global Shape Descriptors", IEEE/ACM Transactions on Computational Biology and Bioinformatics, under review.

Conference publications:

1. **A. Axenopoulos**, P. Daras, G. Papadopoulos, E. Houstis, "Protein-Protein Docking using the Shape Impact Descriptor", 5th Conference of Hellenic Crystallographic Association (HECRA2010), Larissa, Greece, September 2010.
2. **A. Axenopoulos**, P. Daras, G. Papadopoulos, E. Houstis, "3D Protein-Protein Docking using Shape Complementarity and Fast Alignment", IEEE Int. Conference on Image Processing, ICIP 2011, Sep 11-14, Brussels.

Contents

Chapter 1

Introduction

1.1 Protein Docking

Proteins are the workhorse molecules of life, since numerous biological processes are mediated by them. Among others, proteins are the motors that cause muscle contraction, they drive life-sustaining chemical processes and they hold cells together to form tissues and organs. Enzymes are proteins that catalyse chemical reactions, while regulatory proteins control the location and timing of gene expression. Information between cells is transmitted by signaling proteins such as cytokines and hormones. Structural proteins provide support for cells and organs and form large structures such as hair, nails and skin. In the human genome, the protein-coding genes lead to the production of human proteins. If the mechanism of activity of all these proteins was fully understood, we would be able to understand the causes of several diseases, such as cancer, amyotrophic lateral sclerosis, Parkinson's, heart disease etc. All of the above indicate that protein science is still an active research field [1].

The structure of a protein is the arrangement of a polypeptide chain (sequence of aminoacids) in the three-dimensional (3D) space. The 3D structure of the protein and the way its polypeptide chain folds in the 3D space are crucial to understand the protein's biological function [2]. There are several techniques for determination of the protein structure. The most commonly used is X-ray crystallography, which is applicable to molecules/complexes of any size. A drawback of this method is that it provides only information about the native structure of the protein under the particular experimental conditions, thus, it does not include information about the flexibility of the molecule. On the other hand, Nuclear Magnetic Resonance (NMR) spectroscopy can detect more than one structures for the flexible parts of the protein, but it is limited to proteins smaller than 25-30 kilodaltons (kDa). Finally, Electron Diffraction is another option, however, it is rarely used for protein structure determination.

Protein functions are carried out through their interactions with other biological molecules, such as proteins, nucleic acids, lipids, sugars, nucleotides, ions and water. A failure to create the appropriate complex, during a protein interaction, may be the cause of several serious diseases, such as Alzheimer's disease, Huntington's disease, cystic fibrosis, etc. Thus, it is not surprising that research in protein interactions has attracted special interest from the scientific community for decades and still remains a hot research topic in Biochemistry, Biophysics and Bioinformatics. Despite the fact that a vast number of protein interactions have been discovered in the last years, only in a little portion of them the crystallized structures of the resulting complexes are currently available. Thus, it is not surprising that the number of solved complexes in Protein Data Bank (PDB[1]) is orders of magnitude smaller than those of structures of individual proteins [4]. The main reason for this is that X-ray crystallography and NMR spectroscopy encounter difficulties in dealing with structures of complexes. The former because the flexibility of the complex formation makes the crystallization difficult and the latter because the size and weight of the resulting complex usually becomes prohibitive for NMR [5]. Thus, there is an increasing interest to investigate computational methods that automatically predict the structures of protein-protein complexes [6]. These approaches constitute the field of *Protein Docking*, which has attracted increasing interest during the past two decades and still remains a hot research topic in Bioinformatics.

Protein docking deals with the prediction of the conformation and orientation of one protein (ligand) within the binding site of another (receptor). In other words, it calculates the three-dimensional (3D) structure of a protein complex starting from the individual structures of the constituting proteins [7]. The optimized conformation and orientation should be such that the free energy of the overall system is minimized. Most docking algorithms include the following components: a search technique to find the optimal placement (pose) of one protein (ligand)

[1] http://www.rcsb.org/

with respect to the other (receptor); and a scoring function to rate each pose and provide a rank list of candidate complexes. Detecting the optimal pose, during the search stage, is based on the "lock-and-key" concept, that is, the surfaces of the receptor and ligand at their binding site should have geometric complementarity. The "lock-and-key" principle was initially introduced in 1894 by Fischer [3] to explain the function of enzymes and it remains valid until today. Although it is a simple concept, its implementation is not a trivial task. Searching for complementary surfaces usually involves an exhaustive search of the rotational and translational space of one protein with respect to the other, resulting in a six-dimensional search, which is highly time-consuming.

Despite the extensive research in protein-protein docking, a complete solution has yet to be achieved due to the large complexity of the problem. It has been proposed that shape complementarity alone cannot achieve highly accurate docking predictions [7], but it should be used in combination with physicochemical factors, such as Coulomb potentials, van der Waals forces, hydrophobicity, etc. Additionally, protein interactions can involve significant conformational changes, thus, docking techniques should take into account the side-chain and the backbone flexibility. A fast computational method that will accurately predict protein-protein interactions would become a valuable tool for biologists and biochemists. Successful docking will predict binding site amino acids crucial for the complex stability, which will assist biochemists perform concrete mutations in order to test their impact for the protein function. Moreover, a search for docking partners for a protein in structural data bases and prediction of its binding mode could suggest a possible function for this protein.

1.2 Rational Drug Design

As explained above, several biological processes involve binding of a protein to a target molecule. This binding may be the part of a signalling mechanism between cells or a mechanical operation (e.g. muscle contraction), it can mediate a catalytic event, or it could be part of another process. An approach that is usually followed in drug discovery is *competitive inhibition*, which is based on finding an inhibitor able to bind to a protein instead of its natural binding partners, in order to interrupt whatever process the protein mediates.

The process of drug discovery in the laboratory is highly time-consuming and expensive. The candidate drugs must be synthesized and assayed on the target protein for activity, as well as with non-targets for cross-reactivity. In order to speed-up the drug design process, computational techniques would be beneficial. In general, computational approaches comprise two main categories [1]:

De novo design: these methods are based on building a molecule from scratch to fit the binding site of a protein. This usually involves identifying molecular fragments complementary to specific parts of the binding site and connecting them into a single molecule.

Docking: the process starts with a database of known molecules and attempts to place each one in the binding pocket of the protein. When a candidate pose is achieved, the affinity of the binding is estimated using a scoring function. Eventually, a list of the best-binding molecules for the target protein is returned.

In industrial drug design, the economic impact of protein-protein docking is very high, since an accurate and fast docking algorithm will enable rapid scanning of structural data bases for matches with specific targets, which will speed-up the design process of new drugs and increase productivity. Thus, it is not surprising why protein-protein docking is still a very hot research topic and a lot of effort is put towards investigation of a more accurate computational docking solution, which is expected to provide additional insight into the nature of macromolecular recognition.

1.3 Scope and Outline of the Dissertation

As previously explained, one of the key factors for protein docking is the geometric complementarity of the interacting proteins' surfaces at their binding site. However, methods based on exhaustive search of the rotational and translational space of the molecule are highly complex and become prohibitive when search for potential binding partners of a protein is performed on large databases of ligands, as is the case in drug design. A common technique in rational drug design is the process of *virtual screening*, where a search is performed in libraries of small molecules in order to identify those structures which are most likely to bind to a drug target. Several variations of virtual screening are available. The most common are *structure-based virtual screening* and *ligand-based virtual screening*. Structure-based virtual screening involves docking of candidate ligands into a protein target followed by applying a scoring function to estimate the likelihood that the ligand will bind to the protein with high affinity [120]. Ligand-based virtual screening is based on searching molecules with shape similar to that of known actives, as such molecules will fit the target's binding site and hence will be likely to bind the target [121]. The latter technique is based on the *similarity property principle* [8], according to which similar molecular structures are likely to have similar properties.

The scope of the research that is presented in this dissertation is to investigate novel approaches to assist protein-protein docking, particularly in the part of geometric complementarity search. The main objective is to develop algorithms for *Molecular Shape Comparison* to facilitate geometric docking, without the need for exhaustive translations and rotations of the interacting molecules. Special attention is given to the conformational changes that take part during the interaction, thus, the proposed methods will partially support molecular flexibility. Molecular shape comparison will be based on 3D pattern recognition. More specifically, the molecular surface of a protein is treated as a 3D object. After proper 3D content processing, a set of low-level features (descriptors) is extracted from the object, which uniquely characterize the shape of the protein. These features are in the form of low-dimentional vectors and provide a highly compact representation of the specific protein. The descriptors are applied to local regions of the molecular surface and allow for partial or global shape matching. Proteins or parts of proteins with similar low-level features will also have similar (or complementary) shape. This reduces significantly the computation time and provides a fast geometric filtering to speed-up the overall docking process.

The rest of the dissertation is organized as follows.

In Chapter 2, a state-of-the-art analysis in the field of protein-protein docking is given, with emphasis on methods based on geometric complementarity of surfaces. The advantages and disadvantages of each approach are discussed and the motivation for the algorithms proposed in this dissertation is explained. Furthermore, the problem of molecular shape comparison is introduced and the most representative approaches in the field of 3D pattern recognition, which are appropriate for protein similarity matching, are presented.

In Chapter 3, an approach to protein-protein docking is introduced, which is based on geometric complementarity matching of the molecular surfaces. The molecular surface (of either the receptor or the ligand) is segmented into equally-sized geometric patches and a shape descriptor is extracted for each patch. The basic idea of this algorithm is to use a descriptor that measures 3D shape similarity for computing shape complementarity, since equally-sized complementary surface patches tend to have similar shapes. Complementarity matching is achieved through pairwise comparison of the local shape descriptors, thus, providing a fast geometric filtering and avoiding the exhaustive translational and rotational search. To produce the candidate poses, alignment of the ligand is implemented by superimposing either a ligand patch onto a complementary receptor patch or a pair of neighboring ligand patches onto a pair of complementary receptor patches.

Chapter 4 extends the work presented in Chapter 3 in several aspects. Protein docking, as initially proposed in Chapter 3, is based on accurate surface complementarity, which produces high quality poses in the rigid-body case. However, when it comes to non-rigid docking, which is the real case, accurate surface complementarity has been proven to be inappropriate. Thus, a new matching framework is proposed in Chapter 4, which allows binding of surfaces with approximate surface complementarity. Eperiments proved that the improved algorithm is more robust to conformational changes. Furthermore, a new scoring function is presented, which combines geometric complementarity with physicochemical factors, such as Coulomb potentials, van der Waals forces, hydrophobicity. The scoring function takes as input the list of candidate poses and a new rank list is produced having more near-native poses in the first positions.

In Chapter 5, a new approach for molecular shape comparison is presented. It aims to assist the problem of virtual screening, a process that is commonly used in rational drug design. More specifically, a compound shape descriptor is introduced to represent the global shape of a molecule. The advantage of the specific descriptor is that it combines features for both rigid and non-rigid shape matching, thus, it is appropriate for similarity search of flexible ligands, while at the same time is robust in shape comparison of rigid proteins. Due to its compactness, the new descriptor enables fast screening of similar ligands to a target molecule from very large compound databases. A comparison with existing shape-based methods for virtual screening demonstrates the superiority of the proposed method.

In Chapter 6, overall conclusions are drawn and some ideas for future work are proposed.

Chapter 2

Related Work

2.1 Protein Docking Approaches

Protein docking has been evolved into a distinct computational discipline, bringing together techniques from a broad spectrum of sciences such as physics, chemistry, biology, mathematics and computing, with the objective to model in silico how proteins behave [7]. Protein-protein docking methods generally consist of three components: a) *Molecular Surface Representation*; b) *Conformational Space Search*; and c) *Scoring of Potential Solutions*. These components and their role to docking are analysed in the following subsections.

2.1.1 Molecular Surface Representation

The basic description of the protein (or ligand) surface is the atomic representation of exposed residues. A protein structure file (e.g. PDB file) usually provides no more information than a list of atom locations in space and their types. This representation allows for visualization, but it cannot help to distinguish which parts of which atoms are on the surface of the protein and which are buried inside the structure. Knowledge of the surface atoms is crucial for docking since only those take part in the interaction. Thus, an additional tool is needed to capture notions of interior and exterior and spatial adjacency.

One of the most common ways to represent the molecular surface is by its geometric features, such as the *Solvent Excluded Surface* (SES) [9], also known as "*Connolly Surface*". SES is calculated by rolling a probe sphere (of size equal to the size of the solvent molecule) over the exposed contact surface of each atom. Everywhere the center of the sphere goes is the *Solvent Accessible Surface* (SAS), while everywhere the sphere touches (including empty space) constitutes the SES (Figure 2.1). Depending on how the probe sphere touches the Van der Waals atoms, there are three different types of surface regions (Figure 2.2): a) convex (green), where the probe sphere touches one atom; b) concave (yellow), where it touches three atoms and c) saddle (purple), where it touches two atoms. A common tool for extraction of SES is the Maximal Speed Molecular Surface (MSMS) [10] algorithm.

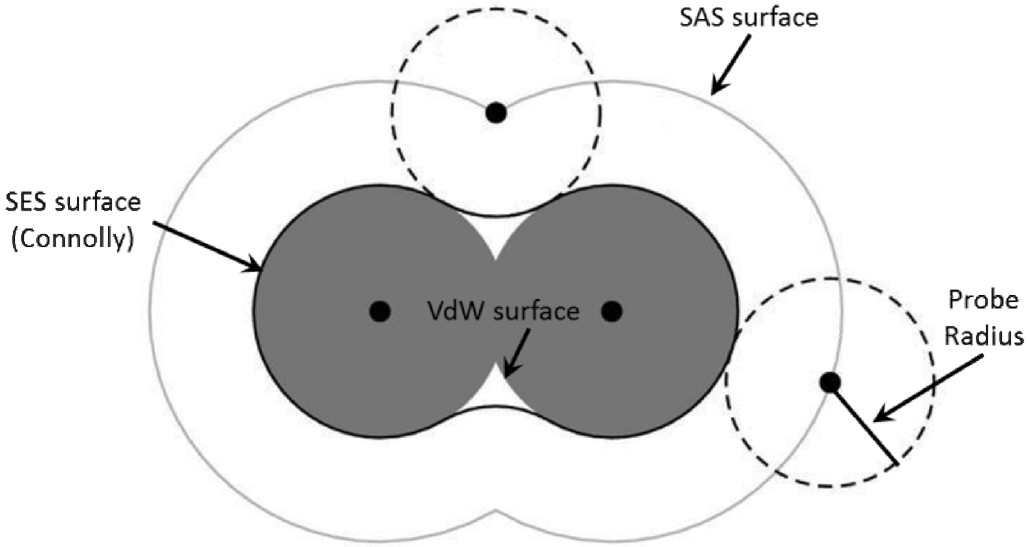


Figure 2.1: Calculation of SES and SAS surface using a probe sphere

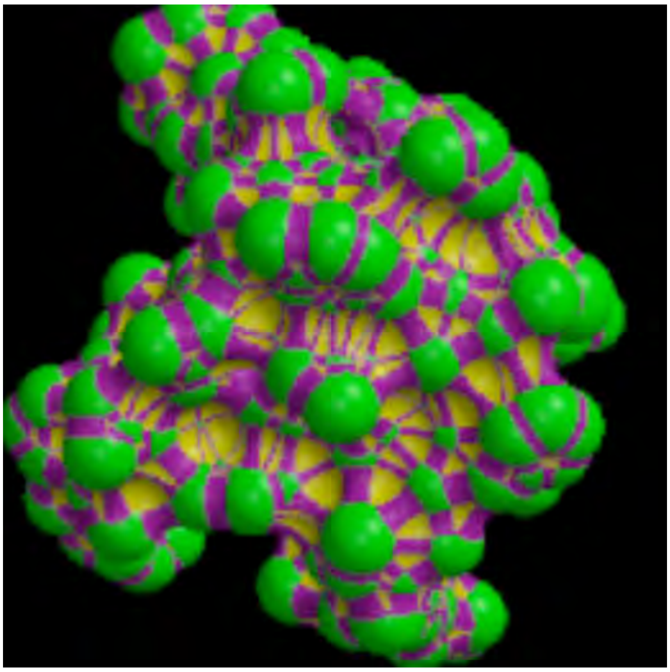


Figure 2.2: The SES surface of a protein, where convex, concave and saddle regions are coloured in green, yellow and purple, respectively.

Apart from the Connolly surface, the *Alpha Shapes* technique [11] is also used for molecular surface representation. In 2D space, two points are “alpha-exposed” if there exists a circle of radius α such that the two points lie on the surface of the circle and the circle contains no other points from point set. Thus, finding all the alpha-exposed regions of the point set defines an enclosed region (Figure 2.3). Expanding this principle in 3D space, we can use alpha shapes to define the surface of the molecules. Alpha Shapes provide a coarser representation of the Connolly surface, however, since there are fewer points to consider, the approximation of the surface as well as the surface matching run faster.Additionally, by tuning α, we can define the degree of matching (i.e. coarse or fine).

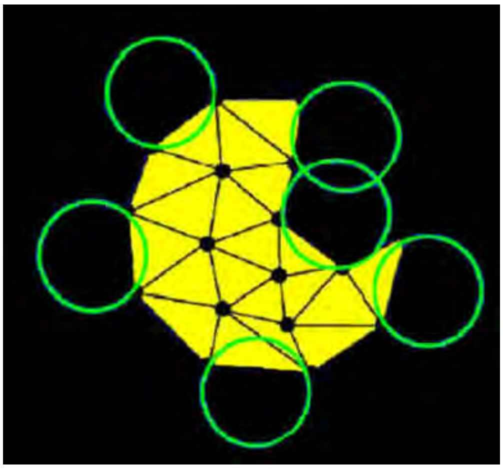


Figure 2.3: Defining an enclosed region using Alpha Shapes.

Other representation approaches are the *Lenhoff* technique [12] and the *Kuntz et al. Clustered Spheres* [13]. The former computes the possible locations of the ligand that will be bound. Thus, it computes a “complementary” surface for the receptor. The latter generates a sphere for every pair of points *i* and *j* that lie on the surface. The generated sphere is centered on the normal at point *i*. Regions where these generated spheres overlap define possible areas of cavities on the receptor and protrusions on the ligand.

2.1.2 Conformational Space Search

During this step, an appropriate technique is used to place the ligand in various candidate poses in the binding site of the receptor. The goal is to achieve at least one near-native pose among the resulting candidate poses. Although each placement could be completely random and independent, most algorithms either use heuristics based on the chemistry or geometry of the atoms involved or use a standard optimization technique to avoid exhaustive translational and rotational search. A wide spectrum of docking algorithms include Fast Fourier Transform (FFT) correlations [14], geometric hashing [15], and Monte Carlo (MC) [16] techniques has been utilized in current docking algorithms.

Regarding geometric docking, two broad categories of algorithms can be identified: a) *brute-force scanning* of the transformation space and b) *local shape feature matching*. Brute force algorithms [17], [18], [19] search the entire 6-dimensional transformation space of the ligand. They begin with a simplified rigid body representation of protein shape obtained by projecting each protein onto a regular 3D Cartesian grid, and by distinguishing grid cells according to whether they are near or intersect the protein surface, or are deeply buried within the core of the protein. Then, docking search is performed by scoring the degree of overlap between pairs of grids in different relative orientations. The running times of those algorithms may reach days of CPU time. In order to make the procedure faster, several techniques have been utilized, such as the FFT [20]. 3D FFT has been incorporated in several correlation-based docking algorithms [21], [22], [23]. A recent overview of the principles of grid-based FFT docking approaches is given in [24]. In [25], a grid-free Spherical Polar Fourier (SPF) approach is introduced which allows rotational correlations to be calculated rapidly using one-dimensional (1D) FFTs.

Towards the direction to improve the computation time in brute-force algorithms, ZDOCK [26] introduces a shape complementarity scoring function called Pairwise Shape Complementarity (PSC). The method computes the total number of receptor-ligand atom pairs within a distance cutoff. In contrast with traditional FFT based methods, PSC does not explicitly explore the entire rotational space resulting in low computation times. Finally, there are also non-deterministic methods in the category of brute-force docking approaches that use genetic algorithms [27], [28]. One of the most recent approaches of this category is presented in [35]. The so-called F^{2}Dock is an extension of a NFFT-based docking algorithm, where an adaptive search phase (rotational and translational) has been incorporated to achieve faster running times.

Since they are based on exhaustive scanning of translational and rotational space, brute-force methods are able to detect at least one near-native pose in almost every complex. On the other hand, this may lead to an extraordinary big number of candidate docking poses, where, due to the existence of false positives, the near-native poses may not be ranked at the first positions. Such phenomena could be avoided with the use of local shape feature matching methods, which detect points of interest on the protein surfaces. These methods require a representation of the molecular surface, attempting to find critical patches on the surface. Then, pairwise complementarity matching is applied on these patches. One of the first docking approaches, based on local shape feature matching, was introduced in 1982 [29]. In [30], a method to match local curvature maxima and minima points was presented. This technique has been extended in [31], [32]. In [33], a method based on geometric hashing [15] is presented. Each protein surface is first pre-processed to give a list of critical points (“pits”, “caps”, and “belts”) which are then compared, using geometric hashing, to generate a relatively small number of trial docking orientations for grid scoring. The method requires low computation times, comparing with other docking algorithms, however, it is not so efficient in predicting the correct pose, since the pits, caps and belts do not enclose significant shape information.

A more recent approach extracts local features from the solvent excluded surface of a protein and is called context shapes [34]. These are boolean data structures and correspond to significantly large parts of the protein surface. Complementarity shape matching is achieved using efficient boolean operations (Figure 2.4). The method demonstrates superior performance over other similar approaches in predicting the correct docking pose using only geometric criteria. However, the exhaustive search of relative orientations for each local feature, even with the use of a pre-calculated lookup table, increases the computational cost as well as the memory requirements. In [36], the method LZerD is introduced, which is based on 3D Zernike Descrptors (3DZD). These are a series expansion of a 3D function (i.e. protein surface) allowing for a compact representation of the 3D function. 3DZD are extracted on local patches that are derived on uniformly distributed points of the protein surface. Partial matches are computed using geometric hashing. Surface Histograms (shDock) [37] is a local shape descriptor, which captures the local geometry around a set of two points with given normals on the surface of a protein. The docking pose is obtained automatically by matching two surface histograms. shDock has achieved the best performance among existing methods in Docking Benchmark 2.4, in the bound docking case, i.e. where the candidate proteins are taken directly from the crystallized complex. However, when dealing with the unbound case, the performance of shDock decreases significantly.

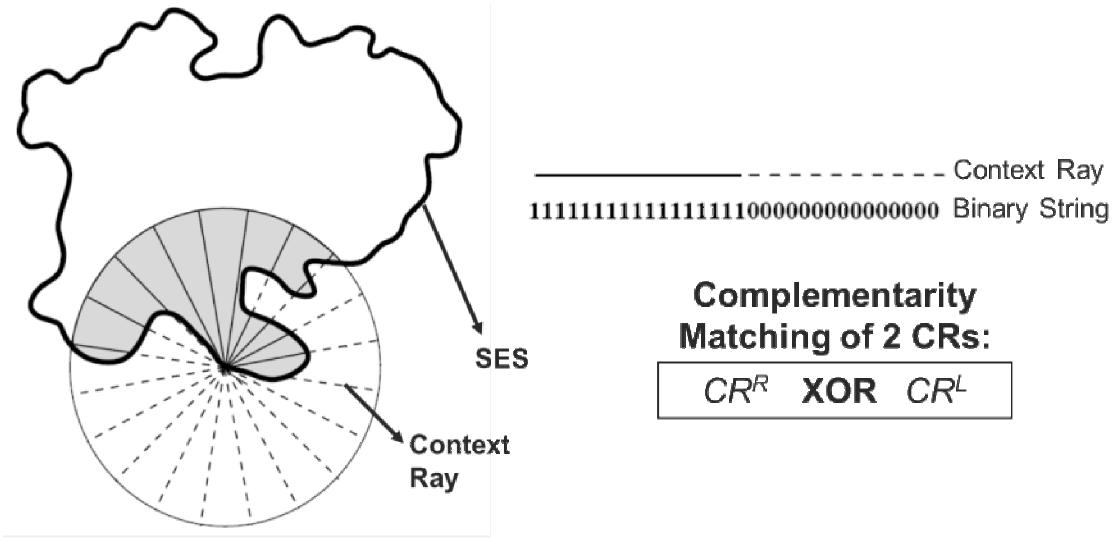

Context Ray
11111111111111111000000000000000 Binary String
Complementarity Matching of 2 CRs:
CR^R XOR CR^L
SES
Context Ray

Figure 2.4: The Context Shapes approach for protein docking.

2.1.3 Scoring of Potential Solutions

The scoring function provides a way to rank the candidate poses of the ligand. Ideally, the score should correspond directly to the binding affinity of the ligand for the protein, so that the best scoring pose is a near-native pose. In order to evaluate the feasibility of each pose several scoring functions have been introduced, based either on geometric complementarity or other non-geometric factors such as desolvation, hydrophobicity, and electrostatics [38], [39].

In [33], the candidate poses produced by geometric hashing are ranked using a geometry-based scoring function. This function relies on the creation of a 3D distance transform grid, where the receptor is placed. Each voxel (cubic bin) of the grid is assigned a value equal to the distance from the receptor’s molecular surface. Then, the translated and rotated ligand enters the grid. The total score is the sum of scores of all ligand surface points accessing the grid voxels (Figure 2.5). Ligand points that lie within the Buried Surface Area (BSA) designate a region where the two surfaces are complementary, thus, increase the total score of the pose. Ligand points that penetrate the surface of the receptor, i.e. create steric clashes, are not allowed, thus, decrease the total score of the pose. A similar geometric scoring function is also described in Chapter 3.

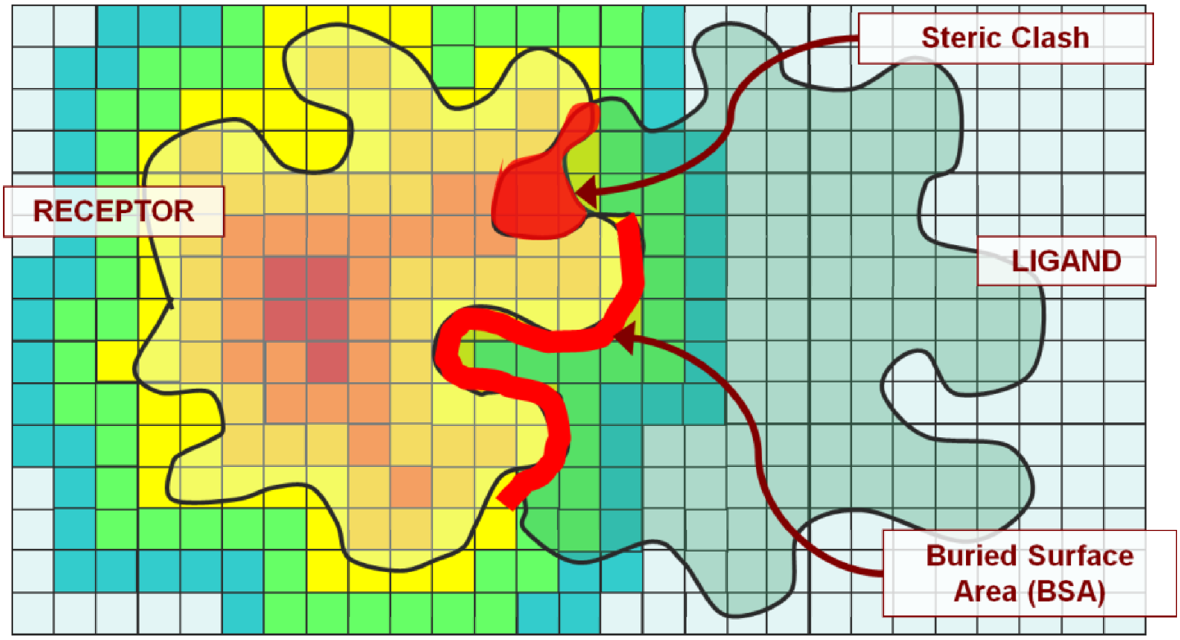


Figure 2.5: Geometric scoring using distance transform grid.

It has been proven that shape complementarity alone does not provide the best possible results, thus, non-geometric factors such as desolvation, hydrophobicity, and electrostatics have been also investigated in order to improve the scoring functions [38]. Recent attempts focus on combining geometric and physicochemical properties in order to produce more accurate predictions. In [39], shape complementarity matching along with knowledge-based potentials, electrostatics, atom desolvation energy, residue contact preferences and Van-derWaals potential are combined, demonstrating remarkable results on a test set of 68 bound and 30 unbound test cases. Although the contribution of each individual non-geometric factor was not assessed in [39], an important conclusion can be drawn: shape complementarity should be combined with physicochemical complementarity to increase the accuracy of docking predictions. F2Dock [35] computes separately shape complementarity scores and electrostatics scores and combines them. This leads to an improvement of shape-only docking in 54% of the complexes of Docking Benchmark 2.0.

The most straightforward way to incorporate geometric and non-geometric properties is to represent the final scoring function as a weighted sum of those factors and determine the optimal weights that each factor contributes to the overall scoring. In [40], an empirical scoring function is proposed, which is a linear combination of the energy score, interface propensity and residue conservation score. The scoring function is the weighted sum of these three values, where the weights are optimized using a simple grid method (exhaustive search of the optimal combination). Similarly in [38], the scoring function is a linear weighted sum of van der Waals attractive and repulsive energies, electrostatics short and long range attractive and repulsive energies, and desolvation. To optimize the weights, a downhill simplex minimization algorithm is used.

Towards the direction of improving existing docking approaches and investigating new approaches, the CAPRI experiment [41] (Critical Assessment of Predicted Interactions) has become an ideal arena for testing docking algorithms. More specifically, in CAPRI, new protein-protein complexes are subjected to structure prediction before they are published. The complexes are submitted by several predictor groups and they are assessed by comparing their geometry to the original structure. Some of the most well-known docking algorithms, such as PatchDock and ZDock, have participated in CAPRI experiment producing acceptable solutions for several CAPRI targets [42], [43].

2.2 Molecular Shape Comparison

As it has been explained in the previous chapter, the 3D structure of a protein is very important in order to understand its function and biological action. Comparison of the 3D molecular structures is useful in a variety of applications such as protein function prediction, computer aided molecular design, rational drug design and protein docking. Following the *similarity property principle* [8], according to which similar structures are likely to have similar properties, several approaches for molecular structure comparison have been proposed, using different representations of the molecules. As an example, in rational drug design, the process of *virtual screening* is usually applied, where given a target molecule, a search is performed in a large database for compounds that are most similar to the target. Since these compound databases range from thousands to millions of structures, an ideal method should provide accurate and at the same time rapid similarity matching. Among the various existing structural comparison methods [44], [45] those that are based on comparison of structures by their mainchain orientation [46] or the spatial arrangement of secondary structure [47] are quite slow, thus, similarity search in large molecular databases can be time-consuming. Therefore, in order to accelerate the search time, methods of 3D shape matching have been proposed in the literature.

Techniques for similarity matching of molecular structures can be classified into different categories based on the molecular representation [48]. The most commonly used representations include backbone Ca positions [44], distance maps [49], secondary structure elements [47] and backbone torsion angles [50]. The technique/algorithm that is used for comparison highly depends on the chosen representation. As an example, for backbone representations, a common technique is dynamic programming [44]; spatial arrangements are used with secondary structure elements [47], while Monte Carlo algorithms are used with distance maps [49]. In general, most of the aforementioned methods are based on comparing coordinates of corresponding residues, which requires a structure superimposition (e.g. by using dynamic programming) as a preprocessing step. This can be time consuming when search is performed in large-scale molecular databases with thousands of structures. As the need for rapid and accurate comparison is becoming even more critical, due to the increasing size of the databases, Molecular Shape Comparison (MSC) techniques have been introduced [51]. MSC methods extract low level features (descriptors) that capture the spatial profile of the protein as a multidimensional feature vector. In this case, similarity matching is reduced to descriptor comparison using a common distance measure, which obviates the need for any feature correspondence or prealignment. Since the work presented in this paper belongs to the category of MSC techniques, a more detailed state-of-the-art analysis of these methods is provided in the sequel.

In shape-based approaches, the protein (or molecule in general) is treated as a three-dimensional (3D) object, on which an appropriate algorithm is applied to extract low-level descriptors that uniquely characterize its shape. A common representation that is extensively used is the molecular surface [52]. Considering the molecular surface as input, several features can be generated, such as Spin Images [53] or Shape Histograms [54]. Spin Images are local 2D descriptions of the surface based on a reference frame that is defined by the associated surface points. Shape Histograms, on the other hand, exploit global geometric properties of the protein captured in the form of a probability distribution sampled from a shape function (e.g. angles, distances, areas). In [55], the protein surface is given as input and 2D views of the surface are taken from 100 uniformly sampled viewpoints. Comparison is performed by multi-view matching using 2D Zernike moments and Fourier descriptors for each 2D view. Multi-view representation has been proven quite efficient for shape matching of 3D objects; however, the optimal performance is achieved when the database objects have symmetries, i.e. in retrieval of generic objects [56]. In the case of molecular shapes, these symmetries are not present. Apart from the molecular surface, other representations are also possible. The method presented in [57], [58], describes the shape of a molecule through its set of interatomic distances, which is encoded as a geometrical descriptor vector. The method achieves very fast comparison times and is appropriate for virtual screening problems.

An interesting category of shape-based approaches comprises methods that extract moments from the 3D object. These have been successfully applied in pattern recognition problems [59]. The moment-based representations result in compact descriptor vectors with high discriminative power. Examples of moments are based on the theory of orthogonal polynomials, such as 2D/3D Zernike moments and Legendre moments [60]. These descriptors allow also reconstruction of the object from its moments [61]. The method in [62] takes as input the volume of the 3D molecular structure producing a new domain of concentric spheres. In this domain, 2D Polar-Fourier coefficients and 2D Krawtchouk moments are applied, resulting in a completely rotation-invariant descriptor vector. Spherical Harmonics have been widely used in molecular similarity comparison problems such as virtual screening [63], protein structure representation and comparison [64] and molecular docking [65][66]. Spherical Harmonics have the advantage of allowing the surface information to be encoded in a compact form as an orthonormal 1D vector of real numbers allowing fast comparison. Their main disadvantages are: a) they represent only star-shape surfaces; and b) the handling of alignment problems is associated with the fast comparison of objects [67]. Recently, 3D Zernike descriptors (3DZD) have been introduced as a representation of the protein surface shape [68]. These are based on a series expansion of a given 3D function. 3DZDs are rotation invariant, with the protein structures not necessarily being aligned to perform the molecular shape comparison. Another advantage of 3DZDs is that they allow other characteristics of a protein surface, such as electrostatic potentials, to be incorporated into the descriptor vector [68]. 3DZDs have been used in problems of protein structure retrieval, protein-protein docking [36] and virtual screening [69] with quite satisfactory results.

In all methods for molecular shape comparison described above, the 3D molecules are treated as rigid objects. A drawback of these approaches is that they are not robust to shape deformations of flexible molecules. Since many molecules are flexible and this flexibility is part of their function, it should by no means be underestimated. To address such problems, methods for non-rigid shape matching should be utilized. Some initial attempts have been proposed, which exploit non-rigid 3D shape descriptors, such as the Local-diameter Descriptor [70] and the Inner Distance Shape Signature [71][72], however, they are not able to efficiently handle shape deformations of molecules with topological changes. In [73], authors propose a method for flexible molecular shape comparison using diffusion distances. The Diffusion Distance Shape Descriptor (DDSD) is a histogram of the diffusion distances between all sample point pairs on the molecular surface. Experiments in a database of flexible molecules show that DDSD outperforms similar approaches. Another approach, which was introduced in [74] for fast screening of proteins, is based on extraction of local patches from the protein surface and computation of a geometric fingerprint (distribution of curvatures) for each patch. This method exploits local surface similarities and achieves rapid shape comparisons.

Methods for non-rigid shape matching have been introduced to address problems that include articulation of the 3D objects (e.g. different human or animal poses in generic 3D object retrieval, molecular flexibility in bioinformatics), since rigid shape descriptors have been proven inappropriate [75]. The two main categories of non-rigid approaches are: a) global-shape-based and b) local-shape-based methods. Global approaches usually transform the Euclidiean space or Euclidean metrics to a metric space where the pairwise distances between points of the 3D object surface are invariant to deformations of the 3D object. Examples include canonical forms [76], geodesic distances [77] or diffusion distances [73]. Local-shape-based methods sample the surface and extract descriptors for each of the sampled local regions. Then, a codebook is created and a bag-of-features method is applied to generate a global shape descriptor [78][79]. Although these approaches are appropriate for non-rigid shape matching problems, in rigid shape retrieval they have inferior performance comparing to rigid methods [80]. To increase the robustness of shape matching so as to deal with both rigid and non-rigid problems, a combination of multiple features that capture different properties of the 3D shape should be investigated. It has been recently proven that combining multiple shape descriptors can significantly improve the performance of rigid 3D shape retrieval [81]. Following the same concept, a framework that combines multiple shape descriptors to address both rigid and flexible molecular shape matching problems is presented in Chapter 5 of this dissertation.

Chapter 3

Molecular Docking using Geometric Complementarity Matching

3.1 Introduction

This chapter presents a novel approach for fast rigid docking of proteins based on geometric complementarity. After extraction of the 3D molecular surface, a set of local surface patches is generated based on the local surface curvature. The shape complementarity between a pair of patches is calculated using an efficient shape descriptor, the Shape Impact Descriptor. The key property of the Shape Impact Descriptor is its rotation invariance, which obviates the need for taking an exhaustive set of rotations for each pair of patches. Thus, complementarity matching between two patches is reduced to a simple histogram matching. Finally, a condensed set of almost complementary pairs of surface patches is supplied as input to the final scoring step, where each pose is evaluated using a 3D distance grid. The experimental results prove that the proposed method demonstrates superior performance over other well-known geometry-based, rigid-docking approaches.

The chapter is organized as follows: in Section 3.2, an overview of the proposed method is provided along with its major scientific contributions. In Section 3.3, a new approach for extraction of critical points from the molecular surface is introduced. In Section 3.4, a description of the Shape Impact Descriptor is given, while in Section 3.5, the final step of the algorithm, which includes alignment and geometric scoring is presented. Then, in Section 3.6, the experimental results are presented, where the proposed method is compared with other state-of-the-art approaches. Finally, the chapter is summarized in Section 3.7.

3.2 Overview and Contributions

The proposed method can be summarized as in the block diagram presented in Figure 3.1. The input is the PDB [82] file of the protein, which is used to generate the Solvent Excluded Surface (SES). Then, a set of critical points is extracted from the surface. The critical points correspond to the centers of small elementary patches (either convex or concave). Then, for each critical point, an Extended Surface Patch (ESP) is created, which spreads over a wider surface area around that point. Each ESP that corresponds to a convex (or concave) elementary patch of the receptor protein is matched with all ESPs that correspond to concave (or convex) elementary patches of the ligand protein. For complementarity shape matching a new rotation-invariant shape descriptor, called the Shape Impact Descriptor (SID), is used. Since SID is invariant to rotation, there is no need to rotate the ESP of the ligand with respect to the receptor patch. The pairs of ESPs ranked as most complementary are given as input to the final step of the algorithm, where the candidate poses are scored, using a distance transform grid.

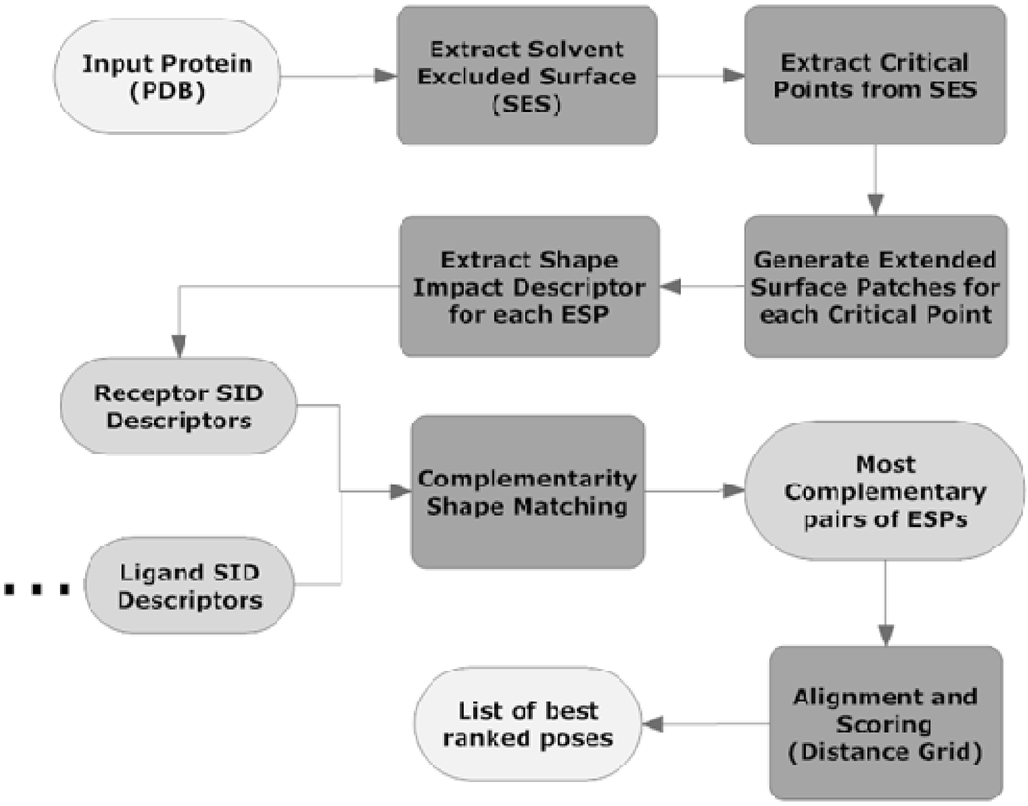

Input Protein (PDB)
Extract Solvent Excluded Surface (SES)
Extract Critical Points from SES
Generate Extended Surface Patches for each Critical Point
Extract Shape Impact Descriptor for each ESP
Receptor SID Descriptors
Ligand SID Descriptors
Complementarity Shape Matching
Most Complementary pairs of ESPs
Alignment and Scoring (Distance Grid)
List of best ranked poses

Figure 3.1: Block diagram of the proposed method

The major strength of the proposed approach is that it introduces a shape similarity descriptor to measure surface complementarity. This is based on the notion that two ESPs with complementary shape can be also regarded as of similar shape if a) they have a specific size and b) the second ESP is turned upside down so that the inner part of the ligand surface matches the outer part of the receptor surface. The size of the ESP should be relevantly large to enclose significant shape information, while at the same time it should be kept within a maximum radius, since with further growth in ESP's size the criterion (b) may not be fulfilled. While there are only few techniques for efficient complementarity surface matching, regarding similarity shape matching a wider variety of algorithms is available. Thus, following the notion described above, it is easier to develop a method for partial surface complementarity by appropriately modifying a shape matching technique. The idea of matching the negative surface of a protein to deal with complementarity matching has been used in the past for similar problems. The DOCK program [83], which is widely used in protein docking, is based on generating a negative image of the receptor's docking site. Then, the shape of a ligand is matched with this negative image in terms of similarity. This approach, which is analysed in [84], differs from the proposed method in the following: the method presented in [84] requires an approximation of the imaginary atoms that lie at the other side of the receptor's negative surface, since mathching is performed by atom-by-atom comparison with the atoms of the ligand. On the other hand, our method is applied directly on the surfaces of the interacting molecules in a more efficient way.

Another innovative feature is that the proposed Shape Impact Descriptor is invariant to any rotation of the matching ESPs, which obviates the need for an exhaustive search of relative orientations, during the pairwise complementarity matching of ESPs. This reduces significantly the computation time and provides an efficient fast filtering for the final scoring stage.

The reduction of computation time is of crucial importance for a docking algorithm, however, the prediction accuracy should by no means be underestimated. The proposed method achieves significant improvement in prediction accuracy by introducing two conceptually simple features in the geometric scoring stage. The first involves a set of additional translations, after superimposition of the two ESPs. The reason is that an actual contact point may not always coincide with a critical point. In fact the actual contact point may lie in a small area close to the critical point. By slightly moving the ligand ESP within a small area close to the critical point, it is more likely to find a pose, which is close to the original pose. The second feature is a slight modification of the scoring function. More specifically, instead of using the ligand surface points to access the distance grid, the triangle centers of the ligand surface are used. The contribution of each triangle to the total score is multiplied by the area of the triangle. This results in a more accurate scoring, taking into account that the point distribution is not uniform across the 3D mesh of the molecular surface.

The idea behind the proposed approach was inspired by the method presented in [34]. The concept of pairwise complementarity matching of equally sized surface patches is common to both approaches; however, the proposed method introduces several innovative features. First of all, in [34], the authors adopt the method in [32] in order to generate an initial set of sparse critical points, while, in this work, a new method is developed (Section 2), which provides a more approximate representation with sparse points and it can be applied also to non-molecular 3D meshes. Furthermore, the two methods use different local descriptors to measure the shape complementarity of surface patches. In [34], the Context Shapes are used, which require an exhaustive set of rotations of the ligand patch with respect to the receptor. In the proposed approach, a new descriptor is introduced, the Shape Impact Descriptor, which is rotation-invariant, thus, it does not require several rotations of the ligand. This provides a fast geometric filtering, keeping only a very small subset of candidate poses for the final scoring step. Finally, the proposed method provides an additional scoring step, which is an improvement of the distance grid used in [33], in order to produce more accurate results.

3.3 Molecular Surface Representation and Critical Points Extraction

A local shape feature matching algorithm for protein docking requires, as a first step, an appropriate representation of the molecular surface. In this work, the Solvent Excluded Surface

(SES) [9] has been used, which efficiently represents the shape of a protein. SES is calculated by rolling a probe sphere (of size equal to the size of the solvent molecule) over the exposed contact surface of each atom. In order to generate the SES, the Maximal Speed Molecular Surface (MSMS) [10] algorithm has been utilized.

Given the SES of a protein as input, a set of critical points can be extracted. These are usually the centers of concave (holes), convex (knobs) or saddle areas of the molecular surface. Several approaches have been utilized to derive critical points from SES. One of the most widely used is the sparse surface representation [85]. The sparse surface consists of three types of points called caps, pits and belts. These points correspond to the face centers of convex, concave and flat areas of the surface, respectively. The face centers are calculated by projecting the centroid of each face to the surface in the normal direction.

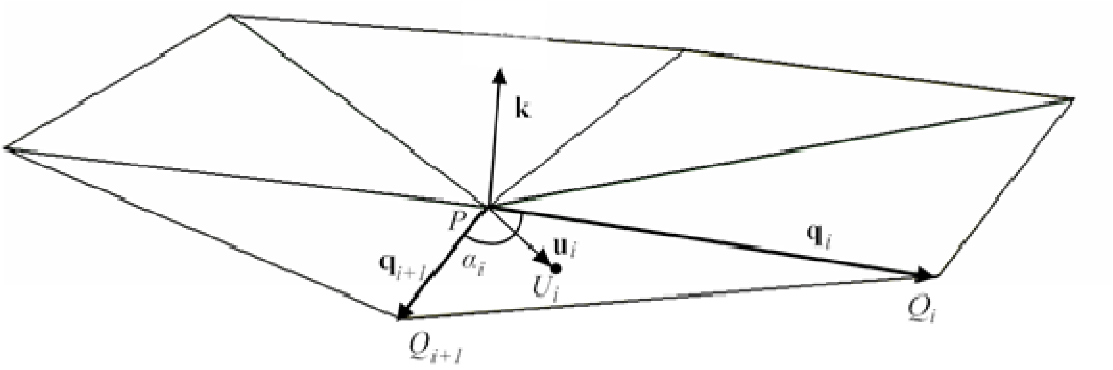


Figure 3.2: Estimation of the local curvature around a point *P*

As an alternative, we propose a method for generating critical points based on the local curvature of the surface. The reason for not adopting the sparse surface [85] to extract critical points is that the proposed method is applied directly to the 3D mesh, while the sparse surface requires additional information about the surface atoms. Thus, the sparse surface can be used to estimate the local curvature only for molecular surfaces extracted using the Connolly algorithm, while the proposed approach is applicable to all types of triangulated meshes.

More specifically, for each point P of the molecular surface, the vector k, which provides a local estimation of the curvature, is calculated as follows (Figure 3.2):

$$\mathbf{k} = \sum_{i=1}^{N} \frac{\mathbf{u}_i}{|\mathbf{u}_i|} \alpha_i \tag{3.1}$$

where *N* is the total number of neighboring points Q_i of *P*, u_i is the vector from *P* to U_i and U_i is the centroid of the triangle PQ_iQ_i+1. The angle α_i is given by:

$$\alpha_i = \arccos\left(\frac{\mathbf{q}_i \cdot \mathbf{q}_{i+1}}{|\mathbf{q}_i||\mathbf{q}_{i+1}|} \right) \tag{3.2}$$

and $\mathbf{q}_i$ is the vector from *P* to Q_i.

For surface points *P* that belong to convex areas, their corresponding vectors **k** point at the inner part of the molecule, while the vectors of points that belong to concave areas point at the outer part of the molecule (Figure 3.2). In flat areas, the vectors are almost tangential to the surface (they point neither at inner nor outer part of the molecule). This can provide an initial segmentation of the SES into three distinct regions according to the curvature (convex, concave and flat regions), which is reduced to selecting continuous regions where the vectors point at the same direction (inner, outer or tangential to molecular surface). In Figure 3.3, a Connolly surface, segmented into different regions according to the curvature, is depicted. Convex areas are marked with red, concave areas with blue and flat areas with green color, respectively.

These areas need to be further segmented into smaller patches. The centers of these patches will eventually provide the set of critical points. The algorithm for the segmentation of these areas (Figure 3.4) consists of the following steps:

Step 1: select a continuous region of surface points of the same type (convex, concave or flat).

Step 2: rank all region points according to their distance from the region contour and select those with the maximum distance as seed points. In Figure 3.4 (a), the two selected seed points are marked with the blue dots.

Step 3: expand each seed point uniformly to all directions along the surface until the region contour is reached. In the example shown in Figure 3.4 (b), the contour is reached at the second level of expansion for both seed points. The set of surface points, which are grouped around a seed point, constitute an elementary patch (convex, concave or flat) centered at the seed point

(Figure 3.4 (c)). If a seed point is already included in a group centered at another seed point, it is removed from the seed points list.

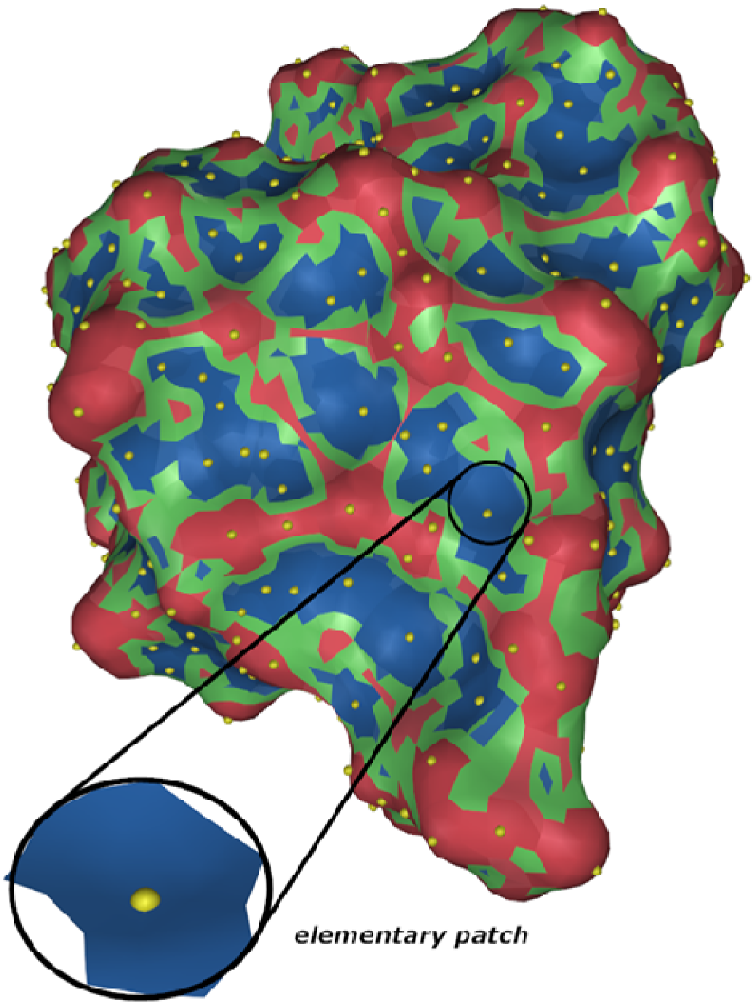


Figure 3.3: Segmentation of SES into convex, concave and flat regions. The critical points are represented by yellow dots.

In Figure 3.3, the yellow points represent the centers of elementary patches after the segmentation step. The procedure described above results in a sparse set of critical surface points. These can be characterized as convex, concave or flat, according to the type of their corresponding elementary patches. Critical points provide a sufficient approximation of the molecular surface, which significantly reduces the search space in local shape feature matching algorithms. In our approach, the convex points of the receptor are matched with the concave points of the ligand and vice versa (excluding flat points) in order to find candidate poses. The matching relies on the shape complementarity between the extended patches which correspond to each critical point. The Shape Impact Descriptor used for complementarity matching is described in the following subsection.

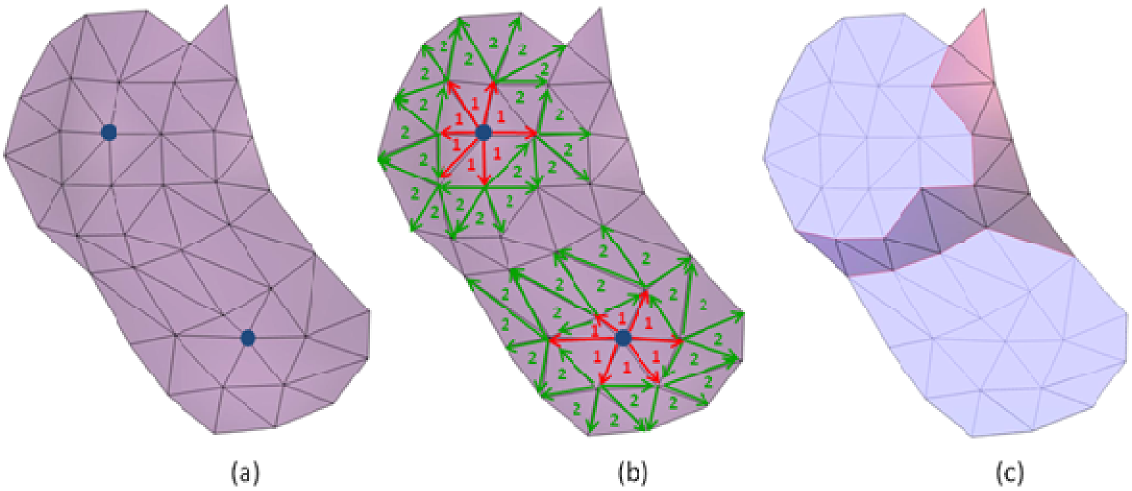


Figure 3.4: The steps for segmenting a continuous region of surface points of the same type: (a) select the most distant points from the region contour as seed points (b) expand uniformly to all directions until the region contour is reached; the numbers represent the level of expansion around the seed point (c) group all surface points covered by the expansion around each seed point; these sets of points constitute the elementary patches.

3.4 The Shape Impact descriptor

The idea of local shape complementarity matching that we propose is similar to the one presented in [34]. More specifically, we are interested in finding one or more Possible Contact Points (PCPs) from the receptor and their corresponding points from the ligand. These PCPs can be derived from the sparse critical surface points of each molecule, since sparse critical surface provides a good approximation of the molecular surface. If two PCPs, one from the receptor and one from the ligand, are actual contact points, the ligand is translated so that its PCP coincides with the receptor's PCP. Then the ligand is appropriately rotated around that point in order to find the optimal pose.

It can be easily inferred from the above that for a pair of actual contact points, the ESPs, which are centered at these points, should be parts of the actual binding site and reveal shape complementarity. Thus, in order to identify candidate poses, a complementarity matching of all potential pairs of ESPs takes place. In the proposed approach, the ESPs of the receptor centered at convex critical points are matched with the ESPs of the ligand centered at concave critical points and vice versa. This is due to the assumption that a convex critical point is highly probable to match with a concave critical point, while other combinations (convex-convex, concave-concave, convex-flat, concave flat) are less likely to happen. Finally, the case of flat-flat critical points is not taken into account, even if it is very likely to happen. The reason is that the discriminative power of a complementarity matching algorithm cannot be fully exploited in this case, since two flat-only ESPs can be both complementary and similar at the same time. Therefore, at least one convex-concave or concave-convex combination should appear in every pair of matching ESPs.

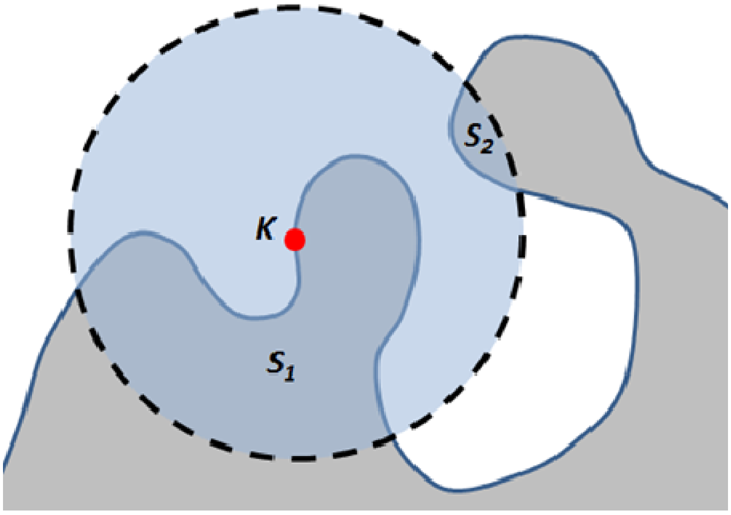


Figure 3.5: Removal of unconnected surface parts using geodesic distance: taking into account only Euclidean distances from the center *K* of the ESP, both S_1 and S_2 surface parts are enclosed. However, points of S_2 have geodesic distances greater than the predefined threshold G_{max}, thus, they are discarded.

3.4.1 Preprocessing

Given the SES of a protein along with the set of critical points, an ESP is extracted as follows:

Firstly, a sphere of a given radius E centered at a critical point is created. The ESP consists of the part of the SES (points/triangles) enclosed within the sphere. In order to discard small unconnected surface parts enclosed within the sphere, an additional filtering based on the geodesic distance G from the center is applied. Geodesic distance between two surface points is the shortest path on the surface connecting these points. In Figure 3.5, the creation of an ESP is depicted. Based only on Euclidean distance between the ESP center K and all surface points, both S_1 and S_2 surface parts are included. However, points that belong to the unconnected surface part S_2 are very far from the ESP center in terms of geodesic distance, thus, they should be discarded. Surface points with geodesic distance greater than a predefined threshold ($G_{\max}$) are excluded from the ESP. The value of $G_{\max}$ has been experimentally determined and the value that was used for the experiments is given in Table 3.4.

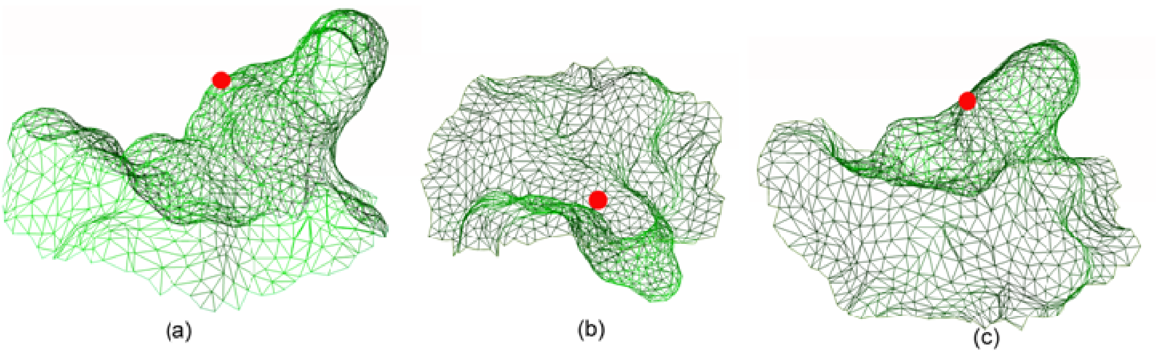


Figure 3.6: a) an ESP of the receptor of the 1CGI complex (large protrusion); b) the ESP of the ligand (deep cavity) centered at a critical point which is a point of actual contact with the ESP in a); c) the ESP of b) turned upside down so that the inner surface is visible. The patches in a) and c) have approximately similar shapes.

In Figure 3.6, a pair of complementary ESPs of the 1CGI complex is depicted. Their centers (red spheres) are actual contact points in the SESs of the two interacting proteins. Note that in both Figure 3.6a and Figure 3.6b, the outer parts of the surface patches are shown. In Figure 3.6c, the inner part of the ligand ESP is depicted. It is obvious that the latter patch has similar shape with the receptor ESP (Figure 3.6a), if its inner part is treated as outer and vice versa. Based on this observation, the complementarity matching of ESPs can be reduced to a similarity matching problem, using a shape similarity descriptor, the Shape Impact Descriptor.

The Shape Impact Descriptor was first introduced in [86] as a shape similarity measure for 3D objects. In the present work, the 3D objects are the ESPs of the receptor and the ligand. In order to proceed to descriptor extraction, the triangulated mesh representation of the ESPs has to be transformed into a binary 3D function. More specifically, the triangulated mesh, after translation, is placed inside a cubic grid (Figure 3.7). The binary 3D function $f(i,j,k)$ for each voxel $[i,j,k]$ of the cubic grid is given as:

$$f(i,j,k)=\begin{cases}1, \textit{when at least one surface point lies inside the voxel}\\0, \textit{otherwise}\end{cases}$$

Note that in the above equation voxels that lie inside the molecule are not taken into account, since only surface points lead to non-zero values of $f(i,j,k)$. Note also that scaling normalization of the 3D mesh is not required in this case since all ESPs have the same size.

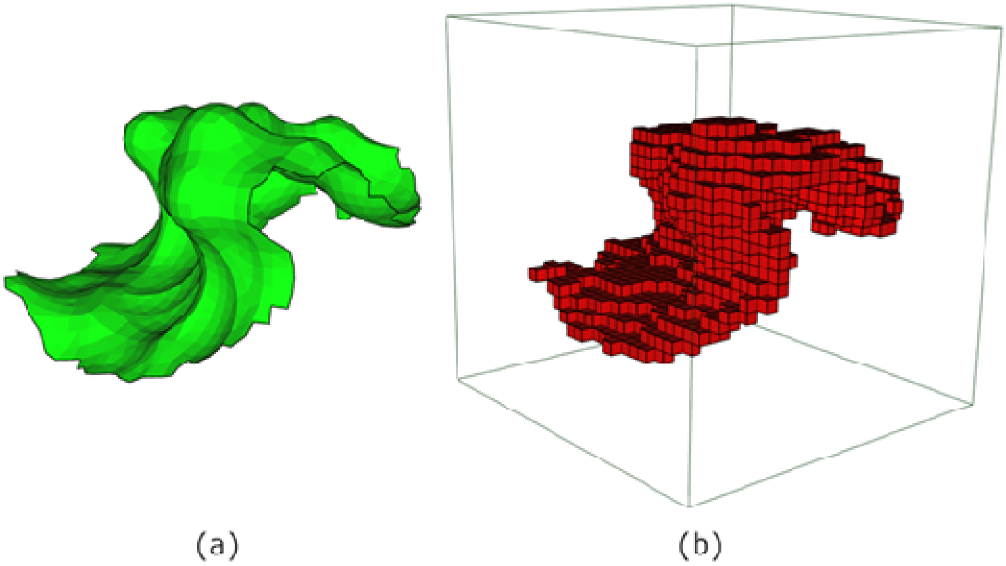


Figure 3.7: a) An ESP of the receptor of the 1AY7 complex (triangulated mesh); b) the same ESP represented as a binary 3D function *f*. Here only the voxels (red boxes) where *f* has non-zero values are depicted.

3.4.2 Descriptor Extraction

The key idea of the Shape Impact Descriptor (SID) is the description of the resulting phenomena that occur by the insertion of the 3D object in the space. It is expected that similar objects will result in similar physical phenomena. Regarding the specific problem of complementarity matching between two ESPs, SID can provide an efficient geometric descriptor. Some obvious selections are the traditional electrostatic force field (following the Coulomb law) and the Newtonian force field. More sophisticated selections could involve the generalized Einstein field theory, or the Maxwell electromagnetic field theory [87].

In order to compute a field, a cause for the field existence should be selected. Thus, every voxel of the 3D object is considered as point mass, (or, equivalently as a point charge). Any 3D object can be considered as a distributed mass (or a distributed charge) with a specific distribution, resulting in a static field around it. More specifically, in every point $\mathbf{x}=[x\,y\,z]^T$ of the 3D space that is not occupied by the object, the density and the potential of the field can be computed according to:

$$E(x)=C\sum_{i=1}^{N}\frac{1}{|x-x_i|^{r+1}}(x-x_i) \tag{3.3}$$

$$\phi(x)=C\sum_{i=1}^{N}\frac{1}{|x-x_i|^{r-1}} \tag{3.4}$$

where $r=1,2,\ldots$ is a free parameter that defines the field's law. It is obvious that for $r=2$, the generalized field is identical to the classical Newtonian/Coulombian field. The constant parameter has been selected to be $C=1$, without any loss of generality. Equations (3.3) and (3.4) are applied to all points $\mathbf{x}=[x\,y\,z]^T$ of the 3D space not occupied by the object, i.e. those points lying at the centers of the voxels $[i,j,k]$ of the cubic grid where $f(i,j,k)=0$. The parameter N in (3.3) and (3.4) represents the number of all non-zero voxels, i.e. where $f(i,j,k)=1$. With the voxel-based representation, a uniform distribution of field points around the 3D object is easily obtained.

The introduction of the parameter r in the field's equations offers a great flexibility: different values of r result in different ways that every point of the object contributes to the resulting field. Generally, the static field at a point is mainly the result of the mass that is included in an area centered at this point and its size depends on the value of r, due to the quantity $|x-x_i|^{r-1}$ in the denominator of (3.3) and (3.4). For lower values of r, the area that affects the value of the field in a specific point is larger, while for greater values of r, the area is smaller. In general, when the value of r is low, the resulting field captures more global information while greater values of r result in a more local object description.

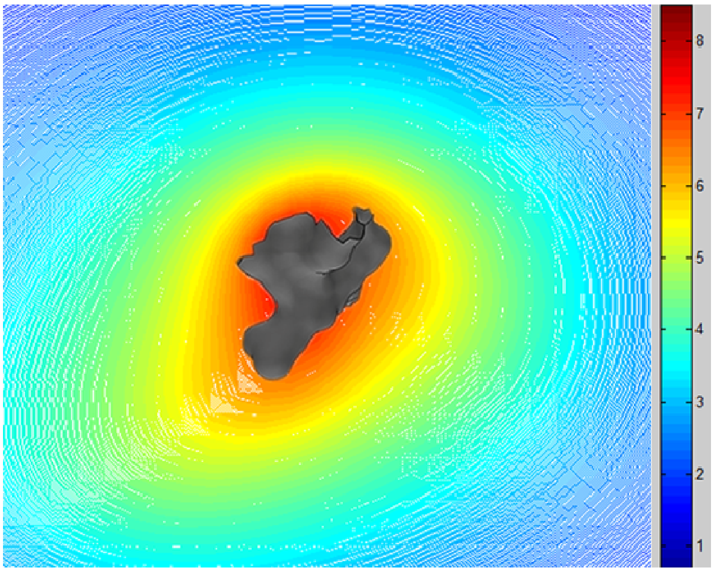


Figure 3.8: The field's potential f(x) produced from the surface of an ESP

The field is computed in various points in the exterior of the object. The key point in the presented approach is the selection of the appropriate observation areas in the exterior of the 3D object to create histograms. By examining (3.3) and (3.4), it is observed that the field vanishes and tends to be homogeneous as the point under suspicion in the exterior of the 3D objet is moved away from the object. This effect is clearly depicted in the equipotential areas around the object (Figure 3.8). Thus, the field at points that are closer to the surface of the object presents more variations and, thus, the resulting descriptor corresponding to these points is intuitively more discriminative.

In the proposed approach, SID is composed of three major histograms created by:

The field potential values, computed in points that are equidistant from the object surface. A point $\mathbf{x}$ belongs to a set of equidistant points of distance d from the object, if its distance to the closest non-zero voxel is equal to d. For the computation of the sets of equidistance points, the voxel-based distribution is used, where points $\mathbf{x}$ lie at the centers of zero valued voxels.

$$\left\{\phi(x) : x \in R^3, \min(x - x_i) = d\right\} \tag{3.5}$$

The field density Euclidean norms, computed in points that are equidistant from the object surface.

$$\left\{|E(x)| : x \in R^3, \min(x - x_i) = d\right\} \tag{3.6}$$

The radial component of the field density, computed in points that are equidistant from the object surface.

$$\left\{E(x)\cdot n_r(x): x\in R^3, \min(x-x_i)=d\right\} \tag{3.7}$$

where $n_r(x)=\frac{x-x_c}{|x-x_c|}$ and x_c is the mass center of the 3D object.

The computation of the histograms involves only relative distances, thus the resulting histograms are invariant under rotation of the 3D object. In fact, very slight variances in the values of SID descriptors between an ESP at the initial pose and the same ESP under rotation are observed. In general, the creation of a 3D voxel grid results in information loss due to discretisation errors. Therefore, the resulting voxel grids are not completely invariant under rotation of the original ESPs (surface points). However, if an adequate level of resolution is chosen for the 3D grid (64^3 voxels), these variances are insignificant (0.001% dissimilar) comparing with the dissimilarity values between two SID descriptors of different ESPs.

In our implementation, the ESPs are described as binary 3D functions in a $M\times M\times M$ grid. The size M of the grid was determined experimentally. More specifically, several resolutions of the binary 3D function were tested ($M=32,64,128,256$). For $M<64$, the resolution was not high enough to efficiently describe cavities and protrusions of the ESP, while for $M>64$, the descriptor extraction time became dramatically high. Finally, $M=64$ was selected as the optimal grid size.

Each ESP’s descriptor is composed of eight histograms of potential values, eight histograms of field’s density and eight histograms of field’s radial component. More specifically, each of the above three measures (potential values, field’s density and field’s radial component) is calculated for $r=1,2,5,6$ field’s laws, examined at points that are $d=1$ and $d=2$ far from the object surface. Therefore a total of $3\times 4\times 2=24$ histograms are calculated. Every histogram consists of 75 bins. The values of r have been appropriately chosen so as to capture both global ($r=1,2$) and local ($r=5,6$) features. Based on the notion that similar 3D objects will result in similar physical phenomena, these sets of histograms are expected to efficiently capture the geometry of the ESP patch. For a more elaborate analysis of how these values were selected, the reader could refer to [87], which describes the extraction of the SID descriptor in detail.

3.4.3 Matching

Due to the different nature of the histograms described above, several comparison metrics have been utilized. More specifically, for the potential related histograms, the normalized distance, presented in [88], has been utilized:

$$dis(H_1,H_2)=\sum_{i=0}^{K}\frac{2\left|H_1(i)-H_2(i)\right|}{H_1(i)+H_2(i)} \tag{3.8}$$

where K is the number of histogram bins. For the other two types of histograms (field's density and field's radial component), the diffusion distance [89] was used. In diffusion distance, the difference between two histograms H_1 and H_2 is treated as an isolated temperature field and a metric for its diffusion is computed.

The object descriptors are compared in pairs. Each SID descriptor consists of 24 histograms (8 histograms of potential values, 8 of field's density and 8 histograms of field's radial component). Every histogram is compared to the appropriate histogram of the other object and "sub-dissimilarities" are computed using the aforementioned dissimilarity metrics. The final dissimilarity metric between two objects is the sum of the sub-dissimilarities.

Let now R and L be the receptor and ligand protein and N_R, N_L the number of critical points of their SESs respectively. We also define the extended surface patches $ESP_R(i)$ and $ESP_L(i)$, as well as the Shape Impact Descriptors $SID_R(i)$ and $SID_L(i)$ for each critical point, where $i=1,\ldots,N_R$ and $j=1,\ldots,N_L$. All pairwise dissimilarities dis_{ij} between convex (or concave) critical points i of the receptor and concave (or convex) critical points j of the ligand, are computed:

$$dis_{ij}=dis\left(SID_R(i),SID_L(j)\right) \tag{3.9}$$

where the dissimilarity between two SID descriptors is computed using the comparison metrics described above. Pairs of ESPs with low values of dis_{ij} have similar shape and should constitute pairs of complementary surface patches. In order to keep only pairs of complementary patches, the array of pair dissimilarities dis_{ij} is sorted in ascending order and the *k*-first pairs are selected for the final scoring step.

In the final scoring step, for each of the selected complementary ESP pairs, a set of candidate poses is calculated and a score for each pose is computed. The process of final geometric scoring is much more time consuming than the dissimilarity matching between SID descriptors. Therefore, only a significantly small subset of patch pairs should be selected as *k*-first, in order to avoid high computation times. On the other hand, the number of *k*-first pairs should not be very small, so that at least one pair of actual contact points is among these pairs.

In order to determine an optimal value for *k*-first, an experiment has been performed using a set of 10 arbitrarily chosen complexes from the Docking Benchmark v2.4 [90]. The results are shown in Table 3.1. In the second column, the total number of ESP pairs (convex-concave and concave-convex) between receptor patches and ligand patches is depicted. In the third column, the rank of the first ranked pair of actual contact points is shown. In order for a pair of patches to be a pair of actual contact points, the following inequality must be fulfilled:

$$dis_{EUCL}\left(C_R, C_L\right) < \varepsilon \tag{3.10}$$

where dis_{EUCL} is the Euclidean distance between the centres C_R and C_L of the receptor and ligand ESPs, respectively. The coordinates of C_R and C_L are the absolute coordinates in the original complex and ε should be a very small value (less than 1.5Å but not zero in order to compensate for small translations around the contact points). From this table it can be inferred that just 0.1% of the total ranked pairs suffice to derive at least one pair of actual contact points. Moreover, the number of *k*-first selected pairs is not a constant value but it depends on the sizes of the two interacting molecules.

Table 3.1: The rank of the first ranked pair of actual contact points along with the percentage over the total number of ESP pairs for 10 arbitrarily chosen complexes from the Docking Benchmark v2.4.

Complex	Total Pairs	First ranked pair of Actual Contact Points	Percentage (%)
1AVX	438010	96	0.022
1CGI	238932	68	0.028
1F51	685587	891	0.1
1FAK	1033662	783	0.005
1FSK	1105528	726	0.065
1GCQ	91749	12	0.013
1HE1	353156	19	0.005
1JPS	1246835	333	0.026
1MLC	704308	207	0.03
1WEJ	719943	20	0.0027

3.5 Alignment and Final Geometric Scoring

In this section, the final stage of the proposed docking approach is described, which involves alignment and scoring of candidate poses. More specifically, the ligand L is translated and rotated with respect to the receptor R and the feasibility of each pose is calculated.

3.5.1 Alignment

Translation is performed by superimposing the centers of each pair of ESPs. Only the *k*-first ranked pairs of ESPs (i.e. the most complementary pairs, according to SID results) are taken into account.

While candidate translations can be easily retrieved from the SID results, the optimal rotation estimation for each translation is not straightforward. This is due to the fact that the SID descriptor is a rotation-invariant shape measure, thus, it does not provide information about the relative rotation between two interacting ESPs. In order to avoid the use of an exhaustive set of rotations, an initial alignment based on *solid vectors* [34] takes place. The solid vector of an ESP is defined below:

Let P and E be the center and radius of an ESP, respectively. Let also V be the solvent excluded volume of the molecule enclosed by the sphere $S(P,E)$, which is regarded as a homogeneous mass, and M its mass center. The solid vector $\mathbf{v}$ is the vector from P to M, as shown in Figure 3.9. For the alignment of two superimposed ESPs with respect to rotation, their corresponding solid vectors ($\mathbf{v}$ and $\mathbf{v}'$) are placed such that their angle ω is 180 degrees.

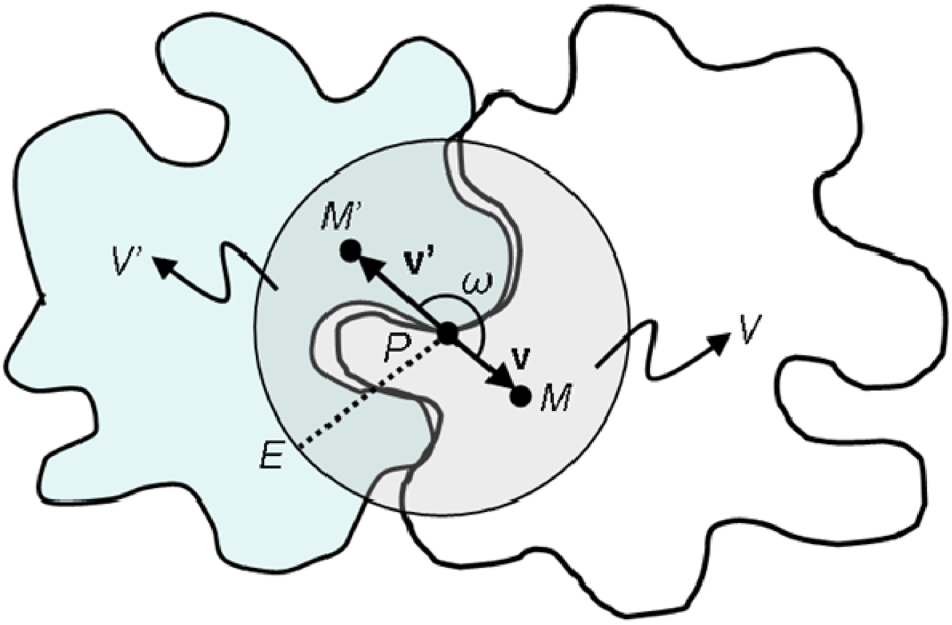


Figure 3.9: Alignment of two ESPs based on their solid vectors v and v'. The angle ω between the two solid vectors is 180 degrees.

The translation and rotation estimation described above provide only an approximation of the final pose. Small translations and rotations (after the initial alignment) should be also taken into account so as to achieve the best pose. Regarding rotation, the ligand ESP is firstly rotated about its solid vector in φ degrees intervals (Figure 3.10). This results in a set of $360/\varphi$ different poses. Then, the solid vector is rotated by θ degrees from its initial position and the ESP is rotated again around the solid vector, resulting in $360/\varphi$ more poses. The procedure is repeated several times, keeping the direction of the solid vector within a region of solid angle Ω (Figure 3.10).

Eventually, a set of N_{θ} uniformly sampled positions of the solid vector are retained, resulting in a total of $(N_{\theta} \times (360/\varphi))$ rotations.

Furthermore, the ligand, after the final superimposition, is translated from the receptor's possible contact point along several directions. The step is kept small (1Å), while the set of directions can be derived from the vertices of a regular polyhedron of radius 1 (e.g. icosahedron) in order to be uniformly distributed. If the 12 vertices of a regular icosahedron are used to model the set of small translations, a total of 13 translations is required. If it is combined with the set of $(N_{\theta} \times (360/\varphi))$ rotations, it results in $N_{Poses} = 13 \times (N_{\theta} \times (360/\varphi))$ different poses for each pair of ESPs. For each of these N_{Poses} poses, a scoring is computed based on the distance transform grid and the pose with the best score is finally selected.

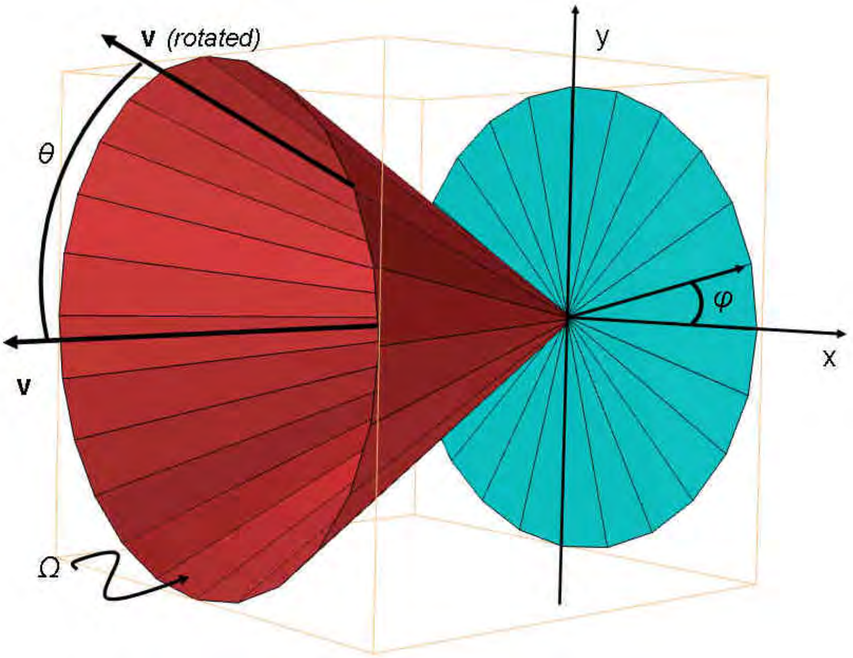


Figure 3.10: Rotations of the ligand ESP, after first alignment based on solid vector: angle φ corresponds to rotations about the solid vector v. Angle θ corresponds to rotations of the solid vector from its initial position. The direction of the solid vector is kept within a region of solid angle Ω.

In order to avoid the use of N_{Poses} poses, an alternative alignment method has been also investigated. The method was inspired by the approach presented in [119] for 3D surface registration. According to this method, two points **p**$_1$, **p**$_2$ from one surface along with their normal vectors **n**$_1$, **n**$_2$ and two corresponding points **p'**$_1$, **p'**$_2$ from another surface and their normal vectors $\mathbf{n'}_1$, $\mathbf{n'}_2$ are required to directly align the two surfaces without the need for exhaustive rotations. In our case, the points $\mathbf{p}_1$, $\mathbf{p}_2$, $\mathbf{p'}_1$ and $\mathbf{p'}_2$ are the centers of ESPs and the alignment is performed only between pairs of complementary ESPs, i.e. those retained from the complementarity matching step.

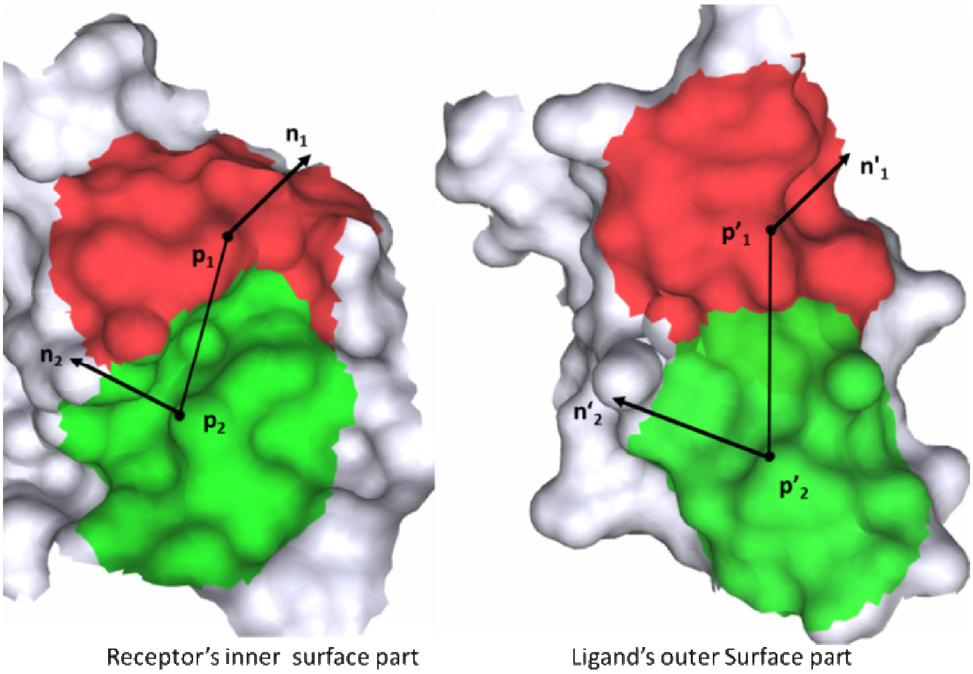


Figure 3.11: Alignment of two protein surfaces (complex 1AHW) based on two pairs of complementary ESPs.

In Figure 3.11, the alignment of the receptor and ligand surfaces of the complex 1AHW, based on two pairs of complementary ESPs, is shown. The alignment procedure consists of the following steps:

Step-1: if $|| \mathbf{p}_1 - \mathbf{p}_2 || \approx || \mathbf{p'}_1 - \mathbf{p'}_2 ||$, then, proceed to *step-2*, otherwise, reject that pair ($\mathbf{p'}_1$ and $\mathbf{p'}_2$) and repeat the procedure for another pair.

Step-2: if $(\mathbf{n}_1 \wedge \mathbf{n}_2) \approx (\mathbf{n'}_1 \wedge \mathbf{n'}_2)$, then, proceed to *step-3*, otherwise, reject that pair and repeat the procedure.

Step-3: compute the rotation matrix *R* that aligns the normal vectors $\mathbf{n}_1$ and $\mathbf{n'}_1$.

Step-4: estimate the rigid transformation that aligns points $\mathbf{p}_1$, $\mathbf{p}_2$, with $\mathbf{p'}_1$, $\mathbf{p'}_2$. This consists of translating $\mathbf{p}_1$, $\mathbf{p}_2$ by $\mathbf{p'}_1 - \mathbf{p}_1$, rotate them so that $\mathbf{n'}_1$ is aligned with $\mathbf{n}_1$, and, finally, rotate around $\mathbf{n'}_1$ to align $\mathbf{p}_2$ with $\mathbf{p'}_2$. A more detailed description of the algorithm is available in [119].

3.5.2 Geometric Scoring

For the geometric scoring of each pose, a method based on a 3D distance grid [33] has been implemented. The SES of the receptor R is inserted in a bounding rectangle divided in equally sized voxels and a 3D function $DT(i,j,k)$ is used to represent the value of each voxel. The sign of $DT(i,j,k)$ is given as:

$$DT(i,j,k)=\begin{cases} 0, \textit{if at least one surface point lies inside the voxel} \\ <0, \textit{if the voxel lies inside the molecule} \\ >0, \textit{if the voxel lies outside the molecule} \end{cases}$$

The absolute value in each voxel corresponds to the Euclidean distance from the closest surface point. Then, the distance grid is divided into shells according to the distance from the molecular surface. In our implementation, 5 shells are used, which are presented in Table 3.2. The ranges of the shells have been experimentally determined.

Table 3.2: The shells in which the distance grid is divided.

Shell 1	[1.4, ∞)	The range (in Å) of the first shell of the distance grid
Shell 2	[-0.8, 1.4)	The range of the second shell of the distance grid
Shell 3	[-1.8, -0.8)	The range of the third shell of the distance grid
Shell 4	[-3.2, -1.8)	The range of the fourth shell of the distance grid
Shell 5	[–∞, -3.2)	The range of the fifth shell of the distance grid
$w_{1\text{-}5}$	0, 1, -7, -10, -27	The values of the weights in the scoring function (equation 12)

The scoring of each pose is calculated as follows: the molecular surface of the ligand L, after translation and rotation, enters the 3D distance grid of the receptor R. L's surface points access the voxels of the 3D grid and are assigned a value according to the distance from R's molecular surface. The score of the transformation is given by:

$$Score = \sum_{i=1}^{5} w_i N_i \tag{3.11}$$

where N_i is the number of L points in shell i of the distance grid and w_i the weight of *i*-th shell (Table 3.2). The above equation can be modified to better represent the surface of the ligand L in each shell, as follows:

$$Score = \sum_{i=1}^{5} w_i \left(\sum_{j=1}^{N_i} s_{ij} \right) \tag{3.12}$$

where N_i is the number of ligand triangles whose centroids lie in *i*-th shell, w_i the weight of *i*-th shell and s_{ij} the area (in Å^2) of *j*-th triangle of *i*-th shell.

The 3D distance grid provides an accurate measure for geometric scoring of candidate poses. The computation time required for this process is proportional to the size as well as the resolution of the ligand's molecular surface. In the alignment step of the proposed method N_{Poses} different poses of the ligand are taken for each pair of complementary ESPs. In order to achieve low computation times without affecting the accuracy of scoring, two different resolutions of the ligand molecular surface are used. For the low-resolution surface, a point density of 1 point per Å^2 was chosen as parameter to MSMS algorithm [10], while for the high-resolution surface a density of 4 points per Å^2 was chosen. The low-resolution surface is used to score the entire set of N_{Poses} poses, during the first step of the scoring procedure. After filtering out the majority of poses, only the poses with the highest scores are used for high-resolution scoring. Finally, the pose with the highest score is kept for each pair of ESPs. The first scoring step may become even faster if instead of the entire SES of the ligand only the part that belongs to the corresponding ESP is used. In this case, the filtering criteria to exclude poses at the first step are given below:

$$\frac{\sum_{j=1}^{N_1} s_{1j}}{\sum_{j=1}^{N_{Total}} s_j} > 0.5 \tag{3.13}$$

$$\frac{\sum_{j=1}^{N_2} s_{2j}}{\sum_{j=1}^{N_{Total}} s_j} < 0.1 \tag{3.14}$$

$$N_3 = 0, N_4 = 0 \tag{3.15}$$

where N_{Total} is the total number of triangles of the ligand ESP and s_j is the area of each triangle. The first criterion implies that at least half of the area of the ligand ESP should lie within a region close to the surface of the receptor, while the last two criteria imply that very deep penetrations are not allowed.

3.6 Experimental Results

The proposed method was experimentally evaluated using the protein-protein docking benchmark v2.4 [90]. This dataset consists of 84 known complexes, with 63 rigid-body cases, 13 cases of medium difficulty, and 8 cases of high difficulty with substantial conformational change.

To evaluate the performance of the method, for each complex of the dataset, the receptor and ligand are separated from each other and the ligand is translated and rotated arbitrarily. Then, the docking algorithm described in the previous sections is applied to generate a set of candidate poses of the ligand. A predicted pose is called a hit if the interface Root Mean Square Deviation (RMSD) between the ligand in that pose and the ligand in the original complex is less than a predefined threshold. The interface RMSD is calculated over the interface *Ca* atoms of the ligand. The value of the predefined threshold was selected to be 2.5Å.

3.6.1 Comparison with Context Shapes, ZDOCK and PatchDock

The results of the proposed method were compared to those of the following three methods: a) Context Shapes (CS) [34], b) ZDOCK (PSC) [26] and c) PatchDock [33]. The first and the third method belong to the category of "local shape feature matching" approaches, while the second is a brute force approach. ZDOCK(PSC) returns a maximum of 3600 predictions, therefore, only the top 3600 predictions are taken into account for all methods. More specifically, for the proposed approach, the number of *k*-first selected pairs after the SID complementarity matching was set to 3600 in order to be comparable to the other methods. In our experiments, the *R-bound/L-bound* case was evaluated. In this case, the receptor and ligand are both bound, i.e., the receptor and the ligand from the co-crystallized protein complexes are used. The performance of the above three methods was computed by using the executables taken from the home pages of the authors: http://www.cs.rpi.edu/~zaki/software/ContextShapes/ for Context Shapes, http://zlab.bu.edu/zdock/ for ZDOCK v2.1 and http://bioinfo3d.cs.tau.ac.il/PatchDock/ for PatchDock.

The method has been optimized by training on a small dataset (20 complexes) of the docking benchmark v0.0 [91]. This dataset was selected so as not to include complexes common in benchmark v2.4. The dataset is depicted in Table 3.3.

The set of parameters that required optimization is given in Table 3.4. Each of these parameters has been assigned several values, during the training procedure. Those values that produced better docking results on this dataset were selected for the experiments in benchmark v2.4.

Table 3.3: Selected training dataset from Docking Benchmark v0.0.

1	1CHO(E:I)	11	1BQL(LH:Y)
2	2PTC(E:I)	12	1NMB(LH:N)
3	1TGS(Z:I)	13	1MEL(B:M)
4	1CSE(E:I)	14	2VIR(AB:C)
5	2KAI(AB:I)	15	1EO8(LH:A)
6	1BRC(E:I)	16	1AVZ(B:C)
7	1BRS(A:D)	17	1MDA(LH:A)
8	1UGH(E:I)	18	1SPB(S:P)
9	1FSS(A:B)	19	1BTH(LH:P)
10	1AVW(A:B)	20	1FIN(A:B)

Table 3.4: The set of parameters that required optimization. In the first column, the abbreviation of the parameter as stated in the text is given. In the second and third column, the optimal value and the description of each parameter are given, respectively.

Abbreviation	Optimal Value	Description
E	10Å	The radius of the sphere that determines the

		size of an ESP (Section 3.1)
G_{max}	12 Å	The maximum allowed geodesic distance from the center of ESP (Section 3.1)
φ	22.5^{O}	The angle interval (in degrees) for rotations of the ESP about the solid vector (Section 4.1)
Ω	0.068π	The solid angle within which the solid vector is rotated (Section 4.1)
N_{ϑ}	9	The number of uniformly sampled positions of the solid vector (Section 4.1)
N_{Poses}	1872	The total number of different poses for each pair of ESPs (Section 4.1)

In Table 3.5, the performance of the proposed method in benchmark v2.4, compared with the other three methods, is depicted. In the first column for each method, the rank of the best ranked hit is presented. This is not necessarily the hit with the smallest RMSD value, it is the first result of the rank list that produces RMSD less than 2.5Å. In the second column for each method, the RMSD value of the best ranked hit is given. In complexes where these values are missing, the method failed to return a hit within the first 3600 predictions. In 7 cases none of the four methods returned a hit in the first 3600 predictions, thus, they are not stated in Table 3.5.

Table 3.5: R-bound/L-bound: Comparisons between the proposed method, Context Shapes, ZDOCK(PSC) and PatchDock on 84 test cases from Benchmark v2.4. PDB gives the PDB id for the protein complex. RMSD and Rank give the RMSD and rank of the best ranked hit (using 2.5 Å cut-off). In 7 cases none of the four methods returned a hit in the first 3600 predictions, thus, they are not stated.

	Proposed Method		Context Shapes		ZDOCK(PSC)		PatchDock	
PDB	Rank	RMSD	Rank	RMSD	Rank	RMSD	Rank	RMSD
1A2K	**17**	**0.92**	40	1.08	570	2.41	300	1.47
1ACB	13	1.89	8	2.32	**6**	**0.82**	10	1.60
1AHW	167	1.8	**7**	**1.20**	56	1.18	40	1.55
1AK4	**908**	**0.52**	2925	2.08	3471	1.14	-	-
1AKJ	**174**	**1.67**	265	2.15	448	1.88	-	-
1ATN	223	1.64	**49**	**2.10**	558	1.15	-	-
1AVX	7	1.71	10	1.76	**1**	**1.96**	43	2.14
1AY7	**23**	**2.69**	193	1.23	46	1.68	24	2.07
1B6C	**3**	**2.26**	11	1.78	24	1.69	40	1.92
1BGX	**1**	**2.51**	**1**	**1.96**	-	-	-	-

1BJ1	-	-	**1**	**1.05**	3	1.42	-	-
1BUH	**49**	**1.58**	61	1.55	393	1.43	83	1.14
1BVK	249	5.05	**45**	**1.69**	1087	1.43	131	2.12
1BVN	**1**	**0.99**	**1**	**1.55**	10	1.24	**1**	**0.75**
1CGI	**1**	**0.72**	**1**	**1.37**	**1**	**1.12**	**1**	**1.08**
1D6R	**2**	**1.31**	4	1.68	35	1.04	-	-
1DE4	538	2.17	**13**	**1.21**	452	1.62	-	-
1DFJ	**1**	**1.08**	-	-	-	-	-	-
1DQJ	49	1.12	67	1.65	**19**	**2.00**	83	1.71
1E6E	34	2.4	**1**	**1.58**	58	2.06	2	2.29
1E6J	**526**	**2.31**	1337	1.92	699	2.02	1706	1.43
1E96	**809**	**2.32**	1206	1.84	-	-	1767	1.44
1EAW	**1**	**1.95**	**1**	**1.41**	**1**	**1.75**	**1**	**0.99**
1EER	**1**	**1.16**	**1**	**1.62**	-	-	**1**	**1.66**
1EWY	**103**	**2.45**	518	2.26	-	-	139	1.42
1EZU	**1**	**2.07**	**1**	**1.60**	-	-	**1**	**0.94**
1F34	**1**	**2.4**	**1**	**1.99**	-	-	**1**	**1.90**
1F51	3	1.18	7	2.01	-	-	**1**	**1.92**
1FAK	**119**	**1.47**	1997	1.70	-	-	-	-
1FC2	**2**	**2.34**	7	1.85	55	2.18	49	1.24
1FQJ	**8**	**1.62**	12	1.94	120	1.94	248	1.48
1FSK	145	0.99	**9**	**2.06**	19	1.70	218	1.57
1GCQ	**1**	**1.16**	2	1.26	382	1.81	-	-
1GP2	551	2.07	**53**	**1.86**	-	-	-	-
1GRN	**1**	**1.66**	**1**	**1.84**	7	2.26	3	1.45
1H1V	49	1.75	**14**	**2.37**	1510	2.40	-	-
1HE1	**1**	**0.8**	**1**	**1.44**	7	1.67	**1**	**1.06**
1HIA	8	1.05	2	1.07	**1**	**1.70**	14	1.19
1I2M	**1**	**0.86**	6	1.36	14	1.80	-	-
1I4D	1278	2.48	**104**	**1.42**	793	2.08	167	1.05
1I9R	**142**	**1.38**	-	-	1271	2.04	-	-
1IB1	**1**	**1.87**	2	1.48	-	-	-	-
1IBR	**1**	**2.01**	**1**	**2.05**	-	-	-	-
1IQD	531	1.09	**14**	**1.19**	55	1.83	-	-
1JPS	216	1.68	**2**	**1.26**	23	2.30	96	1.87
1K4C	712	2.21	**5**	**0.88**	30	1.16	337	1.53
1K5D	59	1.98	**2**	**2.06**	10	2.11	-	-
1KAC	**14**	**1.5**	-	-	381	1.52	-	-
1KKL	**158**	**2.68**	226	1.67	-	-	-	-
1KLU	**899**	**2.28**	1108	1.80	-	-	-	-
1KTZ	**240**	**1.84**	2280	1.41	-	-	-	-
1KXP	**2**	**1.87**	3	2.17	-	-	-	-

1KXQ	**3**	**1.35**	229	1.51	30	1.60	29	1.63
1M10	**6**	**2.4**	-	-	33	2.23	-	-
1MAH	**1**	**1.44**	**1**	**1.45**	**1**	**1.91**	**1**	**1.27**
1ML0	532	2.46	569	1.91	75	1.94	**7**	**0.58**
1MLC	955	1.86	**30**	**1.15**	1205	1.37	516	1.79
1N2C	8	2.12	**3**	**1.36**	-	-	-	-
1NCA	76	2.21	**3**	**1.77**	20	1.48	-	-
1NSN	**149**	**1.15**	-	-	-	-	-	-
1PPE	**1**	**1.38**	**1**	**2.32**	2	1.21	**1**	**1.03**
1QA9	**3**	**2.18**	972	1.30	-	-	-	-
1QFW	1013	2.39	1247	2.21	**16**	**2.46**	-	-
1RLB	591	1.6	**311**	**1.63**	-	-	3143	2.32
1SBB	**962**	**1.72**	-	-	-	-	-	-
1TMQ	18	2.27	**1**	**2.32**	8	1.79	**1**	**1.52**
1UDI	**1**	**1.58**	3	1.52	**1**	**1.50**	**1**	**1.97**
1VFB	73	1.06	**8**	**1.50**	-	-	-	-
1WEJ	897	2.04	**496**	**1.25**	1120	1.11	-	-
1WQ1	**1**	**2.32**	**1**	**1.14**	4	2.04	**1**	**0.84**
2BTF	**2**	**1.45**	4	1.13	21	1.21	137	1.82
2JEL	377	2.15	**56**	**1.40**	532	1.77	282	1.65
2MTA	469	1.64	**21**	**1.45**	1447	2.26	115	1.71
2PCC	**5**	**2.28**	-	-	-	-	-	-
2SIC	**2**	**0.73**	4	1.36	9	1.19	-	-
2SNI	**1**	**1.78**	2	1.27	4	2.50	13	2.10
7CEI	**2**	**1.14**	123	1.90	5	2.18	-	-

Summing up the results of Table 3.5, the proposed approach failed to return a hit in 8 out of 84 cases, while Context Shapes failed in 13 cases, ZDOCK in 29 and PatchDock in 42 cases. In Table 3.6, the number of successful predictions for all methods is presented. It is clear from the results that the proposed method managed to return a hit in most of the cases, outperforming the other three methods. If we relax the RMSD cutoff threshold to 5Å, it is obvious that all methods achieve more successful predictions. Again, the proposed method outperforms the other three, since it fails only in three cases.

Table 3.6: R-bound/L-bound: Number of test cases where a hit is found within the top 3600 predictions, for each method, and the number of test cases where all three methods fail.

Proposed Method	Context Shapes	ZDOCK	PatchDock	All Fail

RMSD ≤ 2.5Å				
76	71	55	42	7
RMSD ≤ 5Å				
81	76	71	63	2

In Table 3.7, the 7 cases where all four methods failed are presented. It is worth to mention that in none of these cases could any of the above methods return a near-native solution among a set of 3600 predicted poses. These examples can really help towards improving existing docking approaches. Additionally, they provide an indication that geometric complementarity is not always the dominant factor in protein-protein docking but other non-geometric parameters (desolvation, hydrophobicity, electrostatics, etc.) should be also taken into account.

Table 3.7: R-bound/L-bound: The 7 cases where all three methods fail.

PDB IDs of the complexes where all methods failed (RMSD ≤ 2.5Å)			
1FQ1	1GHQ	1HE8	1IJK
2HMI	2QFW	2VIS	

In Table 3.8, the win-tie-loss-failure records for the proposed method versus Context Shapes, ZDOCK and PatchDock is presented. Comparing with Context Shapes, the proposed approach returns a better ranked hit in 39 cases, whereas Context Shapes returns a better hit in 25 cases. The methods tie in 13 cases, and both fail in 7 cases. Comparing against ZDOCK and PatchDock, the proposed method clearly outperforms them across all three scenarios; it has 56-17 win-loss record against ZDOCK and 52-13 win-loss record against PatchDock.

Table 3.8: R-bound/L-bound: the win-tie-loss-failure records for the proposed method versus Context Shapes, ZDOCK(PSC) and PatchDock.

Proposed Method vs	Win	Tie	Loss	Both fail
Context Shapes	39	13	25	7
ZDOCK	56	4	17	7
PatchDock	52	11	13	8

In Table 3.9, the results for the first ranked and the 10 best ranked solutions, with RMSD < 5 Å, using the proposed method, are presented. It is obvious that in 51 out of the 83 cases, at least one almost correct prediction with RMSD < 5 Å is ranked among the top 10 solutions.

Table 3.9: The numbers of solutions with RMSD < 5 Å, within the top-1 and top-10 ranked positions, using the proposed method.

PDB	Top 1	Top 10	PDB	Top 1	Top 10	PDB	Top 1	Top 10
1A2K	0	1	1F51	1	2	1KTZ	0	0
1ACB	0	1	1FAK	0	0	1KXP	1	3
1AHW	0	1	1FC2	0	1	1KXQ	0	2
1AK4	0	0	1FQ1	0	0	1M10	0	2
1AKJ	0	0	1FQJ	0	1	1MAH	1	1
1ATN	0	0	1FSK	0	1	1ML0	0	0
1AVX	0	2	1GCQ	1	2	1MLC	0	0
1AY7	0	1	1GHQ	0	0	1N2C	0	1
1B6C	0	2	1GP2	0	0	1NSN	0	0
1BGX	1	2	1GRN	1	2	1NCA	0	1
1BJ1	0	0	1H1V	0	1	1PPE	1	2
1BUH	0	1	1HE1	1	2	1QA9	0	1
1BVK	0	0	1HE8	0	0	1QFW	0	0
1BVN	1	3	1HIA	0	1	1RLB	0	0
1CGI	1	2	1I2M	1	3	1SBB	0	0
1D6R	0	1	1I4D	0	0	1TMQ	0	1
1DE4	0	0	1I9R	0	0	1UDI	1	2
1DFJ	1	2	1IB1	1	2	1VFB	0	1
1DQJ	0	1	1IBR	1	1	1WEJ	0	0
1E6E	0	1	1IJK	0	0	1WQ1	1	3
1E6J	0	0	1IQD	0	0	2BTF	0	1
1E96	0	0	1JPS	0	1	2HMI	0	0

1EAW	1	1	1K4C	0	0	2JEL	0	0
1EER	1	2	1K5D	0	1	2MTA	0	0
1EWY	0	1	1KAC	0	1	2PCC	0	1
1EZU	1	2	1KKL	0	0	2SIC	0	2
1F34	1	3	1KLU	0	0	2SNI	1	2
						2VIS	0	0
						7CEI	0	1

Median RMSD can also provide a useful performance measure. In Table 3.10, the median/min/max RMSD and Rank for the 10 best ranked and 25 best ranked solutions, using the proposed method, are presented. These values were obtained over the entire test dataset.

The above experiments have been performed using the bound molecules of both the receptor and the ligand (R-bound/L-bound). This is due to the fact that none of the above methods, including the one presented in this chapter, is able to efficiently model the side-chain conformations during flexible docking. Experiments for the R-bound/L-bound case were performed to measure the efficiency of the geometric-only algorithms in the ideal case of rigid-body docking. In order to measure the robustness of the proposed method with respect to conformational changes, a set of experiments were performed in Benchmark v2.4 for the R-unbound/L-bound case. In this case, the receptor is taken from the unbound form of the protein, while the ligand is taken from the bound co-crystallized complex.

Table 3.10: The median/min/max RMSD and Rank for the best solution, within the top-10 and top-25 ranked positions, using the proposed method.

	Top 10	Top 25
Median RMSD	3.29	2.34
Minimum RMSD	0.72	0.6
Maximum RMSD	21.03	13.91
Median Rank	5	8
Minimum Rank	1	1

Maximum Rank	10	25

In Table 3.11, the number of successful predictions for all methods, for the R-unbound/L-bound case, is presented. It is clear that the performance of all methods is significantly reduced, comparing with the R-bound/L-bound case. However, the performance of the proposed method is still higher.

Table 3.11: R-unbound/L-bound: Number of test cases where a hit is found within the top 3600 predictions, for each method, and the number of test cases where all three methods fail.

Proposed Method	Context Shapes	ZDOCK	PatchDock	All Fail
RMSD $\leq$ 2.5Å				
46	43	33	22	31
RMSD $\leq$ 5Å				
54	52	52	50	18

Similar conclusions can be drawn in the win-tie-loss-failure records (Table 3.12). Comparing with Context Shapes, the proposed approach returns a better ranked hit in 29 cases, whereas Context Shapes returns a better hit in 24 cases. Both methods fail in 31 cases. Comparing against ZDOCK and PatchDock, the proposed method outperforms them; it has 31-20 win-loss record against ZDOCK and 34-12 win-loss record against PatchDock. For the R-unbound/L-unbound case, where both the receptor and the ligand are unbound, all four methods fail to return a hit in more than half of the complexes, which implies that a solution able to efficiently deal with flexibility is needed.

Table 3.12: R-unbound/L-bound: the win-tie-loss-failure records for the proposed method versus Context Shapes, ZDOCK(PSC) and PatchDock.

Proposed Method vs	Win	Tie	Loss	Both fail
Context Shapes	29	0	24	31
ZDOCK	31	0	20	33
PatchDock	34	0	12	38

3.6.2 Performance Analysis of the proposed method

In Table 3.13, the numbers of the ESPs (centered at convex and concave critical points) for the receptor and ligand, as well as the total number of ESP pairs are presented for the 84 test cases of benchmark v2.4.

The numbers of receptor and ligand atoms for each complex are also included for the sake of completeness. The table shows that the number of generated ESPs as well as the number of ESP pairs is almost proportional to the number of receptor and ligand atoms. In Figure 3.12, the scatter-plot of the combined receptor+ligand size (number of atoms) versus the number of ESP pairs is depicted. It can be inferred that when the total number of receptor and ligand atoms increases, then the number of ESP pairs increases as well.

Table 3.13: Number of ESPs and ESP pairs for the receptor and ligand in benchmark v2.4. The numbers of atoms are also shown for completeness.

PDB	Number of Atoms		Number of ESPs				Number of ESP pairs
	Receptor	Ligand	Receptor		Ligand		
			Convex	Concave	Convex	Concave	
1A2K	1990	1570	542	780	447	628	689036
1ACB	1769	522	482	676	187	233	238718
1AHW	3304	1612	876	874	526	531	924880
1AK4	1266	1062	371	467	403	499	373330
1AKJ	3075	1814	905	916	524	528	957824
1ATN	2907	2035	771	843	505	579	872124
1AVX	1630	1286	418	477	394	404	356810
1AY7	746	720	252	342	231	290	152082
1B6C	831	2602	274	352	719	803	473110
1BGX	3245	6570	806	864	1452	1465	2435318
1BJ1	3307	1522	906	939	507	588	1008801
1BUH	2311	605	634	714	200	285	323490
1BVK	1744	1001	454	549	318	376	345286
1BVN	3907	536	879	938	202	241	401315
1CGI	1799	440	463	528	195	206	198338
1D6R	1629	427	413	503	189	223	187166
1DE4	3063	10044	772	803	1949	1980	3093607
1DFJ	951	3411	316	423	815	893	626933
1DQJ	3244	1001	856	862	302	388	592452
1E6E	3518	859	960	1017	302	367	659454
1E6J	3275	1639	894	904	596	671	1138658

1E96	1419	1502	408	585	429	515	461085
1EAW	1864	2310	500	524	174	198	190176
1EER	1291	3328	456	581	661	751	726497
1EWY	2492	749	705	763	249	301	402192
1EZU	1656	2198	425	491	787	832	740017
1F34	2423	1074	622	772	405	502	624904
1F51	2993	940	873	927	277	296	515187
1FAK	2782	1495	722	732	468	513	712962
1FC2	354	1656	150	184	561	636	198624
1FQ1	1439	2402	450	572	693	751	734346
1FQJ	2611	1111	728	803	368	380	572144
1FSK	3347	1230	729	802	411	514	704328
1GCQ	468	558	199	226	192	243	91749
1GHQ	2417	987	590	632	367	453	499214
1GP2	2788	3021	859	926	768	771	1373457
1GRN	1494	1586	479	633	446	519	530919
1H1V	2875	2539	791	859	718	762	1219504
1HE1	997	1374	340	432	398	433	319156
1HE8	6070	1326	1501	1610	397	416	1263586
1HIA	1787	353	469	570	173	160	173650
1I2M	1346	2899	410	526	714	800	703564
1I4D	3004	1381	819	873	424	467	752625
1I9R	3276	3297	790	852	884	940	1495768
1IB1	3642	1404	1045	1192	438	488	1032056
1IBR	1371	3573	418	471	1036	1117	954862
1IJK	2071	1595	597	613	410	511	556397
1IQD	3089	1246	839	888	367	422	679954
1JPS	3247	1611	858	884	518	578	953836
1K4C	3252	765	887	980	322	381	653507
1K5D	2868	2698	790	871	716	732	1201916
1KAC	3805	625	396	461	304	341	275180
1KKL	1401	959	974	983	218	266	473378
1KLU	3028	1880	838	921	530	626	1012718
1KTZ	653	840	267	333	289	359	192090
1KXP	2736	3431	723	851	993	923	1512372
1KXQ	3910	916	803	844	187	301	399531
1M10	1601	2087	433	571	594	607	602005
1MAH	4116	460	940	1025	193	206	391465
1ML0	5706	515	2150	2169	302	332	1368838
1MLC	3290	1001	310	394	902	1090	693288
1N2C	15926	4132	854	913	667	749	1248617
1NSN	3282	1108	920	991	337	426	725887

1NCA	3329	3075	873	901	669	747	1254900
1PPE	1629	222	410	503	125	101	104285
1QA9	846	776	305	406	351	278	227296
1QFW	1762	1476	504	580	478	581	570064
1RLB	3760	1453	997	1008	444	518	963998
1SBB	1826	1975	563	662	531	630	706212
1TMQ	3598	881	793	820	298	359	529047
1UDI	1818	654	489	594	241	302	290832
1VFB	1730	1001	459	561	312	381	349911
1WEJ	3340	868	901	1055	270	293	548843
1WQ1	2533	1322	719	873	418	528	744546
2BTF	2917	1044	780	885	292	333	518160
2HMI	7630	3264	1377	1346	869	877	2377303
2JEL	3297	640	883	1004	238	285	490607
2MTA	3853	807	938	1059	256	328	578768
2PCC	2371	847	622	712	292	395	453594
2SIC	1938	764	441	585	277	334	309339
2SNI	1938	513	442	560	187	242	211684
2VIS	3261	2076	895	993	548	652	1127704
7CEI	698	1026	248	331	365	499	244567

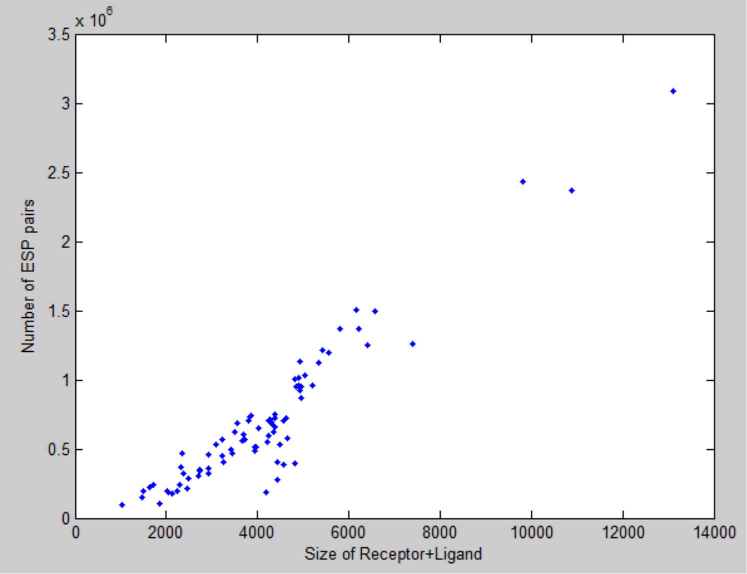


Figure 3.12: Scatter-plot of receptor plus ligand size versus the total number of ESP pairs for each complex. When the total number of receptor and ligand atoms increases, then the number of ESP pairs increases as well.

In Table 3.14, the average computation times for various tasks of the proposed approach are presented. The average time required for extraction of the SID descriptor for an ESP is 0.6s. The SID descriptor extraction time, as well as the time for SES computation is not included in the average running time. These tasks belong to the pre-processing step and are computed off-line. Likewise, the times to calculate the context shapes in Context Shapes method, the SES for PatchDock method and surface residues for ZDOCK method, are also not included in the average running time.

The time required for SID-based matching between a pair of ESPs is less than 0.02ms, since it is based on simple histogram matching. It is obvious that SID descriptor matching is 10000 times faster than the geometric scoring based on distance grid, which demonstrates the importance of the SID descriptor as a fast filtering stage, during the docking procedure. This is made clearer in Table 3.15, where the average running times for the four methods across all 84 test cases are presented. In our approach, the running time is the sum of the time required for SID descriptor matching and the time needed for geometric scoring. Even though geometric scoring is applied to a much smaller set of ESPs (the 3600 first ranked pairs), it lasts longer than SID matching. Comparing with the other methods, the proposed docking approach achieves faster computation time. It is more than two times faster than the Context Shapes approach, more than three times faster than ZDOCK and faster than PatchDock.

Table 3.14: Average computation times for various tasks of the proposed approach

Activity	Average Computation Time
SID Descriptor Extraction / ESP	0.6s
SID matching of a pair of ESPs	0.019ms
Scoring (distance grid) of a pair of ESPs	196ms

The average pre-processing time for a protein in benchmark v2.4, using the proposed method, is about 720s and for a pair of interacting proteins is about 1440s. This results in a total pre-processing and running time of 2280s. This is still faster than ZDOCK and comparable to ContextShapes, while PatchDock, which involves fewer steps in preprocessing, is faster than the proposed method.

Table 3.15: Average running time over all 84 test cases

Method	Average Running Time (SID matching)	Average Running Time (Geometric Scoring)	Average Running Time
Proposed Approach	135s	705s	840s
Context Shapes			2031s
ZDOCK			2914s
PatchDock			1098s

The times were obtained using a PC with a dual-core 2.4 GHz processor and 8GB RAM. The executable files of the proposed method can be downloaded for testing from the authors' website (http://3d-test.iti.gr:8080/3d-test/Images/ProteinDocking.zip).

3.7 Summary

In this chapter, a new framework for fast geometric protein-protein docking was presented. After extraction of the Solvent Excluded Surface, a set of critical points is formed based on the local curvature of the surface. Then, for each critical point an Extended Surface Patch (ESP) is generated, centered at the critical point with radius 10Å. The shape complementarity of all pairs of ESPs between the receptor and the ligand is measured using the Shape Impact Descriptor (SID), which is a fast rotation-invariant shape descriptor. The complementarity matching between two patches is reduced to a simple histogram matching of their SID Descriptors, without the need for taking an exhaustive set of rotations for each pair of patches. For the final scoring step, only a very small subset of the most complementary ESP pairs is given as input, significantly reducing the computation time. The proposed approach was evaluated against three state-of-the-art methods for geometric docking. Not only it achieved more successful predictions in benchmark v2.4, but also reduced two or even three times the computation time, due to the efficiency of the Shape Impact Descriptor.

Chapter 4

SP-Dock: Protein-Protein Docking using Shape and Physicochemical Complementarity

4.1 Introduction

In this chapter, a framework for protein-protein docking is proposed, which exploits both shape and physicochemical complementarity to generate improved docking predictions. Shape complementarity is achieved by matching local surface patches. However, unlike existing approaches, which are based on single-patch or two-patch matching, we developed a new algorithm that compares simultaneously, groups of neighboring patches from the receptor with groups of neighboring patches from the ligand. Taking into account the fact that shape complementarity in protein surfaces is mostly approximate rather than exact, the proposed group-based matching algorithm fits perfectly to the nature of protein surfaces. This is demonstrated by the high performance that our method achieves especially in the case where the unbound structures of the proteins are considered. Additionally, several physicochemical factors, such as desolvation energy, electrostatic complementarity, hydrophobicity, Coulomb potential and Lennard-Jones potential are integrated using an optimized scoring function, improving geometric ranking in more than 60% of the complexes of Docking Benchmark 2.4.

The chapter is organized as follows: in Section 4.2, an overview of the method and its major scientific contributions is presented. In Section 4.3, the preprocessing phase is described, which includes the surface representation and extraction of local patches, as well as the local shape descriptor extraction for each patch. Section 4.4 analyzes the new group-based matching and alignment algorithm, while in Section 4.5 the geometric and physicochemical scoring procedure of the candidate docking poses is given. Concerning the physicochemical scoring, the optimization process that assigns a set of weights for each physicochemical factor is provided.

Then, in Section 4.6, the experimental results are presented, where the proposed method is compared to other existing docking approaches. Finally, the chapter is summarized in Section 4.7.

4.2 Overview and Contributions

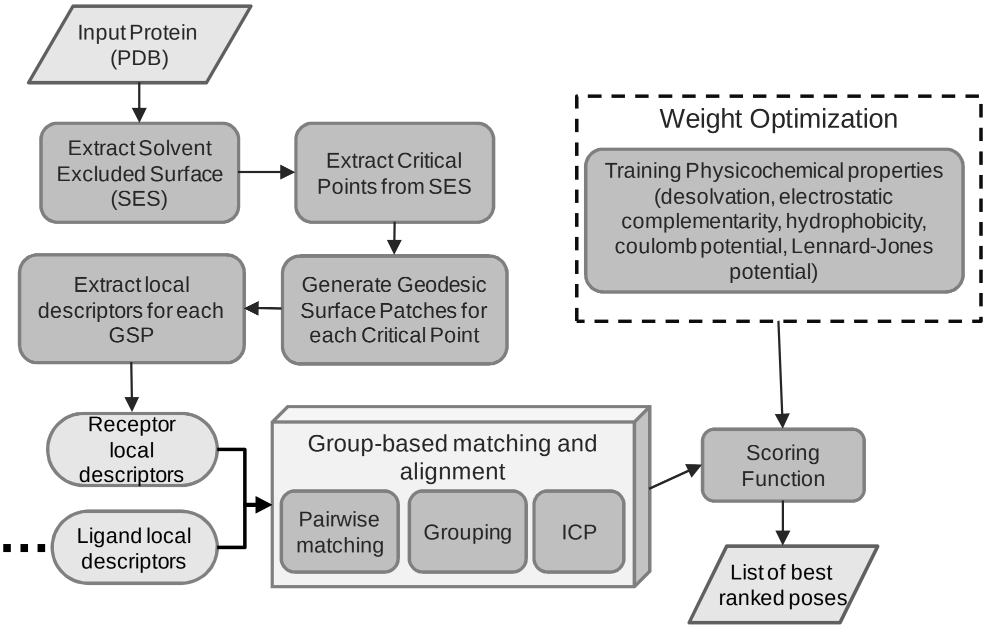

Input Protein (PDB)
Extract Solvent Excluded Surface (SES)
Extract Critical Points from SES
Generate Geodesic Surface Patches for each Critical Point
Extract local descriptors for each GSP
Receptor local descriptors
Ligand local descriptors
Group-based matching and alignment
Pairwise matching
Grouping
ICP
Weight Optimization
Training Physicochemical properties (desolvation, electrostatic complementarity, hydrophobicity, coulomb potential, Lennard-Jones potential)
Scoring Function
List of best ranked poses

Figure 4.1: Block diagram of the proposed method.

In Figure 4.1, the block diagram of the proposed method is depicted. The PDB files of the receptor and ligand proteins are given as input and their Solvent Excluded Surfaces (SESs) are extracted. Then, by computing the curvature of the SES, a set of critical points is extracted, which correspond to the centers of small elementary patches (either convex or concave). Each elementary patch is expanded in size in order to cover a wider area producing a Geodesic Surface Patch (GSP). For each GSP an appropriate local shape descriptor is extracted, which uniquely characterizes its shape. During complementarity matching, each GSP that corresponds to a convex (or concave) elementary patch of the receptor protein is matched with all GSPs that correspond to concave (or convex) elementary patches of the ligand protein. As a next step, several neighboring GSPs are grouped together to generate candidate binding regions on the surfaces of the receptor and the ligand. For aligning the two proteins, one candidate region of the ligand is aligned with respect to a complementary candidate region of the receptor using the Iterative Closest Point (ICP) algorithm. At the final step of the algorithm, the aligned poses are scored using both geometric and physicochemical properties. The weights of the scoring function are optimized via training to achieve improved docking results.

Comparing the proposed SP-Dock (Shape-Physicochemical Docking) method with the approach presented in the previous chapter, both are based on local shape feature matching of surface patches corresponding to convex or concave elementary shape patches. However, numerous novel features are introduced in SP-Dock, which are explained below.

First of all, *a more discriminative local surface descriptor* has been adopted in SP-Dock for patch complementarity matching, instead of the Shape Impact Descriptor (SID) that was initially used. Some of the most well-known local descriptors [92] were compared to SID in a dataset of known complexes to select the most appropriate descriptor. The Local Spectral Descriptor [93] has been proven to be the most discriminative among others.

Another notable innovation of SP-Dock is *the group-based matching algorithm*. It introduces a new approach for shape complementarity matching beyond traditional local shape feature matching techniques. It has been inspired by the fact that the shape complementarity between a pair of local surface patches (one from the receptor and one from the ligand), which correspond to a near-native pose, is mostly approximate rather than exact, while at the same time there are plenty of pairs of patches corresponding to non-native poses that have similar or even better shape complementarity than the near-native ones. Thus, existing local shape matching approaches, which rely on single-patch-to-single-patch or two-patch-to-two-patch complementarity matching, may predict a large number of false-positive docking poses and fail to detect near-native poses. The approach presented in this chapter intuitively groups neighboring patches from both the receptor and the ligand so as to create larger candidate binding regions. This increases the confidence of a receptor patch to be complementary to a ligand patch, since, according to the grouping criterion, the neighbours of the receptor patch should be complementary to the neighbors of the ligand patch as well. The effectiveness of the proposed approximate complementarity matching is convincingly reflected in the unbound docking case where SP-Dock clearly outperforms similar docking approaches.

Additionally, SP-Dock *proposes the adoption of the Iterative Closest Point (ICP) algorithm for fast alignment of the complementary candidate regions*. ICP has been extensively used for surface registration in 3D reconstruction problems. Although 3D reconstruction involves alignment of surfaces with near-exact similarity, we prove that ICP is also appropriate for aligning surfaces with approximate similarity, as is the case of geometric docking. It is the first time, to the best of our knowledge, that ICP has been used for alignment of protein surfaces. It is also worth mentioning that surface similarity is equivalent to surface complementarity, if the surface of the ligand is turned upside-down, as it has been already proven in the previous chapter.

Finally, *the contribution of physicochemical factors to achieve more accurate docking predictions is assessed*. Several non-geometric factors, namely the Atom Desolvation Energy, Interface Residue Contact Preferences, Generic Residue Contact Preferences, Electrostatic Complementarity, Coulomb Potential, Hydrophobicity and Van der Waaks Potential, were computed and combined with the geometric properties into a unified scoring function. These factors have been already discussed in previous works and are summarized in [39]. In this work, the contribution of each factor is assessed and the optimal weight, with which each factor participates in the scoring function, is estimated using an appropriately selected optimization method. The improvement of docking predictions by combining the geometric with the physicochemical factors, in Docking benchmark 2.4, is impressive.

As it is a local feature matching method, the proposed algorithm shares similarities with the well-known PatchDock method [33]. More specifically, the step of critical points extraction produces similar sparse surface representations for both PatchDock and SP-Dock (although a different algorithm is used in each case to extract the critical points). The geometric scoring step is also similar in both methods, since they generate a 3D distance grid around the receptor, which is accessed by the surface points of the ligand. On the other hand, their surface complementarity matching stages, which constitute core parts of the docking process, are completely different. First of all, SP-Dock does not rely on shape matching of the small convex and concave patches of the sparse surface, but it generates bigger surface patches (the GSPs), which cover a wider surface area around a critical point. These GSPs enclose more significant shape information than the local patches of PatchDock, which is important especially in filtering out a lot of false positive matches. Additionally, instead of the rather simple geometric features that describe the shape of a patch (or a pair of patches) in PatchDock, SP-Dock utilizes state-of-the art local shape descriptors, which makes the method more discriminative in terms of local complementarity matching. Then, in order to match multiple complementary pairs simultaneously and enhance the certainty of pairwise matches, the proposed SP-Dock method does not use geometric hashing (as in PatchDock) but it introduces a new grouping algorithm. This algorithm groups intuitively pairs of complementary GSPs and allows for slight flexibility in the relative positions of the corresponding GSPs within a group. The latter increases the robustness of the method and makes it more appropriate for unbound docking cases, where slight side-chain flexibilities are allowed. The superiority of the proposed method over PatchDock is demonstrated in the experiments section, where SP-Dock outperforms PatchDock especially in the unbound case.

4.3 Preprocessing

This section describes the preprocessing procedure, which involves two phases: during the first phase, an appropriate representation of the molecular surface is generated from the input PDB file, a set of critical points is extracted and a GSP is created for each critical point. The second phase involves the extraction of low-level geometric descriptors for each GSP, which uniquely characterize its shape.

4.3.1 Surface Representation and Extraction of Local Patches

Extraction of 3D shape descriptors from a protein initially requires an appropriate representation of its 3D structure. Several representations have been proposed so far, namely the volumetric representation [66], the Solvent Excluded Surface (SES) [9], Sparse Surface [85] and Alpha Shapes [94]. In this work, the SES method has been selected, which produces a 3D triangulated surface of the protein. In order to generate a SES, the Maximal Speed Molecular Surface (MSMS) [10] algorithm has been utilized.

Computation of critical points on the SES offers a sufficient approximation of the protein surface and constitutes a preliminary step that is followed by almost all the local shape feature matching approaches. We have followed a method for generating critical points based on the local curvature of the surface. This approach has been introduced in our previous work (Chapter 3), it is applied directly to the 3D triangulated mesh and it is applicable to all types of triangulated meshes. The extracted critical points are the centers of concave and convex regions of the molecular surface.

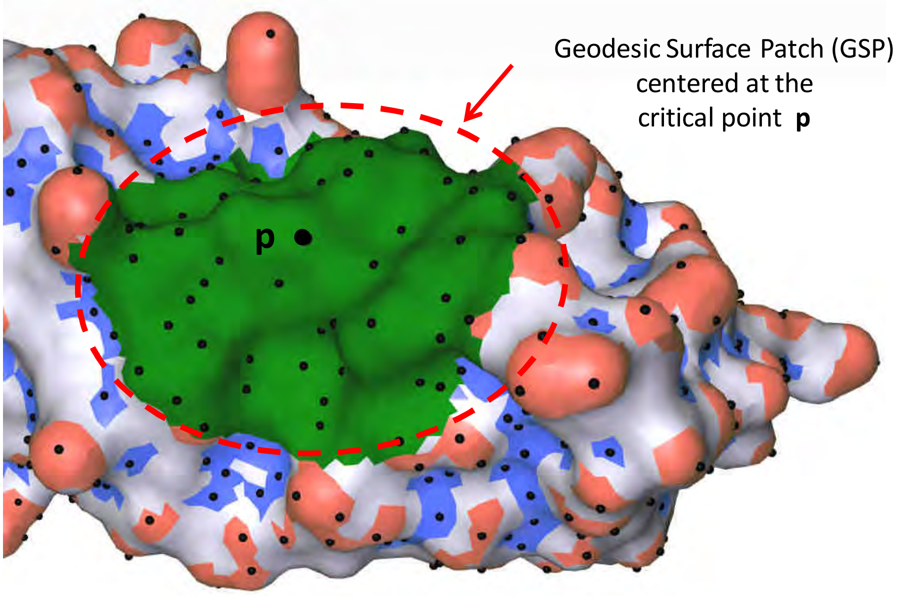


Figure 4.2: A Geodesic Surface Patch (GSP) is centered at the critical point **p**.

For each critical point, a GSP is created (Figure 4.2), which spreads over a wider surface area around that point. More specifically, a GSP consists of all points of SES whose geodesic distance from the critical point is less than a predefined threshold (G_{max}). GSP differs from the Extended Surface Patch (ESP) that was defined in our previous work in the sense that the latter uses the Euclidean distance as an initial threshold, while the geodesic distance is used only as a post-filter to remove unconnected surface parts. However, it was proven experimentally that the GSP-based approach achieves better accuracy than the ESP-based approach. It was also experimentally found that an optimal value for G_{max} is 16Å.

4.3.2 Local Descriptor Extraction

In protein-protein docking problems, local-shape-feature-based methods rely on pairwise matching of local surface regions between the receptor and the ligand. The most complementary surface regions are, then, selected as candidate poses. The approach presented in this chapter uses shape similarity descriptors to measure surface complementarity. It has been proven in the previous chapter that complementarity matching of surface patches can be reduced to a similarity matching problem, if the inner surface part of the ligand patches is treated as outer and vice versa.

In the approach presented in this chapter, the GSPs of the receptor that correspond to convex (or concave) critical points are matched with the GSPs of the ligand that correspond to concave (or convex) critical points. The matching relies on the shape complementarity between the GSPs. Unlike our previous method, where only the SID was used for shape similarity, in this work, three local shape descriptors have been tested in order to find the most appropriate one for our problem. The most well-known local shape descriptors for 3D meshes have been presented in SHREC 2011 (Shape Retrieval Contest on Non-rigid 3D Watertight Meshes) [92]. Two local shape descriptors that achieved high accuracy in SHREC 2011 have been tested in our docking framework and compared with the Shape Impact Descriptor. The selection of descriptors has been performed according to the following criteria:

Rotation Invariance: the local patches of the two protein surfaces have arbitrary orientations. In order to be matched, they should be either aligned or a rotation-invariant descriptor can be used. Alignment is usually based on the directions of the patch normals; however, the latter do not provide a robust measure, which leads in inaccurate alignment. Rotation-invariant descriptors are able to match two surface patches irrespective of their pose.

Compactness and fast extraction: local descriptors are applied to a relatively big number of surface patches. This implies that descriptor extraction and pairwise matching of single patches should be extremely fast. Fast matching is achieved by using very compact descriptor vectors (usually up to 100 values). Thus, shape descriptors with high computational complexity are not appropriate in our case.

Finally, the candidate shape descriptor *should be applied to the surface of the protein*, which automatically excludes descriptors based on the volume of a 3D object. The descriptors that will be described in the sequel, for the sake of completeness, fulfil all the above requirements.

Local Spectral Descriptor

This local descriptor has been proposed by G. Lavoue in [93] for retrieval of non-rigid 3D meshes. It is based on the extraction of geometric descriptors from a surface patch P_i centered around a sample point $\mathbf{p}_i$ on the mesh. The method computes the Fourier spectra of the patch by projecting the geometry on the eigenvectors of the Laplace-Beltrami operator (LBO). LBO is defined as the divergence of the gradient for functions that are defined over manifolds. The eigenvalues and eigenvectors of this operator satisfy the following equation:

$$-\mathbf{Q}\mathbf{h}^k = \lambda_k \mathbf{D}\mathbf{h}^k \tag{4.1}$$

where λ_k is the k^{th} eigenvalue, $\mathbf{h}^k$ is the k^{th} eigenvector $\mathbf{h}^k = [H_1^k, \ldots H_m^k]$ and m is the total number of vertices of the surface patch. $\mathbf{D}$ is the Lumped Mass matrix and $\mathbf{Q}$ is the Stiffness matrix that are described in [95]. In order to compute the k^{th} spectral coefficient, the inner product between the patch surface and the k^{th} eigenvector is calculated:

$$\tilde{x}_k = < \mathbf{x}, \mathbf{h}^k > = \sum_{i=1}^{m} x_i D_{i,i} H_i^k \tag{4.2}$$

where x_i is the x-coordinate of the i^{th} vertex of the surface patch. Similar equations hold for $\tilde{y}_k$ and $\tilde{z}_k$, which correspond to the y and z coordinates, respectively. Finally, the k^{th} spectral coefficient is given by:

$$c_k = \sqrt{(\tilde{x}_k)^2 + (\tilde{y}_k)^2 + (\tilde{z}_k)^2} \tag{4.3}$$

The Local Spectral Descriptor for patch P_i around point $\mathbf{p}_i$ is the vector $\mathbf{c}^i = [c_1^i, \ldots c_n^i]$, where $k = 1, \ldots, n$ the first spectral coefficients. The dimensionality of the descriptor has been experimentally found to be $n = 50$.

ShapeDNA

The ShapeDNA descriptor has been proposed by M. Reuter et al. in [96] for non-rigid shape analysis. It presents similarities with the previous approach of Local Spectral Descriptors in the sense that they are both based on solving the eigenvalue problem of the Laplace-Beltrami operator. However, in ShapeDNA the descriptors are the first smallest *N* eigenvalues, which are the solutions of the Laplacian eigenvalue problem (4.1), while in Local Spectral Descriptors, the descriptors are extracted by projecting the geometry of the surface on the eigenvectors of the Laplace-Beltrami operator.

The ShapeDNA descriptor is insensitive to noise, it can deal with objects containing cavities and it is isometry invariant. To compute the first eigenvalues of LBO, the simple linear Finite Element Modeling (FEM) [95] is utilized. In general, a small number of egenvalues (10 to 15) provide a sufficient number of descriptors (less than 10 is not discriminative enough, while including higher values increases influence of noise and non-isometric deformations). In our experiments, $N = 14$ was experimentally found to give the optimal results. To compute the ShapeDNA descriptor on the GSPs of the receptor and ligand proteins, a remeshing on the surface patches has been applied, since mesh quality can degrade the accuracy of the linearly approximated eigenvalues. A detailed description of the ShapeDNA descriptor is available in [96].

Shape Impact Descriptor (SID)

SID was firstly introduced in [86] and extended in [87] as a shape similarity measure for 3D objects. The key idea of SID is the description of the resulting phenomena that occur by the insertion of the 3D object in the space. It is expected that similar objects will result in similar physical phenomena. Some obvious selections of surrounding fields are the traditional electrostatic force field and the Newtonian force field. Any 3D object can be considered as a distributed mass (or a distributed charge) with a specific distribution, resulting in a static field around it.

SID is composed of three major histograms created by a) the field potential values $\phi(x)$, b) the field density Euclidean norms $E(x)$ and c) the radial component of the field density ($E(x)\cdot n_r(x)$), computed in points **x** that are equidistant from the object surface. The computation of histograms involves only relative distances, thus the descriptor is rotation-invariant. A more detailed description of SID is available in the previous chapter and in [86], [87].

4.4 Group-based Matching and Alignment

Most of the existing local shape feature docking approaches are based on either one-patch-to-one-patch or two-patch-to-two-patch complementarity matching between the local patches of the receptor and the ligand. Then, the most complementary pairs are aligned in order to produce the final poses as follows: a) for methods based on single-patch matching [34], the ligand is translated so that the patch center of the ligand patch coincides with the patch center of the receptor patch; b) for methods based on two-patch matching [33], the ligand is translated so that its first critical point coincides with one of the receptor and then it is rotated so that its second critical point coincides with the second critical point of the receptor. These approaches suffer from the following limitations:

Alignment is not always accurate: the patch centers (critical points) of the receptor and the ligand do not always coincide with their real contact points, producing docking poses that may be far from the near-native poses. An approach that is usually followed is to increase the number of samples on the protein surfaces, which in turn dramatically increases the computation time.

Low shape complementarity between surface patches: this is due to the fact that shape complementarity in protein surfaces is mostly approximate rather than exact. This results in relatively low complementarity scores of patches that correspond to near-native poses comparing to scores of patches that correspond to non-native poses, causing a high number of false positive predictions.

Instead of applying single-patch or two-patch matching, we propose a novel approach, where several neighboring GSPs are grouped together to generate candidate binding regions on the surfaces of the receptor and the ligand. This increases the confidence of a receptor patch to be complementary to a ligand patch, since, according to the grouping criterion, the neighbours of the receptor patch should be complementary to the neighbors of the ligand patch as well.

4.4.1 Creating Groups of Neighboring Complementary GSPs

The steps of the group-based matching algorithm are summarized in Figure 4.3. Let N_R, N_L be the GSPs of receptor and ligand, respectively, and D_R^i , D_L^j their corresponding local shape descriptors, where $i=1,\dots,N_R$ (or N_L). Let, also, the function that represents the convexity or concavity of a GSP be:

$$Cur(i)=\begin{cases}1, if\ i\ is\ a\ convex\ GSP\\ -1, if\ i\ is\ a\ concave\ GSP\end{cases} \tag{4.4}$$

```
INPUT:
  $N_R$, $N_L$ the GSPs of receptor and ligand, respectively
  $D_R^i$, $D_L^i$ their local shape descriptors, $i$=1,…$N_R$ (or $N_L$)
OUTPUT:
  $G=\{G_1,G_2,\dots,G_M\}$ the set of patch groups
ALGORITHM:
  Set $G\leftarrow\{\}$
  For each receptor GSP $i$
    For each ligand GSP $j$
      If $Cur(i)\cdot Cur(j)=-1$
        Calculate $dis(D_R^i,D_L^j)$
    Sort GSPs of ligand
    Keep $k$-first ligand GSPs and create ranked list $RL_R^i$
    For each ligand GSP $j$ of $RL_R^i$
      For each group $G_k\in G$
        If pair ($i,j$) fulfils Grouping Criterion for $G_k$
          Then add pair ($i,j$) to $G_k$
      If ($i,j$) not added to any group
        Then create new group and add to $G$
```

Figure 4.3: The Group-based Matching algorithm.

Each receptor GSP i is matched with all ligand GSPs j of different type ($Cur(i) \cdot Cur(j) = -1$). A dissimilarity metric is calculated for each pair (i, j) as:

$$Dissimilarity(i, j) = dis(D_R^i, D_L^j) \tag{4.5}$$

where $dis()$ is an appropriate distance metric applied on the descriptor vectors D_R^i and D_L^j. The distance metric depends on the selected descriptor. In our experiments, the Manhattan distance (L_1), the Euclidean distance (L_2) and the diffusion distance [89] have been selected for matching of the Local Spectral Descriptors, the ShapeDNA descriptors and the SID descriptors, respectively. After computation of the dissimilarities, the GSPs of ligand are sorted with respect to similarity to the receptor GSP i and the *k*-first are selected to form a ranked list RL_R^i. It is worth mentioning that similarity of the local descriptors is equivalent to complementarity.

The output of the algorithm is a set of groups G, which is defined as follows:

$$G = \{G_1, G_2, \ldots, G_M\} \tag{4.6}$$

where $G_k = \{(I_R^1, I_L^1), (I_R^2, I_L^2), \ldots, (I_R^g, I_L^g)\}$ is a group that consists of the pairs (I_R^i, I_L^i), I_R^i is the index of a receptor GSP ($I_R^i = 1, \ldots N_R$) and I_L^i is the index of a ligand GSP ($I_L^i = 1, \ldots N_L$). In order for the above pairs to form a group, the following grouping criterion must hold:

$$d_{Geod}(I_R^i, I_R^j) < gThres \qquad i, j \in [1, \ldots g] \text{ and}$$

$$d_{Geod}(I_L^i, I_L^j) < gThres \qquad i, j \in [1, \ldots g] \tag{4.7}$$

where d_{Geod} is the geodesic distance between two GSPs of either the receptor or the ligand and $gThres$ an appropriately selected geodesic threshold.

The candidate pairs of a group G_k are created by combining each receptor GSP i with the k most similar ligand GSPs, i.e. those included in the ranked list RL_R^i. Consequently, a group G_k consists of neighboring receptor and ligand GSPs and each receptor GSP of the group is complementary with at least one ligand GSP of the group. This is illustrated in Figure 4.4 (a) and (b). In Figure 4.4 (a), the group of the receptor consists of four patches, whose centers are represented by blue spheres. Three of these patches are complementary with one patch of the ligand group, while the second patch of the ligand group is complementary with the fourth patch of the receptor. In general, the proposed grouping algorithm allows many–to-many correspondences of local patches increasing the confidence of complementarity between pairs of groups.

The ranges of $gThres$ and k values have been experimentally determined to be at the ranges of 4-6 Å and 8-11, respectively. Higher values of $gThres$ result in a smaller number of larger groups, while lower values of $gThres$ result in a larger number of smaller groups. In the former case, the algorithm may fail to predict some near-native poses, while in the latter case, more false positive results may occur. Similar observations are made for k , if we decrease it ($k < 8$) or increase it ($k < 11$), respectively. For the experiments, *gThres*=5Å and *k*=10 lead to the best results.

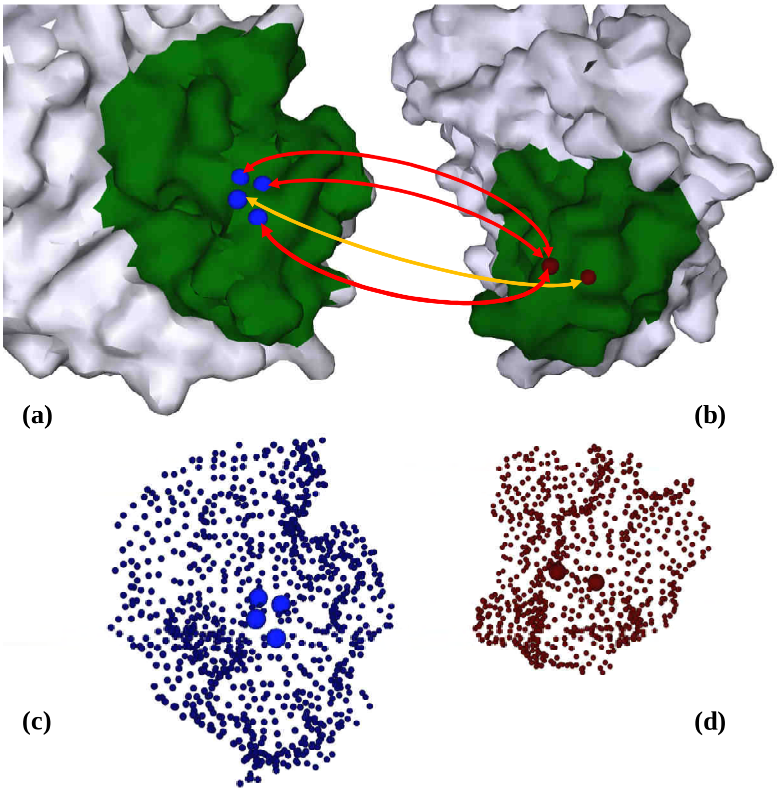


Figure 4.4: (a) and (b): a pair of complementary groups, one from receptor and one from ligand, respectively, for the 1AVX complex. The first patch of the ligand is complementary with one patch of the receptor and the second patch of the ligand is complementary with three patches of the receptor. (c) and (d) the corresponding point clouds that are given as input to ICP for the alignment step.

The above process produces M groups, which will be given as input to alignment and final scoring function and will result in M predicted docking poses. There is no need to define an additional cutoff threshold, as required in our previous work to select the first most complementary pairs of patches, since the average number of M is just 2500-3000. Another advantage of the proposed grouping algorithm, comparing with our previous method, is that patch pairs that lead to almost the same docking poses are not taken as separate cases but are grouped together (due to the neighborhood criterion). This results in a significantly smaller number of false positive predictions, which improves the final rank of the near-native predictions.

4.4.2 Iterative Closest Point (ICP) Alignment

During the alignment phase, a rigid transformation of the ligand is computed for each of the M groups created using the group-based matching algorithm. Let C_R^i (or C_L^i) be the point cloud that consists of all points of the i^{th} receptor GSP (or ligand GSP). The receptor point cloud GC_R^k of group G_k is given by:

$$GC_R^k = C_R^{I_R^1} \cup \ldots \cup C_R^{I_R^g} \tag{4.8}$$

i.e. it is the union of the receptor point clouds C_R^i of the GSPs within group G_k. The ligand point cloud GC_L^k of group G_k is computed in a similar manner. The required rigid transformation translates and rotates GC_L^k so as to optimally fit to GC_R^k. Then, the same rigid transformation is applied to the entire ligand molecule in order to compute the final score of the predicted pose.

The optimal alignment of two point clouds is a surface registration problem. One of the most well-known techniques for surface registration is the Iterative Closest Point (ICP) algorithm [97]. Let $GC_R = \{\mathbf{c}_R^1, \mathbf{c}_R^2, \ldots, \mathbf{c}_R^{n_R}\}$ and $GC_L = \{\mathbf{c}_L^1, \mathbf{c}_L^2, \ldots, \mathbf{c}_L^{n_L}\}$ be the two point clouds to be aligned, and $\left\|\mathbf{c}_L^j - \mathbf{c}_R^i\right\|$ be the Euclidean distance between point $\mathbf{c}_R^i \in GC_R$ and $\mathbf{c}_L^j \in GC_L$. Let also $CP(\mathbf{c}_L^j, GC_R)$ the closest point of GC_R to the point $\mathbf{c}_L^j$. It is useful to launch ICP with an initial estimate T^0 of the rigid transformation. This is usually computed by translating the median point of GC_L to coincide with the median point of GC_R and rotating GC_L so that its average normal (the average of the normals of all points $\mathbf{c}_L^j$) is aligned with the average normal of GC_R. Then, an iterative process is repeated ($t = 1,\ldots,t_{\max}$ iterations) until convergence. For the t^{th} iteration, the set of correspondences is computed by:

$$Corr^t = \bigcup_{i=1}^{n_L} \left\{ \left(\mathbf{c}_L^i, CP\left(T^{t-1}(\mathbf{c}_L^i), GC_R \right) \right) \right\} \tag{4.9}$$

Then, the new transformation T^t that minimizes the mean square error between point pairs in $Corr^t$ is computed. In Figure 4.4 (c) and (d), the point clouds that correspond to the pair of complementary groups (a) and (b) are depicted. These are given as input to ICP at the alignment phase. In Figure 4.5, two results of alignment using ICP are provided for the 1AVX and 1HIA complexes. It is obvious that a highly accurate alignment is achieved.

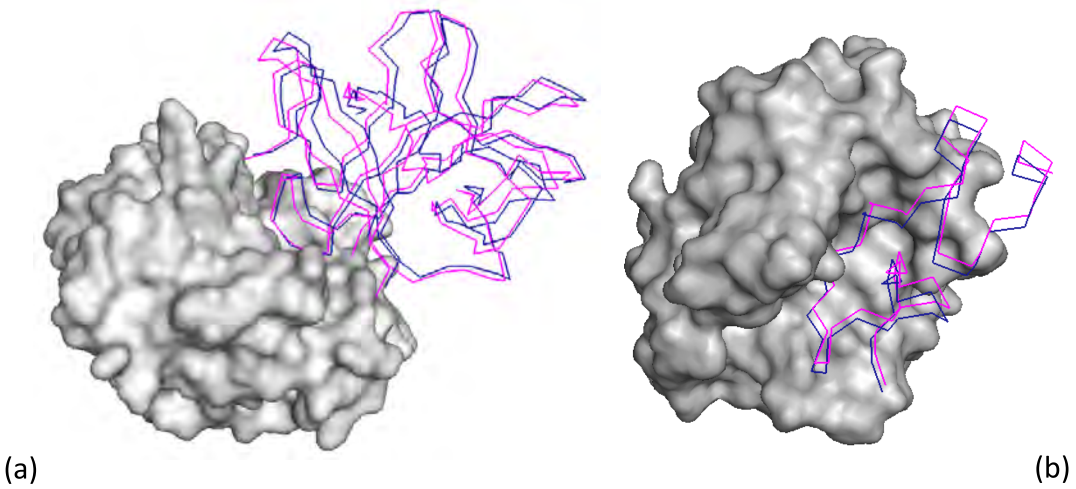


Figure 4.5: Aligment results using ICP for the (a) 1AVX and (b) 1HIA complexes. A surface representation is used for the receptor and a backbone representation for the ligand. The blue line corresponds to the original position of the ligand and the magenta line corresponds to the pose predicted using ICP.

4.5 Scoring of Candidate Poses

In this section, the final stage of the proposed SP-Dock method is described, which involves scoring of the candidate poses that were produced during the group-based matching and alignment phase. Apart from the geometric complementarity, the effect of several (non-geometric) physicochemical factors on the accuracy of docking predictions is also investigated.

The final scoring function is a weighted sum of the geometric score and the scores obtained from each separate physicochemical factor. The predicted docking poses are sorted in descending order, with the poses of the highest overall score to appear first in the ranked list.

4.5.1 Geometric Scoring

For the geometric scoring of each candidate pose, the 3D distance grid, which was presented in Chapter 3, is used. The receptor protein and its surrounding space is represented by a 3D function $DT(i,j,k)$:

$$DT(i,j,k)=\begin{cases}0, & \textit{if at least one surface point lies inside the voxel}\\ <0, & \textit{if the voxel lies inside the molecule}\\ >0, & \textit{if the voxel lies outside the molecule}\end{cases} \tag{4.10}$$

The absolute value of each voxel corresponds to the Euclidean distance from the closest surface point. Then, the distance grid is divided into 6 shells (Table 4.1) according to the distance from the molecular surface. The shell ranges have been experimentally determined. It is worth mentioning that the shell ranges are similar to the ones obtained by the PatchDock method [33] (with a 0.2 Å shift), which was expected since the geometric scoring step is quite similar in both methods.

Table 4.1: The shells in which the distance grid is divided.

Shell 1	[1.2, ∞)	The range (in Å) of the first shell of the distance grid
Shell 2	[-1.2, 1.2)	The range of the second shell of the distance grid
Shell 3	[-2.4, -1.2)	The range of the third shell of the distance grid
Shell 4	[-3.8, -2.4)	The range of the fourth shell of the distance grid
Shell 5	[-5.2, -3.8)	The range of the fifth shell of the distance grid
Shell 6	[−∞, -5.2)	The range of the sixth shell of the distance grid
$a_{1\text{-}6}$	0, 1, -1, -18, -190, -10000	The values of the weights in the scoring function (13)

For each of the docking poses predicted in the group-based matching and alignment phase, the translated and rotated ligand L enters the 3D distance grid of the receptor R . L's surface points access the voxels of the 3D grid and are assigned a value according to the distance from R's molecular surface. The score E_S of the transformation is given by:

$$E_S = \sum_{i=1}^{6} a_i N_i \tag{4.11}$$

where N_i is the number of L points in shell *i* of the distance grid and a_i is the weight of the *i*-th shell (Table 4.1).

After the ICP-based alignment step, *M* different poses of the ligand are taken (equal to the *M* generated groups *G*). For the pose that corresponds to a group G_k, an additional refinement step is applied, which involves +/-2 Å translation of the ligand towards the direction of GC_L^k's average normal and +/-25^O rotation of the ligand around GC_L^k's average normal. This results in a total of 9 poses for each group G_k, which is significantly faster than our previous method that requires 1872 different poses for each pair of complementary ESPs. The reason for taking only 9 poses is that the final transformation has been already approximated using ICP, thus, only a slight refinement is required. Taking also into account that ICP is significantly faster than distance-grid-based scoring, the significance of ICP in our approach is obvious.

The computation time required for the distance-grid-based scoring is proportional to the size and the resolution of the ligand's surface. In order to achieve low computation times, two different resolutions of the ligand SES are used: a) the low-resolution surface with point density of 1 point per Å^2 and b) the high-resolution surface with point density of 4 points per Å^2. The low-resolution surface is used to score all 9 poses for each group G_k, and the high-resolution surface is used for the best among the 9 poses.

4.5.2 Physicochemical Factors Assessment

Among the several non-geometric physicochemical factors that may affect the accuracy of protein-protein docking, the following have been assessed:

Atom Desolvation Energy (ADE): the atomic contact potential, which is used to estimate the desolvation energy for the replacement of protein-water contacts with protein-protein contacts, is given by [98]:

$$E_{ADE} = \sum_{i=1}^{N}\sum_{j=1}^{M} e_{ij} \tag{4.12}$$

where e_{ij} is the non-scaled contact value of a contact between atom *i* from receptor and atom *j* from ligand. The contact values are summed over all atoms of receptor that are within 6 Å distance to at least one atom of ligand and vice-versa.

Interface Residue Contact Preferences (RCP): these are volume-normalized pair probabilities that represent the pairing preferences of aminoacids at the protein-protein interface [99]:

$$E_{RCP} = \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{e_{ij}}{r_{ij} + 1.5} \tag{4.13}$$

where e_{ij} is the volume-normalised pairing preference between aminoacid *i* from the receptor and aminoacid *j* from ligand and r_{ij} is the distance between their corresponding C_{β} atoms. The value 1.5 has been added to avoid unrealistic close contacts.

Generic Residue Contact Preferences (GCP): it is calculated in a similar manner as in the case of Interface Residue Contact Preferences. In this case, e_{ij} is the pairing probability of aminoacids in protein structures [100].

Electrostatic Complementarity (EC): the electrostatic complementarity at the interface is calculated by [101]:

$$E_{EC} = \sum_{i=1}^{N}\sum_{j=1}^{M} e_{ij} \tag{4.14}$$

$$e_{ij} = \begin{cases} 0, \ \ if \ r_{ij} > r_{\max} \\ A(p_i, p_j), \ \ if \ r_{ij} \le r_{\max} \end{cases} \tag{4.15}$$

where $A(p_i, p_j)$ are statistical interaction energies, r_{ij} is the distance between atoms *i,j*, and $r_{\max} = 4$Å if both atoms are apolar and 3.4 Å otherwise.

Coulomb Potential (CP): Coulomb potential is given by the following equation:

$$E_{CP} = \frac{q_i q_j}{(r_{ij} + c)} \tag{4.16}$$

where q_i, q_j are the partial charges of each atom. The constant *c* is equal to 1.5 Å to avoid strong influence of very close atoms [102].

Hydrophobicity (HP): it is calculated using the following equation [103]:

$$E_{HP} = \frac{hh}{hh + pp + hp} \tag{4.17}$$

where *hh* is the number of contacts between hydrophobic atoms, *pp* is the number of contacts between two polar atoms and *hp* is the number of contacts between polar and hydrophobic atoms.

Van-der-Waals Potential (vdW): here, the modified 6-12 Lennard-Jones Potential is calculated by:

$$E_{vdW} = \begin{cases} \varepsilon_{ij}\left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - 2\frac{\sigma_{ij}^{6}}{r_{ij}^{6}}\right), & if\ r_{ij} > 0.6\sigma_{ij} \\ \varepsilon_{ij}(A + (r_{ij} - 0.6\sigma_{ij})B), & otherwise \end{cases} \tag{4.18}$$

$$A = \frac{\sigma_{ij}^{12}}{0.6\sigma_{ij}^{12}} - 2\frac{\sigma_{ij}^{6}}{0.6\sigma_{ij}^{6}}, B = -12\frac{\sigma_{ij}^{12}}{0.6\sigma_{ij}^{13}} + 12\frac{\sigma_{ij}^{6}}{0.6\sigma_{ij}^{7}} \tag{4.19}$$

where σ_{ij} is the sum of van-der-Waals radii and r_{ij} is the distance between atoms *i* and *j*. The potential is calculated for atoms with interatomic distances of less than 6 Å.

It should be stressed that the above factors are not the only ones that affect the protein interactions. A variety of additional physicochemical properties could be also found and integrated into a compound scoring function. An extensive survey on all possible factors is not within the scope of this work, but it constitutes a significant challenge for future research. The factors presented above are also summarized in [39], where it is stated that they are able to improve docking predictions when merged with geometric docking. However, no information about the contribution of each separate factor is given in [39]. In this work, an assessment of each factor is provided through an appropriate optimization method. More specifically, the overall score of each docking pose is given as the weighted sum of the geometric score (Section 4.1) and the scores of the factors described above. The weights are optimized on a training dataset (59 test cases of Docking Benchmark v1.0 [91]) using Particle Swarm Optimization (PSO) [104]. The overall scoring function is given by:

$$Score_{Total} = w_s E_s + w_{ADE} E_{ADE} + w_{RCP} E_{RCP} + w_{GCP} E_{GCP} + w_{EC} E_{EC} + w_{HP} E_{HP} + w_{CP} E_{CP} + w_{vdW} E_{vdW} \quad (4.20)$$

PSO is a global optimization algorithm, similar to a genetic algorithm, motivated by social behavior of organisms such as bird flocking and fish schooling. PSO iteratively tries to improve a candidate solution with respect to a given measure of quality (fitness function). PSO establishes a population (swarm) of candidate solutions, known as particles that move around in the search space, and are guided by the best found positions, updated when better positions are found by the particles.

In our approach, the population of candidate solutions is the 8 weights *w* of (4.20), which can take arbitrary real values within the range $[0,1]$. The values of scores *E* (4.20) have been normalized so that their value range is within $[0,100]$. The key of success of the PSO method is the selection of an appropriate fitness function. In our experiments, two fitness functions are determined. The first one is the *Average Precision* of the first-ranked *hit* for all the complexes of the training dataset. The *Precision* of the first-ranked hit for one complex is given by:

$$F_1 = \frac{n_{hit}}{n_{retrieved}} = \frac{1}{n_{retrieved}} \quad (4.21)$$

where n_{hit} is the total number of hits (i.e. near-native poses) that are retrieved and $n_{retrieved}$ is the total number of predicted docking poses that are retrieved. If we select $n_{retrieved}$ to be equal to the number of retrieved poses until the first hit is retrieved, then the numerator of (4.21) is equal to 1. As an example, if the first hit is retrieved in the fourth position, then the Precision for this complex is $F_1 = 0.25$, or 25%. The Average Precision of the first-ranked hit provides an acceptable metric to be used as a fitness function, however, it suffers from the following limitation: it favors those complexes in which the first ranked hit is retrieved at the first positions (1 - 10), while the complexes, in which the first hit has $rank > 100$, have insignificant contribution to the calculation of the Average Precision. In other words, an improvement of the hit's rank from 200 to 100 contributes with 0.01 to the average precision, while an improvement from 2 to 1 contributes with 1 to average precision. This is not desired since in the former case the improvement is much more significant and should contribute more to the average precision.

To overcome the above limitation a new fitness function was determined, which is given by:

$$F_2 = \frac{1}{N_C} \sum_{i=1}^{N_C} \frac{rank_S^i - rank_{SP}^i}{N_{Poses}^i} \tag{4.22}$$

where N_C is the number of complexes of the training dataset, $rank_S^i$ is the rank of the first hit (of complex *i*) that is retrieved using only shape complementarity, $rank_{SP}^i$ is the rank of the first hit using the weighted score (4.20) and N_{Poses}^i is the number of predicted poses of complex *i*.

4.6 Experimental Results

The proposed (Shape-Physicochemical) SP-Dock method was experimentally evaluated using the protein-protein docking benchmark v2.4 [90], which consists of 84 known complexes (63 rigid-body cases, 13 cases of medium difficulty, and 8 difficult cases). To evaluate the performance of the method, for each complex, the receptor and ligand are separated from each other and the ligand is translated and rotated arbitrarily. In order to increase the confidence of the results, the docking algorithm has been repeated for three different initial rotations of the ligand. Eventually, we observed that these three arbitrary rotations produced only very slight modifications on the final poses (mainly due to the outcome of the ICP algorithm), which did not affect the final rankings. Thus, the result of only one of the three iterations (the first one) is presented in the following subsections. The docking algorithm described in the previous sections is applied to generate a set of candidate poses of the ligand. The predicted pose of the ligand is compared to its original pose in the complex in terms of interface Root Mean Square Deviation (*iRMSD*):

$$iRMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{a}_i^p - \mathbf{a}_i^o\right\|^2} \tag{4.23}$$

where $\mathbf{a}_i^p$ is the i^{th} interface C_a atom (in x, y, z coordinates) of the ligand in the predicted pose and $\mathbf{a}_i^o$ is the corresponding C_a atom of the ligand in the original pose (crystallized complex). Interface C_a atoms of the ligand are those that are within the distance of 10 Å from the receptor. A predicted pose is called a hit if the iRMSD between the ligand in that pose and the ligand in the original complex is less than 2.5 Å.

4.6.1 Evaluation of Local Descriptors and Physicochemical Factors

The choice of the appropriate local shape descriptor is crucial for the accuracy of the docking predictions. In Figure 4.6(a), a comparison of the three shape descriptors of Section 4.3.1 is given. The diagram depicts the distribution of the ranks of the first prediction within 2.5 Å of the native complex structure. As an example, in the case of the Local Spectral Descriptor (blue column), the value of the first bin is 17, which means that in 17 out of 84 complexes of Docking Benchmark 2.4 the algorithm returned a hit at the first position. Similarly, the value of second bin is the number of complexes where a hit is predictred within the first five positions and so on. The value of the last bin is 76, i.e. in 76 complexes the algorithm returned a hit within the first 3600 positions, thus failed only in 8 cases. It is also clear from Figure 4.6(a) that the Local Spectral Descriptor produces better results than ShapeDNA and SID, thus it was eventually selected for our SP-Dock method.

In Figure 4.6(b), the effect of using physicochemical properties along with shape complementarity is demostrated. The red and green columns depict the ranks distribution in Docking Benchmark 2.4 using the unified scoring function of (20) optimized with the fitness functions F_1 (21) and F_2 (22), respectively. In both cases, the use of physicochemical properties improves the docking predictions of the shape-only approach. However, the weighted function optimized with F_1 demonstrates better improvement at the first ranks (1-10), while the weighted function optimized with F_2 is better at the higher ranks (100-2000). This makes sense taking into account the fact that F_1 favors those complexes in which the first ranked hit is retrieved at the first positions. Eventually, the results obtained with the fitness function F_2 were selected since they provide better overall improvement over the shape-only approach (65.3% improvement comparing with 57.3% obtained with F_1).

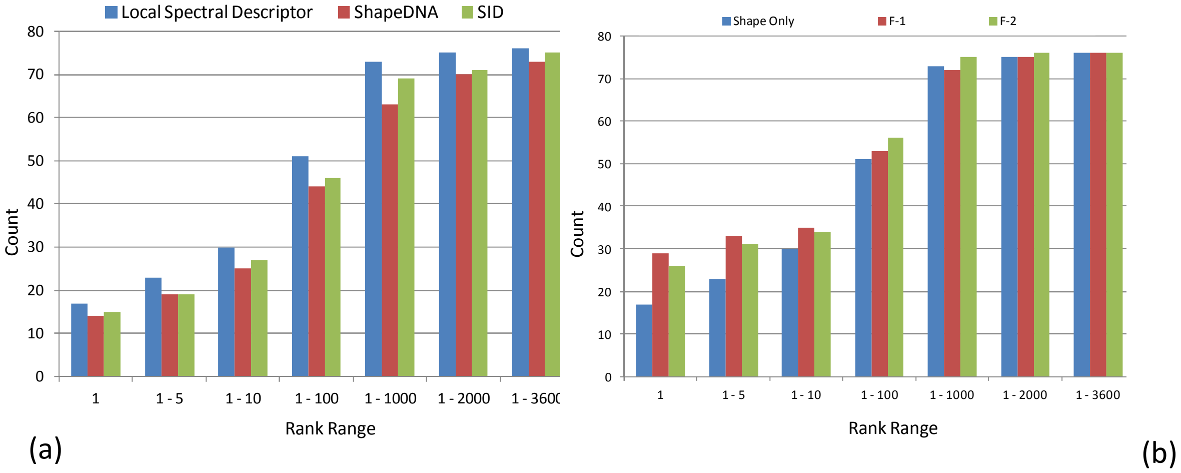


Figure 4.6: Distribution of the ranks of the first prediction within 2.5 Å of the native complex structure for all test cases in Docking Benchmark 2.4, a) for different local shape descriptors and b) comparison of our method using only shape complementarity with our method using shape and physicochemical complementarity and different fitness functions for weight optimization of the scoring function (F-1 is the precision of the first-ranked hit and F-2 is the function described in (22)).

The weights of (20) that produced the results presented in Figure 4.6 (b) have been optimized by training on a dataset (59 complexes) of the docking benchmark v1.0 [91]. These weights for both F_1 and F_2 fitness functions are depicted in Table 4.2.

Table 4.2: The optimized weights for each factor in (20) obtained by the two fitness functions and Particle Swarm Optimization.

Fitness Function	w_S	w_{ADE}	w_{RCP}	w_{GCP}	w_{EC}	w_{HP}	w_{CP}	w_{vdW}
F_1	0.114	0.158	0.008	0.006	0.143	0.01	0.05	0.51

F_2	0.231	0.013	0.007	0.005	0.104	0.079	0.047	0.512

4.6.2 Comparison with PatchDock, ZDock, LZerD, shDock and F²Dock

The results of the proposed method were compared to those of the following five methods: a) Local 3D Zernike descriptor-based Docking (LZerD) [36], b) Surface Histograms (shDock) [37], c) Fast Fourier Protein-Protein Docking (F^2Dock) [35], d) PatchDock [33] and e) ZDock [26]. These are the most recent works related to geometric protein-protein docking and they have achieved the best docking accuracy reported so far. In our experiments, both *R-bound/L-bound* and *R-unbound/L-unbound* cases were evaluated. It is worth mentioning that the last two methods, PatchDock and ZDock, have participated in the CAPRI experiment, a well-established arena for testing docking algorithms.

Two variations of the proposed method have been tested: the first (S-Dock) is based only on geometric properties, while in the second (SP-Dock), both shape and physicochemical properties are integrated. In Table 4.3, the performance of the proposed method in the *R-bound/L-bound* case, compared with the other three methods, is depicted. In "*Rank*" column, the rank of the best ranked hit is presented. This is not necessarily the hit with the smallest iRMSD, it is the first result of the rank list that produces iRMSD < 2.5Å. In "*RMSD*" column, the iRMSD of the best ranked hit is given. When these values are missing, the method failed to return a hit within the first 3600 predictions. In 3 cases none of the methods returned a hit in the first 3600 predictions, thus, they are not presented in Table 4.3. Moreover, two variations of F^2Dock are presented, either with (S-E) or without (S) the use of electrostatics.

Table 4.3: R-bound/L-bound: Comparisons between S-Dock, SP-Dock, LZerD, shDock and F^2Dock on 84 test cases from Benchmark v2.4. PDB gives the PDB id for the protein complex. RMSD and Rank give the iRMSD and rank of the best ranked hit (2.5 Å cut-off). In 3 cases none of the four methods returned a hit in the first 3600 predictions.

	PatchDock		ZDock		LzerD		shDock		F^2Dock (S)		F^2Dock (S-E)		S-Dock		SP-Dock	
PDB	Rank	RMSD	Rank	RMSD	Rank	RMSD	Rank	RMSD	Rank	RMSD	Rank	RMSD	Rank	RMSD	Rank	RMSD
Enzyme–Inhibitor or Enzyme–Substrate																

1ACB	10	1.6	6	0.82	15	1.5	140	1.31	**1**	**0.45**	**1**	**0.45**	142	1.42	20	1.37
1AVX	43	2.14	1	1.96	812	1.8	**1**	**0.73**	46	0.64	10	0.64	11	0.91	**1**	**0.91**
1AY7	24	2.07	46	1.68	46	1.31	11	0.56	1867	0.55	941	0.55	**7**	**2.02**	**7**	**2.14**
1BVN	**1**	**0.75**	10	1.24	**1**	**0.71**	**1**	**0.52**	3	0.98	44	0.98	**1**	**1.14**	**1**	**1.14**
1CGI	**1**	**1.08**	**1**	**1.12**	**1**	**0.61**	**1**	**0.41**	**1**	**0.4**	**1**	**0.4**	**1**	**0.63**	**1**	**0.63**
1D6R	-	-	35	1.04	27	0.52	13	1.48	40	0.35	41	0.35	5	2.01	**3**	**2.12**
1DFJ	-	-	-	-	-	-	-	-	**1**	**0.61**	**1**	**0.61**	44	1.27	6	1.27
1E6E	**2**	**2.29**	58	2.06	11	1.45	23	0.56	34	1.18	3	1.02	22	2.14	16	2.14
1EAW	**1**	**0.99**	**1**	**1.75**	2	0.81	**1**	**0.75**	59	1.14	10	1.14	**1**	**2.27**	**1**	**2.27**
1EWY	139	1.42	-	-	141	0.86	-	-	779	0.73	447	0.62	69	2.03	16	2.03
1EZU	**1**	**0.94**	-	-	**1**	**1.18**	3	0.35	24	1.09	9	1.09	9	0.97	**1**	**0.97**
1F34	**1**	**1.9**	**-**	**-**	**1**	**1.32**	**1**	**0.79**	**1**	**1.35**	**1**	**1.35**	**1**	**2.25**	**1**	**2.25**
1HIA	14	1.19	**1**	**1.7**	3	1.5	3	0.71	**1**	**0.52**	**1**	**0.52**	7	2.23	**1**	**2.15**
1KKL	-	-	-	-	946	0.91	26	0.78	1097	1.38	297	1.38	**22**	**1.03**	103	1.12
1MAH	**1**	**1.27**	**1**	**1.91**	**1**	**0.98**	**1**	**0.64**	-	-	-	-	**1**	**1.49**	**1**	**1.49**
1PPE	**1**	**1.03**	2	1.21	**1**	**0.41**	**1**	**0.43**	**1**	**0.77**	**1**	**0.77**	**1**	**1.32**	**1**	**1.32**
1TMQ	**1**	**1.52**	8	1.79	**1**	**0.65**	4	0.69	302	1.06	254	1.08	4	2.01	**1**	**0.65**
1UDI	**1**	**1.97**	**1**	**1.5**	**1**	**0.8**	**1**	**0.27**	324	1.15	18	0.94	**1**	**2.13**	**1**	**2.13**
2MTA	115	1.71	1447	2.26	92	0.88	81	0.66	269	1.58	305	1.41	56	2.17	**27**	**2.03**
2PCC	-	-	-	-	-	-	788	1.89	503	1.36	**16**	**0.6**	50	1.71	314	1.06
7CEI	-	-	5	2.18	54	1.08	3	0.86	162	0.34	58	0.34	26	1.77	**2**	**1.56**
Antibody–Antigen																
1AHW	40	1.55	56	1.18	1	1.21	43	1.06	-	-	-	-	-	-	-	-
1BGX	-	-	-	-	-	-	-	-	35	1.4	44	1.4	107	2.21	**34**	**2.21**
1BVK	131	2.12	1087	1.43	-	-	259	0.45	1831	0.66	310	0.41	78	2.03	206	1.93
1DQJ	83	1.71	19	2	391	0.67	**1**	**0.36**	3336	2.23	-	-	346	1.18	63	1.18
1E6J	1706	1.43	699	2.02	104	0.62	**46**	**0.52**	-	-	-	-	-	-	-	-
1JPS	96	1.87	23	2.3	21	0.98	**2**	**0.65**	1414	1.51	666	0.85	297	2.24	148	2.24
1K4C	337	1.53	30	1.16	274	0.88	**5**	**0.26**	-	-	-	-	148	2.07	34	2.07
1MLC	516	1.79	1205	1.37	1378	2.49	**10**	**0.67**	-	-	-	-	1274	1.73	252	2.16
1VFB	**-**	**-**	**-**	**-**	**37**	**1.74**	1369	0.27	349	0.59	159	0.59	92	1.33	59	1.14
1WEJ	**-**	**-**	1120	1.11	707	2.4	202	0.55	2266	1.36	2778	1.36	186	2.03	**155**	**2.09**
2VIS	-	-	-	-	-	-	3471	0.59	-	-	-	-	2225	2.15	**454**	**2.15**
Antigen–Bound Antibody																
1BJ1	-	-	**3**	**1.42**	58	0.73	4	0.55	-	-	-	-	180	2.02	57	2.02
1FSK	218	1.57	19	1.7	454	0.94	**1**	**0.52**	1030	1.89	994	1.89	385	2.38	719	2.38
1I9R	**-**	**-**	1271	2.04	**174**	**2.11**	3319	0.6	2794	1.69	1189	1.69	748	2.11	271	2.11
1IQD	**-**	**-**	55	1.83	**1**	**0.35**	28	0.37	772	0.99	81	0.99	334	2.19	200	2.19
1KXQ	29	1.63	30	1.6	47	0.83	4	0.21	1511	1.7	569	1.69	4	1.97	**2**	**1.71**
1NCA	-	-	20	1.48	354	1.21	**8**	**0.46**	-	-	-	-	655	2.21	658	2.21
1NSN	-	-	-	-	**36**	**0.82**	3305	2.2	-	-	-	-	701	2.24	1451	2.24
1QFW	-	-	16	2.46	**14**	**2.1**	1036	0.58	433	0.89	147	0.89	289	0.97	221	0.97

2JEL	282	1.65	532	1.77	**1**	**0.79**	10	0.31	3029	1.05	3124	0.86	297	2.09	161	2.14
Other																
1A2K	300	1.47	570	2.41	826	1.2	**1**	**0.7**	232	0.6	50	0.6	22	1.89	47	1.74
1AK4	-	-	3471	1.14	1024	2.18	-	-	13	0.34	**5**	**0.34**	-	-	-	-
1AKJ	-	-	448	1.88	-	-	7	0.68	32	0.93	12	0.93	**3**	**2.13**	37	2.13
1ATN	-	-	558	1.15	-	-	**11**	**0.43**	-	-	-	-	134	2.27	41	2.27
1B6C	40	1.92	24	1.69	44	0.94	88	0.96	911	0.94	1588	0.94	35	1.03	**1**	**1.13**
1BUH	83	1.14	393	1.43	2378	0.82	130	0.4	8	0.33	2	0.26	5	0.46	**1**	**0.98**
1DE4	-	-	452	1.62	-	-	**4**	**0.32**	51	1.36	38	1.36	32	1.19	**4**	**1.19**
1E96	1767	1.44	-	-	-	-	**87**	**0.68**	946	1.26	1053	1.26	462	2.17	104	2.17
1EER	**1**	**1.66**	-	-	38	1.4	**1**	**0.4**	-	-	531	1.55	117	2.05	94	2.05
1F51	**1**	**1.92**	-	-	26	1.58	26	1.51	642	2.21	782	2.21	2	2.10	4	1.34
1FAK	-	-	-	-	-	-	**1**	**0.59**	974	1.89	818	1.89	38	1.05	30	0.89
1FC2	49	1.24	55	2.18	42	0.52	**3**	**0.94**	2530	0.49	-	-	36	2.11	26	1.63
1FQ1	-	-	-	-	-	-	**2**	**0.7**	187	0.73	-	-	1076	1.15	189	1.15
1FQJ	248	1.48	120	1.94	15	1.78	**1**	**0.85**	-	-	-	-	12	2.23	119	2.23
1GCQ	-	-	382	1.81	119	0.44	**1**	**0.44**	11	0.4	328	0.43	**1**	**1.55**	**1**	**0.82**
1GP2	-	-	-	-	-	-	**1**	**0.34**	2224	1.85	1277	1.42	**1**	**2.17**	**1**	**2.17**
1GRN	3	1.45	7	2.26	**1**	**1.08**	**1**	**0.22**	329	1.21	39	1.21	**1**	**2.07**	**1**	**2.11**
1H1V	-	-	1510	2.4	-	-	818	1.19	-	-	-	-	194	1.3	**46**	**1.3**
1HE1	**1**	**1.06**	7	1.67	15	0.81	**1**	**0.3**	3	0.59	**1**	**0.59**	**1**	**0.74**	**1**	**0.74**
1HE8	-	-	-	-	-	-	-	-	-	-	-	-	**847**	**2.27**	**847**	**2.27**
1I2M	**-**	**-**	14	1.8	**1**	**1.48**	95	0.92	433	0.99	2	0.98	9	2.14	26	2.14
1I4D	167	1.05	793	2.08	695	1.68	**38**	**0.57**	-	-	-	-	-	-	-	-
1IB1	-	-	-	-	5	2.04	**1**	**0.57**	181	0.91	56	0.91	**1**	**1.26**	**1**	**1.26**
1IBR	-	-	-	-	336	1.34	**1**	**0.79**	398	1.87	166	1.74	**1**	**2.27**	**1**	**2.27**
1IJK	-	-	-	-	115	0.9	2958	1.04	277	1	-	-	164	2.11	**86**	**2.11**
1K5D	-	-	10	2.11	2245	2.03	**1**	**0.75**	1370	0.83	42	0.69	**1**	**1.57**	**1**	**1.57**
1KAC	-	-	381	1.52	-	-	534	0.52	1018	0.55	341	0.55	38	0.86	**8**	**0.7**
1KLU	-	-	-	-	-	-	178	0.54	424	1.13	1558	1.13	**24**	**2.09**	28	2.09
1KTZ	-	-	-	-	-	-	603	0.51	2965	0.8	190	0.61	589	2.15	332	2.15
1KXP	-	-	-	-	-	-	**1**	**0.97**	203	0.98	54	0.98	**1**	**1.25**	**1**	**1.25**
1M10	-	-	33	2.23	-	-	33	0.63	197	0.93	**11**	**0.84**	86	2.18	26	2.18
1ML0	**7**	**0.58**	75	1.94	157	1.48	52	0.41	-	-	-	-	42	1.14	35	1.14
1N2C	-	-	-	-	-	-	2	0.52	-	-	-	-	8	2.03	**1**	**2.03**
1QA9	-	-	-	-	-	-	-	-	77	1.25	**22**	**0.84**	28	1.24	86	1.05
1RLB	3143	2.32	-	-	561	0.49	2021	2.22	-	-	-	-	517	2.27	**226**	**2.27**
1WQ1	**1**	**0.84**	4	2.04	**1**	**1.77**	-	-	10	0.49	2	0.49	**1**	**1.38**	**1**	**1.38**
2BTF	137	1.82	21	1.21	40	0.94	4	0.9	-	-	-	-	4	0.71	**1**	**0.71**
2QFW	-	-	54	1.84	-	-	986	0.61	1106	0.91	**364**	**0.91**	-	-	-	-
2SIC	-	-	9	1.19	**1**	**0.88**	**1**	**0.68**	**1**	**0.64**	7	0.64	**1**	**2.14**	**1**	**2.14**
2SNI	13	2.1	4	2.5	4	1.27	2	0.51	**1**	**0.81**	**1**	**0.81**	2	0.88	**1**	**0.73**

Summing up the results of Table 4.3, the proposed approach failed to return a hit in 8 out of 84 cases, while shDock failed in 10 cases, LZerD in 25, F^{2}Dock (S) in 22, F^{2}Dock (S-E) in 25, PatchDock in 39 and ZDock in 23 cases. SP-Dock was ranked first in 40 out of 84 cases, far beyond the shDock method that was ranked first in 32 cases. In Table 4.5, the number of cases, where at least one hit is found at different docking thresholds, is presented for all methods. In the R-bound/L-bound results, the proposed SP-Dock method outperforms all other five methods for thresholds 1, 1000, 2000 and 3600, while for thresholds 5, 10, 100 only shDock outperforms the proposed method. In Table 4.4, the win-tie-loss-failure records for the proposed method versus shDock, LZerD, F^{2}Dock, PatchDock and ZDock is presented. Comparing our shape-only approach (S-Dock) with shDock, S-Dock returns a better ranked hit in 31 cases, whereas shDock returns a better hit in 29 cases. The methods tie in 20 cases, and both fail in 4 cases. Comparing against LZerD, F^{2}Dock (S), PatchDock and ZDock, the proposed method clearly outperforms them; it has 50-21 win-loss record against LZerD, 66-8 win-loss record against F^{2}Dock (S), 54-17 win-loss record against PatchDock and 56-21 win-loss record against ZDock (S). The accuracy of our method is further improved when shape complementarity is merged with physicochemical complementarity (Table 4.4). Note that S-Dock is compared to the shape-only version of F^{2}Dock, while SP-Dock is compared to the F^{2}Dock (S-E), where electrostatics are merged with the geometric properties.

Table 4.4: R-bound/L-bound: the win-tie-loss-failure records for the proposed method versus shDock, LZerD and F^{2}Dock.

S-Dock vs	Win	Tie	Loss	Both fail
shDock	31	20	29	4
LZerD	50	9	21	4
F^{2}Dock (S)	66	4	8	6
PatchDock	54	9	17	4
ZDock	56	4	21	3
SP-Dock vs	Win	Tie	Loss	Both fail
shDock	39	18	23	4
LZerD	51	11	18	4
F^{2}Dock (S-E)	59	6	13	6

PatchDock	58	11	11	4
ZDock	60	6	15	3

The above experiments have been performed using the bound molecules of both the receptor and the ligand. In Table 4.5, the number of cases, where at least one hit is found at different docking thresholds, is presented also for the unbound case. In the R-unbound/L-unbound results, the proposed SP-Dock method is ranked first for all docking thresholds.

Table 4.5: Number of test cases where at least one hit is found for different thresholds (1, 5, 10, 100, 1000, 2000 and 3600) and the average iRMSD, for both R-bound/L-bound and R-unbound/-unbound cases.

	PatchDock	ZDock	shDock	LZerD	F^{2}Dock	F^{2}Dock	S-	SP-
R-bound/L-bound								
Rank = 1	13	6	23	16	8	8	17	**26**
Rank ≤ 5	15	11	**37**	20	10	13	25	31
Rank ≤ 10	17	18	**41**	20	12	17	30	34
Rank ≤ 100	28	37	**57**	39	23	33	51	56
Rank ≤ 1000	39	48	67	56	46	52	73	**75**
Rank ≤ 2000	41	54	69	58	55	57	75	**76**
Rank ≤ 3600	42	55	74	60	62	59	76	**76**
Avg, iRMSD (Å)	1.53	1.73	0.69	1.17	1.01	0.95	1.73	1.68
R-unbound/L-unbound								
Rank = 1	0	**2**	0	1	1	1	**2**	**2**
Rank ≤ 5	1	4	1	2	2	2	3	**6**
Rank ≤ 10	1	5	2	2	2	2	9	**11**
Rank ≤ 100	9	11	6	14	9	11	23	**30**
Rank ≤ 1000	23	30	22	29	24	27	**53**	**53**
Rank ≤ 2000	31	35	33	36	31	33	55	**56**
Rank ≤ 3600	37	42	41	38	33	37	**56**	**56**
Avg, iRMSD (Å)	1.76	1.84	1.89	1.87	1.57	1.59	1.88	1.84

In Table 4.5, the average iRMSD is also presented for both the bound and the unbound cases. The average iRMSD is calculated as follows: for each case that succeeds in finding a hit in the top 3600 predictions, the iRMSD of the best ranked hit is taken. For all these cases, the average iRMSD is computed. It is worth mentioning that the average iRMSD of the proposed method is greater (i.e. less accurate) than the iRMSD of the other methods in the bound case, while it is comparable in the unbound case. This can be explained by the fact that the proposed method provides an approximate estimation of the docking pose, while other methods provide more exact estimations. However, the approximate complementarity matching of SP-Dock allows identification of complementary pairs of patches even after small conformational changes (unbound docking). This is the reason why the proposed method performs better in unbound docking than other methods, while at the same time its iRMSD is not significantly affected (as it happens with the other methods).

Table 4.6: R-unbound/L-unbound: the win-tie-loss-failure records for the proposed method versus shDock, LZerD and F^{2}Dock.

S-Dock vs	Win	Tie	Loss	Both fail
shDock	47	0	18	19
LZerD	46	1	19	18
F^{2}Dock (S)	51	0	13	20
PatchDock	47	0	15	7
ZDock	41	1	22	5
SP-Dock vs	Win	Tie	Loss	Both fail
shDock	49	0	16	19
LZerD	50	1	15	18
F^{2}Dock (S-E)	47	0	17	20
PatchDock	52	0	11	6
ZDock	45	1	19	4

Similar conclusions can be drawn in the win-tie-loss-failure records (Table 4.6). The proposed approach clearly outperforms all five methods, even in the case when only shape complementarity is used (S-Dock). If, instead of the geometric-only scoring, the shape-physicochemical scoring of (20) is used, the hit ranks are improved in 60.4% of the cases of Benchmark 2.4. The performance of all five methods for the unbound case in Benchmark 2.4 is shown in Table 4.7. It should be stressed that the proposed SP-Dock method does not return a fixed number of docking poses. The number of docking poses corresponds to the number *M* of patch groups (Section 4.4.2) and it varies depending on the size of the interacting proteins. In case *M*>3600, only the first 3600 ranked poses are kept and presented in Table 4.7. In complexes where *M*<3600, all poses fulfill the constraint of "first 3600 predictions", thus, all hits can be included in Table 4.7.

Table 4.7: R-unbound/L-unbound: Comparisons between S-Dock, SP-Dock, LZerD, shDock and F^{2}Dock on 84 test cases from Benchmark v2.4. PDB gives the PDB id for the protein complex. RMSD and Rank give the iRMSD and rank of the best ranked hit (2.5 Å cut-off). In 15 cases none of the four methods returned a hit in the first 3600 predictions.

	PatchDock		ZDock		LzerD		shDock		F^{2}Dock (S)		F^{2}Dock (S-E)		S-Dock		SP-Dock	
PDB	Rank	RMSD	Rank	RMSD	Rank	RMS	Rank	RMS	Rank	RMS	Rank	RMS	Rank	RMS	Rank	RMS
Enzyme–Inhibitor or Enzyme–Substrate																

1ACB	–	–	–	–	–	–	–	–	–	–	–	–	26	2.24	**10**	**2.17**
1AVX	2053	2.22	2863	2.23	786	2.41	1199	2.5	1769	1.75	1909	1.75	122	2.21	**97**	**2.21**
1AY7	679	1.14	–	–	1884	1.98	733	1.56	94	0.87	**32**	**0.98**	37	2.15	83	2.15
1BVN	110	1.68	502	1.97	27	2.32	82	2.15	72	1.58	54	1.58	6	1.65	**5**	**1.65**
1CGI	–	–	145	2.44	–	–	–	–	39	2.5	45	2.5	**7**	**2.12**	21	2.12
1D6R	–	–	2951	2.03	2619	2.24	–	–	177	1.45	**170**	**1.45**	634	2.19	449	2.19
1DFJ	–	–	**9**	**2.27**	–	–	–	–	243	1.15	22	1.14	–	–	–	–
1E6E	38	2.12	–	–	52	2.13	1014	1.52	–	–	3526	2.41	728	1.72	**7**	**1.34**
1EAW	59	2.1	**3**	**1.54**	20	2.42	324	2.07	517	1.7	454	1.52	9	1.14	18	1.14
1EWY	88	2.46	259	2.32	349	2.36	175	2.15	**4**	**1.21**	**4**	**1.17**	76	2.12	15	2.12
1EZU	–	–	1100	1.94	824	1.21	**784**	**2.24**	–	–	–	–	–	–	–	–
1F34	30	1.57	5	2.2	–	–	1528	2.14	98	1.34	60	1.34	**1**	**0.72**	**1**	**0.72**
1HIA	–	–	–	–	–	–	–	–	–	–	–	–	**49**	**1.93**	336	1.93
1KKL	–	–	–	–	–	–	–	–	–	–	–	–	**4**	**2.13**	8	2.13
1MAH	1184	0.83	**92**	**1.31**	**92**	**0.87**	2252	2.13	–	–	3327	2.07	1614	1.94	155	1.34
1PPE	12	1.51	**1**	**0.57**	**1**	**0.83**	8	1.86	355	1.12	392	1.12	**1**	**0.62**	**1**	**0.62**
1TMQ	**3**	**1.16**	314	1.88	50	1.45	186	1.18	247	1.63	241	1.63	225	1.56	5	1.56
1UDI	261	1.55	258	2.17	59	2.36	–	–	–	–	3043	1.74	**25**	**1.56**	**25**	**1.56**
2MTA	1086	0.83	–	–	606	1.64	2423	2.11	1378	1.58	1124	1.58	208	1.42	**24**	**1.19**
2PCC	–	–	–	–	–	–	–	–	–	–	843	0.66	**14**	**2.21**	70	2.21
2SIC	113	1.24	173	1.86	**12**	**2.04**	35	1.94	1072	1.79	1429	2.35	–	–	–	–
2SNI	–	–	–	–	–	–	–	–	362	1.92	377	1.92	72	2.15	**51**	**1.98**
7CEI	241	2.49	106	1.97	–	–	1515	2.05	1188	1.04	598	0.85	197	1.46	**39**	**1.46**
Antibody – Antigen																
1AHW	168	1.3	268	2.28	**5**	**1.34**	1419	2.05	–	–	–	–	319	2.29	269	2.29
1BVK	–	–	–	–	–	–	–	–	801	2.21	**560**	**2.21**	935	2.14	576	2.14
1DQJ	–	–	2287	2.48	–	–	–	–	–	–	–	–	**933**	**1.32**	1009	1.32
1E6J	3483	2.29	**15**	**1.56**	439	2.18	3065	2.49	–	–	–	–	631	2.21	83	2.12
1JPS	1185	1.89	**171**	**1.81**	292	0.9	469	1.56	484	1.24	702	1.17	–	–	–	–
1MLC	847	0.98	110	1.19	1834	1.16	1027	0.96	–	–	–	–	524	2.18	**54**	**2.18**
1VFB	1541	2.48	2734	1.79	1303	1.69	**207**	**2.23**	310	0.75	213	0.75	611	2.09	526	2.09
1WEJ	2152	1.25	465	2.37	–	–	–	–	–	–	–	–	386	2.26	**269**	**2.34**
2VIS	–	–	**2747**	**2.49**	–	–	3027	1.45	–	–	–	–	–	–	–	–
Antigen–Bound Antibody																
1BJ1	–	–	129	0.86	298	1.86	2052	1.58	–	–	–	–	143	2.32	**14**	**2.17**
1FSK	420	2.08	**1**	**1.63**	15	2.4	47	0.62	–	–	–	–	46	0.91	64	0.91
1I9R	–	–	**50**	**2.45**	95	2.39	302	2.48	2739	1.51	2090	1.51	71	2.28	96	2.28
1IQD	3228	2.12	612	2.27	**41**	**1.2**	–	–	–	–	–	–	374	2.13	75	2.13
1K4C	–	–	–	–	**1188**	**1.43**	–	–	–	–	–	–	–	–	–	–
1KXQ	11	1.5	212	1.91	73	1.68	**30**	**1.41**	646	1.36	528	1.39	308	2.2	387	2.18
1NSN	1254	1.76	185	1.96	945	2.29	1364	2.03	–	–	–	–	179	2.01	**115**	**2.01**
1NCA	575	1.45	**14**	**1.93**	–	–	600	0.85	–	–	–	–	232	1.21	257	1.21
1QFW	1457	1.85	257	1.14	**108**	**1.24**	759	1.08	1372	1.34	1212	1.34	–	–	–	–
2JEL	1142	0.95	45	1.79	133	2.49	–	–	–	–	–	–	83	2.16	**9**	**1.89**
2HMI	–	–	**237**	**2.5**	–	–	–	–	–	–	–	–	–	–	–	–
Others																
1A2K	–	–	–	–	–	–	237	2.45	–	–	–	–	**13**	**2.21**	27	2.21
1AKJ	–	–	–	–	–	–	292	2.23	102	1.45	**46**	**1.45**	47	2.04	484	2.04
1B6C	201	2.14	1717	2.43	1001	2.41	–	–	1862	1.96	1687	1.96	172	2.06	**159**	**2.06**
1BUH	625	2.37	–	–	–	–	391	1.78	65	0.75	**64**	**0.75**	357	1.62	735	1.62
1E96	–	–	3094	2.26	216	2.14	3526	2.5	300	1.79	**193**	**1.79**	–	–	–	–
1F51	650	2.03	230	2.18	3545	1.58	3561	2.12	–	–	–	–	**79**	**2.21**	154	2.21
1FAK	–	–	–	–	–	–	–	–	–	–	–	–	993	2.11	**146**	**1.87**
1FQJ	3004	2.46	–	–	–	–	–	–	27	2.12	30	2.1	**7**	**2.17**	19	2.17
1GCQ	–	–	–	–	–	–	1787	2.21	–	–	–	–	681	1.93	**231**	**1.93**
1GP2	–	–	–	–	–	–	–	–	–	–	–	–	764	2.21	**288**	**2.21**
1GRN	831	1.54	1704	2.34	1407	2.18	1724	1.61	1264	2.23	674	2.23	**501**	**2.17**	692	2.17
1HE1	33	2.16	–	–	267	1.98	3107	1.41	**1**	**1.12**	**1**	**1.12**	–	–	–	–
1HE8	–	–	–	–	–	–	**646**	**2.27**	–	–	–	–	1589	2.32	898	2.32
1I2M	–	–	–	–	–	–	–	–	–	–	–	–	**210**	**1.57**	482	1.57
1I4D	–	–	–	–	–	–	–	–	–	–	–	–	**647**	**2.14**	1186	2.24
1IB1	–	–	–	–	–	–	–	–	–	–	–	–	16	2.27	**4**	**2.27**

1IJK	–	–	–	–	–	–	1639	2.42	2221	2.5	**1426**	**2.43**	–	–	–	–
1KAC	–	–	2896	2.33	655	2.18	138	2.38	747	1.67	672	1.67	8	2.12	**5**	**2.12**
1KLU	–	–	–	–	–	–	–	–	–	–	–	–	2450	1.89	**1861**	**1.67**
1KTZ	–	–	–	–	–	–	–	–	–	–	–	–	286	1.23	**17**	**1.23**
1KXP	37	2.49	1734	2.36	–	–	**3**	**1.7**	306	2.01	157	2.01	100	0.87	8	0.87
1ML0	450	1.59	**36**	**1.56**	559	2.38	303	1.87	–	–	–	–	714	1.92	522	1.92
1QA9	3039	1.86	–	–	1381	2.19	**1264**	**2.16**	–	–	–	–	–	–	–	–
1WQ1	–	–	1101	2.49	141	1.87	–	–	96	1.95	62	1.95	**7**	**1.76**	102	1.76
2BTF	–	–	–	–	–	–	–	–	–	–	–	–	**236**	**1.72**	363	1.56
2QFW	1018	1.68	832	2.29	**68**	**1.55**	–	–	525	1.18	427	1.18	–	–	–	–

4.6.3 Computational Issues

In Table 4.8, the average computation times for various tasks of the proposed approach are presented. The average time required for extraction of the Local Spectral Descriptor for a GSP is 0.1s. The time required for matching between a pair of GSPs (using the Local Spectral Descriptor) is ~ 0.001ms. It is obvious that descriptor matching is $8 \cdot 10^4$ times faster than the geometric scoring based on distance grid, which demonstrates the importance of the Local Spectral Descriptor as a fast filtering stage.

Table 4.8: Average computation times for various tasks of the proposed approach

Activity	Average Computation
Local Spectral Descriptor Extraction /	100ms
Complementarity matching of a pair of	0.001ms
Scoring (distance grid) of a pose	86ms

The average running time of SP-Dock is the sum of a) the time required for preprocessing and descriptor extraction; b) descriptor matching, grouping and alignment; c) distance-grid-based geometric scoring and d) physicochemical scoring. The most time-consuming parts are the geometric and physicochemical scoring, while the fastest part is the descriptor matching, grouping and alignment. The average running time for small-to-medium-sized complexes is approximately one hour (Table 4.9), while it takes a few hours for large complexes. The times reported were obtained using a PC with a dual-core 2.4 GHz processor and 8GB RAM.

Table 4.9: Average running time of the proposed method

Average Running Time				
Preprocessing/	Descriptor Matching,	Geometric	Physicochemical	Average
300s	85s	1935s	1340s	3660s

Although we did not run the other three methods, we compare our algorithm with the times reported in the related articles. As stated in [36], LZerD requires 1-2 hours for small proteins and it may take longer for larger proteins. These numbers were obtained using a computer with dual-core 2.1 GHz processor with 8 GB RAM, i.e. similar to the PC that we conducted our experiments. Thus, our approach is slightly faster than LZerD, while at the same time it clearly outperforms LZerD. In [37], authors use a computer with i7 quad-core processor at 3.2GHz and 12GB RAM. The average running time for shDock is reported to be 2758s, i.e. a bit less than SP-Dock and LZerD. Taking into account the fact that in shDock a higher performance computer is used, it can be inferred that the average running time is comparable to SP-Dock and LZerD. Finally, in [35], no specific running time is reported for F^2Dock. It should be stressed that the running time for SP-Dock includes also the time for physicochemical scoring, while in the cases of LZerD and shDock only geometric docking is considered. If we keep only the geometric part, our method becomes much faster than LZerD and shDock. On the other hand, if we use both shape and physicochemical properties, we produce much better docking results within approximately the same running time.

4.7 Summary

We have presented a unified framework for protein-protein docking based on both shape and physicochemical complementarity. For shape complementarity, a new approach has been implemented, which utilises an effective local descriptor. The so-called Local Spectral Descriptror is compact, fast to extract and capable to capture similatities of local surface patches. As a next step, multiple pairs of complementary local patches from the receptor and the ligand are grouped together using a new grouping algorithm. The above grouping algorithm was inspired by the observation that shape complementarity in protein surfaces is mostly approximate rather than exact, thus single-patch or two-patch complementary matching generates numerous false-positive predictions. Additionally, shape complementarity is enhanced by physicochemical complementarity. Several non-geometric factors were tested and their contribution to the improvement of the shape-only docking predictions was assessed. Particle Swarm Optimization was applied to train the weights that each factor contributes to the overall scoring function. The most significant improvement is achieved when Atom Desolvation Energy, Electrostatic Complementarity, Hydrophobicity, Coulomb Potential and van der Waals Potential are introduced along with the shape complementarity, while Residue Contact Preferences and Generic Contact Preference seem to have insignificant contribution. This was an initial selection of the most well-known non-geometric factors. More factors that are available in the literature can be tested and assessed in a similar manner, which is planned for future work.

Results performed on the 84 complexes of the Docking Benchmark 2.4 demonstrate the superiority of the proposed SP-Dock method over five similar docking approaches. While in the case of bound complexes our method performs slightly better than the best docking methods reported so far, in the unbound case our approach clearly outperforms them.

Chapter 5

Similarity Search of Flexible 3D Molecules combining Local and Global Shape Descriptors

5.1 Introduction

In this chapter, a framework for shape-based similarity search of 3D molecular structures is presented. The proposed framework exploits simultaneously the discriminative capabilities of a global, a local and a hybrid local-global shape feature to produce a geometric descriptor that achieves higher retrieval accuracy than each feature does separately. Global and hybrid features are extracted using pairwise computations of diffusion distances between the points of the molecular surface, while the local feature is based on accumulating pairwise relations among oriented surface points into local histograms. The local features are integrated into a global descriptor vector using the bag-of-features approach. Due to the intrinsic property of its constituting shape features to be invariant to articulations of the 3D objects, the framework is appropriate for similarity search of flexible 3D molecules, while at the same time it is also accurate in retrieving rigid 3D molecules. The proposed framework is evaluated in flexible and rigid shape matching of 3D protein structures as well as in shape-based virtual screening of large ligand databases whith quite promising results.

The chapter is organized as follows: an overview of the method is given in Section 5.2, along with its major contributions. In Section 5.3, the preprocessing stage is described, which includes computation of molecular surface and selection of sample points and keypoints. The methodology for extracting the global descriptor based on diffusion distances is analysed in Section 5.4. Section 5.5 describes the extraction process of the local and the hybrid local-global features, along with the combined matching scheme that includes the global, the local and the hybrid feature. Experiments performed in four benchmark datasets are reported in Section 5.6. Finally, a summary is provided in Section 5.7.

5.2 Method Overview and Contributions

In Figure 5.1, the block diagram of the proposed method is depicted. The crystal structure of the molecule is given as input (e.g. PDB file) and its Solvent Excluded Surface (SESs) is generated in the form of a triangulated mesh. Then, a mesh simplification step is performed on SES, resulting in two sets of points: a set of N_S oriented points and a set of N_K keypoints ($N_K < N_S$) that provide a coarser representation of the 3D molecule. In the descriptor extraction step, two different descriptor vectors are proposed in this work: the *Bag of Augmented Local Descriptor (BoALD)* and the *Modal Representation of the Diffusion-Distance Matrix* (*DDMR* descriptor). These descriptor vectors are combined into a common distance measure in order to calculate the dissimilarity between the query molecule and the molecules of a database. The main innovative points of the proposed framework are outlined in the sequel.

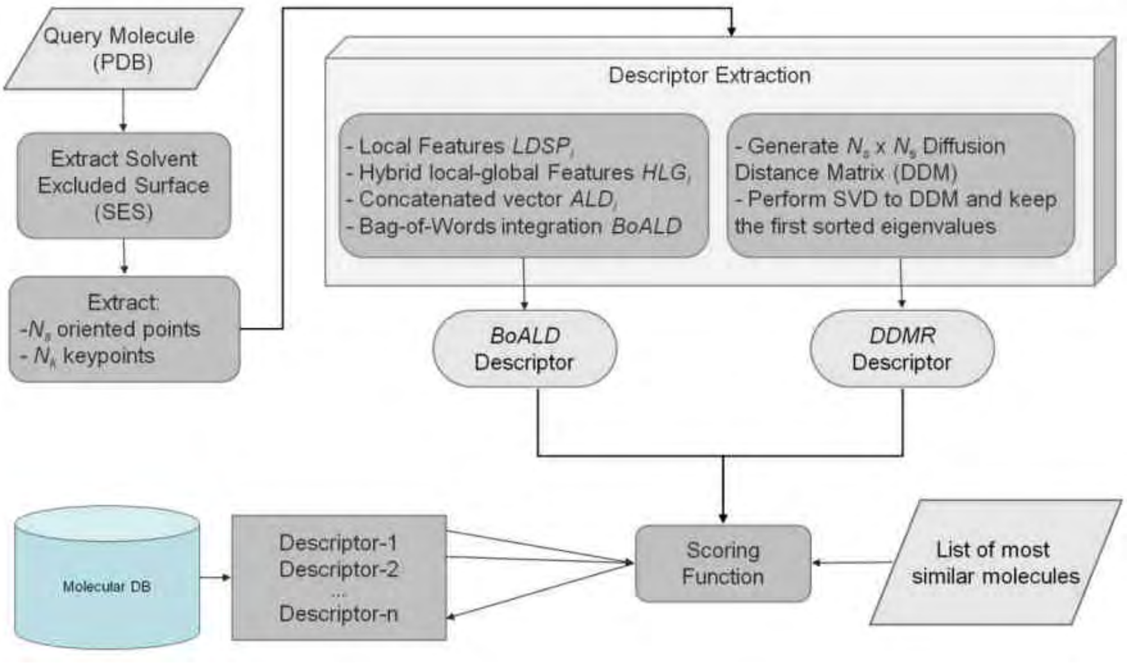

Query Molecule (PDB)
Extract Solvent Excluded Surface (SES)
Extract: -N_s oriented points - N_k keypoints
Descriptor Extraction
- Local Features *LDSP*, - Hybrid local-global Features *HLG*, - Concatenated vector *ALD*, - Bag-of-Words integration *BoALD*
- Generate N_s x N_s Diffusion Distance Matrix (DDM) - Perform SVD to DDM and keep the first sorted eigenvalues
BoALD Descriptor
DDMR Descriptor
Molecular DB
Descriptor-1
Descriptor-2
...
Descriptor-n
Scoring Function
List of most similar molecules

Figure 5.1: Block diagram of the proposed method.

First of all, to the best of the authors' knowledge, it is the first time that *a global, a local and a hybrid local-global feature are combined* into a unified descriptor to address shape similarity search of flexible 3D molecules. This combination captures different properties of the 3D structure of molecules resulting in a robust shape descriptor appropriate for both rigid and non-rigid shape similarity problems.

Although Diffusion Distances have been already exploited to address non-rigid shape retrieval problems [73], we have further extended this work. Instead of computing histograms of diffusion distances between all sample point pairs on the molecular surface, we provide a Modal Representation (MR) by performing singular value decomposition (SVD) of the Diffusion Distance Matrix (DDM), a matrix that summarizes all point-to-point diffusion distances on the molecular mesh. This extension improves the accuracy in similarity matching of flexible molecules over existing diffusion-distance-based approaches.

Combining a local and a hybrid local-global feature for non-rigid shape retrieval was initially proposed in [80], where the local geometrical feature was augmented with its spatial context computed as histogram of diffusion distances computed over mesh surface. In our case, the local geometrical feature of [80] is substituted by a more discriminative shape descriptor, which is based on Surflet-Pair relations [105]. The resulting *Augmented Local Descriptor (ALD)* improves significantly the retrieval accuracy. Additionally, the diffusion distances for the computation of the hybrid local-global feature are directly derived from the DDMR descriptor computation step, instead of using the manifold ranking method of [80], to speed-up the descriptor extraction process.

The proposed unified framework demonstrates superior performance to existing methods for non-rigid shape matching. Experiments performed in a benchmark Database of Macromolecular Movements (MolMovDB) [106] show that our method clearly outperforms other state-of-the-art approaches.

Although the proposed method was initially designed to address non-rigid shape matching problems, it achieves high accuracy in retrieval of rigid molecules as well. More specifically, the proposed framework outperforms existing molecular shape matching approaches in three datasets. Thus, the proposed framework is applicable to both rigid and non-rigid problems.

Finally, it is worth mentioning that the resulting shape descriptors constitute a compact representation of the molecular shape. This makes the method appropriate for problems of virtual screening in very large databases, since it achieves both high accuracy and fast retrieval. This is analysed in Experiments section, where the proposed method is tested on two large-scale virtual screening benchmarks.

5.3 Preprocessing

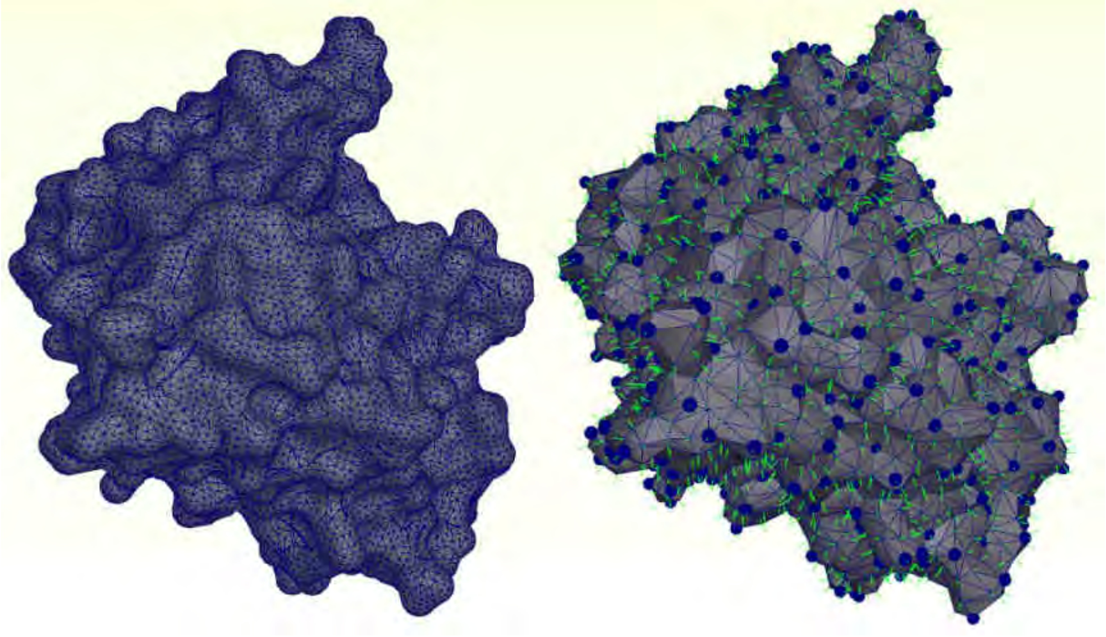
This section describes the preprocessing procedure, which consists of two steps: the first step involves computation of the Solvent Excluded Surface (SESs) of the molecule, while, during the second step, the SES is remeshed so that each molecule is represented by the same number of oriented points. These preprocessing steps are required for both descriptor extraction processes.

Input to the system is the crystal structure of the molecule (e.g. in PDB file format), which represents its atoms in the 3-dimensional space (*x*, *y*, *z* coordinates). In order to generate a SES, the Maximal Speed Molecular Surface (MSMS) [10] software has been utilized, which is based on rolling a probe sphere (of size equal to the size of the solvent molecule) over the exposed contact surface of each atom.

The output mesh is then used for the extraction of global and local shape descriptors. In order to apply the descriptor extraction algorithms, all molecules of the dataset should have the same number of mesh vertices. Since by using the MSMS software we cannot determine the exact number of the extracted vertices, a remeshing step follows to produce a mesh with the exact number of vertices N_S. For this remeshing, the Computational Geometry Algorithms Library (CGAL[2]) has been used. Let $\mathbf{p}_i$ be the i^{th} vertex, $i = 1,\ldots,N_S$. For each $\mathbf{p}_i$ its normal vector $\mathbf{n}_i$ is computed resulting in a set of N_S oriented points $(\mathbf{p}_i, \mathbf{n}_i)$. These oriented points are further sub-sampled to generate a new set of N_K keypoints $\mathbf{q}_i$, $i = 1,\ldots,N_K$, where $N_K < N_S$, that provide a coarser representation of the 3D molecule. Sub-sampling is performed using quasi-random

[2] http://www.cgal.org/

sequence, which is a deterministic sequence that produces sample points more uniformly distributed than a pseudo-random sequence [107]. In Figure 5.2a, the SES of a protein is depicted. This mesh consists of 25144 vertices and 50278 faces. The new surface after the remeshing step consists of 3000 vertices and 5996 faces and it is shown in Figure 5.2b. The normals $\mathbf{n}_i$ of the N_S oriented points are given in green lines, while the dark blue spheres depict the centers of the N_K sub-sampled points.



a) b)

Figure 5.2: a) the SES of a protein that consists of 25144 vertices and 50278 faces; b) the surface that is produced after the remeshing step, consisting of N_S = 3000 vertices and 5996 faces. The dark blue spheres depict the N_K =500 sub-sampled points, while the green lines depict the normals n$_i$.

5.4 A Global Shape Descriptor Based on Diffusion Distances

Distance-based descriptors constitute a category of 3D shape descriptors, which are based on computation of pairwise distances between surface vertices. These distances are usually accumulated into a histogram, which provides the final descriptor vector. The first attempts in this category were using the Euclidean distance, such as the Euclidean Distance (ED) descriptor [108]. However, Euclidean metrics were proven inappropriate to deal with the articulation of 3D objects [72], [73]. To address this issue, the Euclidiean space or Euclidean metrics should be transformed to a metric space where the pairwise distances between points of the 3D object surface are invariant to deformations of the 3D object. Thus, geodesic distances (GD) [77], inner distances (ID) [72] or diffusion distances (DD) [73] have been used instead.

GD refers to the length of the shortest path between two points along the boundary surface. Advantages of GD is its invariance to surface bending, however, it is not quite efficient in capturing shape articulation deformation that commonly exists in macromolecular movements (e.g. hinge motion). ID overcomes some disadvantages of GD. ID distance is computed as the length of the shortest path between landmark points within the molecular shape. It outperforms GD in capturing deformation of molecular structure [72]. A drawback of ID is that it may be significantly affected by topological changes of shape deformation [73]. Invariance to topological changes can be achieved by using diffusion distances (DD). DD is the probability of travelling on the surface from one point to another in a fixed number of random steps.

The difference of DD comparing to GD and ID is that DD is computed as the average length of paths connecting two points, while GD and ID represent the length of the shortest path. Thus, DD is usually more robust to topological changes. In our framework, the diffusion distance was selected as a base to the global shape descriptor. The computation of DD on the molecular surface as well as the extension of DD with a modal representation are provided in the following subsections.

5.4.1 Computing Diffusion Distances on the Molecular Surface

The computation of diffusion distances is performed in three main steps: (a) calculation of the Markov probability matrix; (b) Singular Value Decomposition (SVD) of the matrix to generate the diffusion map space; and (c) computation of the diffusion distances.

Let $\mathbf{p}_i$ be the set of N_S vertices. Let $K(\cdot)$ be a kernel function with bandwidth h. The Gaussian kernel $K(\mathbf{p}_i, \mathbf{p}_j) = \exp\left(-\left\|\mathbf{p}_i - \mathbf{p}_j\right\| / 2h^2\right)$ is one of the most commonly used, where the bandwidth h controls the local scale of each data point's neighborhood and $\left\|\mathbf{p}_i - \mathbf{p}_j\right\|^2$ is the Euclidean distance between surface points i and j. Then, the diffusion matrix $\mathbf{L}$ with elements $L_{ij} = K(\mathbf{p}_i, \mathbf{p}_j)$ is normalized as $\mathbf{M} = \mathbf{D}^{-1}\mathbf{L}\mathbf{D}^{-1}$ by the degree matrix $\mathbf{D}$ with $D_{ij} = \sum_i L_{ij}$. The normalized diffusion matrix $\mathbf{M}$ is a stochastic matrix with all row sums equal to one, and according to [109] it can be interpreted as a random walk on a graph, where the vertices of the graph are the surface points $i = 1,\ldots,N_S$ and the weights of the $\langle i, j \rangle$ edges correspond to M_{ij} values. Thus, M_{ij} denotes the $p(1,i \mid j)$ transition probability from the surface point j to point i in one time step ($t = 1$). For any finite time t the Markov probability matrix $\mathbf{M}^t$ with elements M_{ij}^t is computed as $M_{ij}^t = p(t,i \mid j)$, expressing the probability distribution of reaching surface point i, given a starting point j at time $t = 0$. Thus, the transition probability is given by $p(t,i \mid j) = \mathbf{e}_j \mathbf{M}^t$, where $\mathbf{e}_j$ is a row vector of zeros with a single entry equal to one at the *j*-th coordinate. Let the SVD of matrix $\mathbf{M}^t$ be:

$$\mathbf{M}^t = \mathbf{A}\,\boldsymbol{\Sigma}\,\mathbf{B}^T \tag{5.1}$$

where $\boldsymbol{\Sigma} = diag(\sigma_0, \sigma_1, \ldots, \sigma_k)$ and $\sigma_0 \geq \sigma_1 \geq \ldots \geq \sigma_k \geq 0$ are the $k+1$ singular values of $\mathbf{M}^t$, $\mathbf{A} = [\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_k]$ and $\mathbf{B} = [\mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_k]$ with $\mathbf{a}_i = \{a_i(1), a_i(2), \ldots, a_i(N_S)\}$ and $\mathbf{b}_i = \{b_i(1), b_i(2), \ldots, b_i(N_S)\}$ are the left and right singular vectors, respectively, and $\mathbf{a}_0$ and $\mathbf{b}_0$ are the first left and right eigenvectors, corresponding to the first $(\sigma_0 = 1)$ eigenvalue. Note that following [109], the first eigenvalue and the respective eigenvectors are excluded from the diffusion process and are used only for normalization purposes. The distance between two surface points i, j at time t is calculated as:

$$D_t^2(i,j) = \left\| p(t,y \mid i) - p(t,y \mid j) \right\|_w^2 = \sum_{y \in \mathbf{Y}} \left(p(t,y \mid i) - p(t,y \mid j) \right)^2 w(y) \tag{5.2}$$

where Y is the set of the N_S surface points and $\forall y \in Y$ the $w(y)$ value is treated as weight function and calculated as $w(y) = 1/a_0(y)$.

Since the $D_t^2(i,j)$ distance depends on the random walk on the graph, it denotes the diffusion distance at time t. As it is mathematically proven in [109], the diffusion distance between surface points i, j is calculated by:

$$D_t^2(i,j) = \left\| \mathbf{\Psi}_t(i) - \mathbf{\Psi}_t(j) \right\|^2 \tag{5.3}$$

with $\mathbf{\Psi}_t(i) = \left(\sigma_1^t \cdot b_1(i), \sigma_2^t \cdot b_2(i), \ldots, \sigma_k^t \cdot b_k(i)\right)$ is the mapping of the *i*-th surface point from the original kernel space (formed by the kernel function $K(\cdot)$) to the diffusion map space at time t .

5.4.2 Modal Representation of Diffusion Distance Matrix

Given the computation of diffusion distances between the molecular surface points, the next step is to exploit this intrinsic feature for the computation of a global shape descriptor. A common technique is to accumulate these pairwise distances into a histogram [73]. In this work, we propose an alternative approach based on a modal representation.

As a first step, the Diffusion Distance Matrix $\mathbf{DDM} = \{D_t^2(i,j)\}$, where $i,j = 1,\ldots,N_S$, is computed. Then, SVD of $\mathbf{DDM}$ yields:

$$\mathbf{DDM} = \mathbf{U}\,\mathbf{L}\,\mathbf{V}^T \tag{5.4}$$

where the modal representation, i.e. the singular value matrix $\mathbf{L} = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$, contains the intrinsic information about geometry, and matrices $\mathbf{U}, \mathbf{V}$ contain the information about correspondences between points. The first n singular values $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ constitute the *Modal Representation of Diffusion Distance* (*DDMR*) descriptor $\mathbf{D}^{DDMR}$ of the 3D object. It has been proven in [77] that the eigenvalue matrix is invariant to sampling order of the surface points.

5.5 An Augmented Local Descriptor

The proposed Augmented Local Descriptor (ALD) is computed on each of the N_K keyponts that provide a coarse approximation of the molecular surface. This results in a total of N_K ALD descriptor vectors $\mathbf{D}_i^{ALD}$ (*i*=1,…, N_K) that are extracted for each 3D molecule. Each descriptor $\mathbf{D}_i^{ALD}$ is the concatenation of two vectors: the Local Descriptor based on Surflet-Pair Relations $\mathbf{D}_i^{LDSP}$ and the Hybrid Local-Global feature $\mathbf{D}_i^{HLG}$, i.e. $\mathbf{D}_i^{ALD} = (\mathbf{D}_i^{LDSP}, \mathbf{D}_i^{HLG})$. Finally, a Bag-of-

Features approach applied on the N_K ALD descriptors produces a global descriptor vector, the Bag-of-ALD (BoALD).

5.5.1 A Local Descriptor based on Surflet-Pair Relations

This descriptor takes as input the set of oriented points $(\mathbf{p}_i, \mathbf{n}_i)$, $i = 1, \ldots, N_S$ and the set of keypoints $\mathbf{q}_i$, $i = 1, \ldots, N_K$ (Figure 5.2b). Each local descriptor is defined on a spherical region of radius R centered at each keypoint $\mathbf{q}_i$. Let $Q = \{(\mathbf{p}_1, \mathbf{n}_1), (\mathbf{p}_2, \mathbf{n}_2), \ldots, (\mathbf{p}_N, \mathbf{n}_N)\}$ be the set of oriented points within a spherical region around keypoint $\mathbf{q}$ with $\|\mathbf{p}_1 - \mathbf{q}\| \leq R$ and $\mathbf{n}_q$ the normal vector of $\mathbf{q}$. The *Local Descriptor based on Surflet-Pair relations* (*LDSP*) is computed as follows:

For each oriented point $(\mathbf{p}_i, \mathbf{n}_i) \in Q$ the following values α, β, γ and δ are computed:

$$\alpha = \arctan(\mathbf{w} \cdot \mathbf{n}_i, \mathbf{n}_q \cdot \mathbf{n}_i), \tag{5.5}$$

$$\beta = \mathbf{v} \cdot \mathbf{n}_i, \tag{5.6}$$

$$\gamma = \mathbf{n}_q \cdot \frac{\mathbf{p}_i - \mathbf{q}}{\|\mathbf{p}_i - \mathbf{q}\|}, \tag{5.7}$$

$$\delta = \|\mathbf{p}_i - \mathbf{q}\| \tag{5.8}$$

where $\mathbf{v} = \frac{(\mathbf{p}_i - \mathbf{q}) \times \mathbf{n}_q}{\|(\mathbf{p}_i - \mathbf{q}) \times \mathbf{n}_q\|}$ and $\mathbf{w} = \mathbf{n}_q \times \mathbf{v}$.

The above process is illustrated in Figure 5.3. The N_S points of the molecule are given as a point cloud (small black dots), while the larger dots represent the N_K keypoints. In the magnified view, on the right, the oriented points $(\mathbf{p}_i, \mathbf{n}_i)$ within the spherical region centered at $\mathbf{q}$ are depicted.

Using (5.5)-(5.8), a 4-tuple $(\alpha, \beta, \gamma, \delta)$ is created for each oriented point $(\mathbf{p}_i, \mathbf{n}_i)$ of set Q. Then, all 4-tuples of Q are collected into a 4-dimensional joint histogram. Let k_L be the number of bins for each dimension, then the total number of bins of the 4D histogram is k_L^4. Each 4-tuple

$(\alpha,\beta,\gamma,\delta)$ is assigned to bin (i,j,l,m) with $0 \le i,j,l,m \le k_L - 1$ according to the following inequalities:

$$hmin_\alpha + i \cdot hstep_\alpha \le \alpha < hmin_\alpha + (i+1)\cdot hstep_\alpha\,, \tag{5.9}$$

$$hmin_\beta + j \cdot hstep_\beta \le \beta < hmin_\beta + (j+1)\cdot hstep_\beta\,, \tag{5.10}$$

$$hmin_\gamma + l \cdot hstep_\gamma \le \gamma < hmin_\gamma + (l+1)\cdot hstep_\gamma\,, \tag{5.11}$$

$$hmin_\delta + m \cdot hstep_\delta \le \delta < hmin_\delta + (m+1)\cdot hstep_\delta\,, \tag{5.12}$$

where α ranges from $hmin_\alpha$ to $hmax_\alpha = hmin_\alpha + k \cdot hstep_\alpha$, β ranges from $hmin_\beta$ to $hmax_\beta = hmin_\beta + k \cdot hstep_\beta$ and so on.

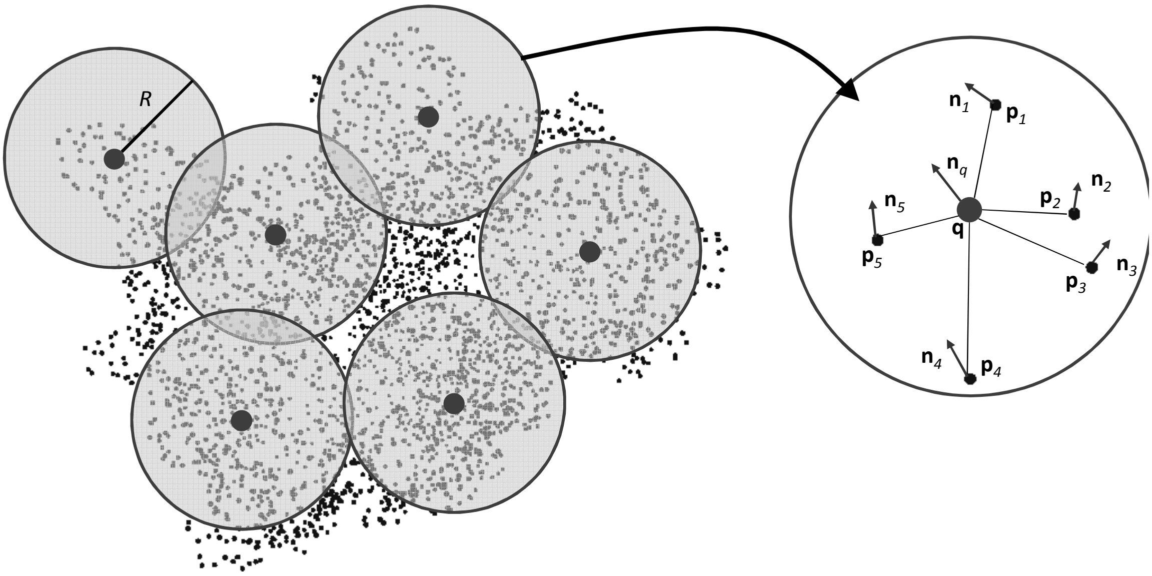


Figure 5.3: The protein surface given as point cloud. The circles represent the local spherical regions centered at the local keypoints N_K. On the right, the oriented points within a local spherical region are given in a magnified view.

The LDSP $\mathbf{D}_i^{LDSP}$ for each keypoint $\mathbf{q}_i$ is a 1D vector of dimension k_L^4. The values of $\mathbf{D}_i^{LDSP}$ are normalized so that their sum equals 1. The selection of parameter k_L should be such that the number of bins is adequate to produce a discriminative descriptor, while at the same time k_L is not very high so as to keep the descriptor dimensionality low. In [105], $k_L = 5$ was reported as an optimal solution, which was also verified in our case: for $k_L < 5$, the discriminative power of the local feature was negatively affected, while for $k_L > 5$, the descriptor dimensionality was increasing dramatically without achieving significant improvement of accuracy. The optimal value for radius R has been estimated in a similar manner: very low values of R result in spherical regions with trivial shape information; for very high values of R, the local character of the descriptor, which gives its robustness to non-rigid problems, disappears. Eventually, an optimal choice for our experiments was $R = 0.4 \cdot R_A$, where R_A is the radius of the 3D molecule's smallest bounding sphere.

Regarding the values $hmin, hmax$ and $hstep$ of (5.9)-(5.12), their selection is straightforward: since α is an arctan function, $hmin_\alpha = -\pi$ and $hmax_\alpha = \pi$; β, γ represent dot products of unit vectors, thus $hmin_\beta = hmin_\gamma = -1$ and $hmax_\beta = hmax_\gamma = 1$; δ is practically the distance of each point $\mathbf{p}_i$ from the center keypoint $\mathbf{q}$, thus $hmin_\delta = 0$ and $hmax_\delta = R$; finally, $hstep = (hmax - hmin)/k_L$ in all cases.

5.5.2 A Hybrid Local-Global Feature

Similar to LDSP, the *Hybrid Local-Global feature* (*HLG*) is computed for each keypoint $\mathbf{q}_i$, $i = 1,\ldots,N_K$. More specifically, the following set is computed for each $\mathbf{q}_i$:

$$DD_{\mathbf{q}_i} = \{dd(\mathbf{q}_i, \mathbf{p}_1), dd(\mathbf{q}_i, \mathbf{p}_2), \ldots, dd(\mathbf{q}_i, \mathbf{p}_{N_S}),\}, \tag{5.13}$$

where $dd(\mathbf{q}_i, \mathbf{p}_j)$ is the diffusion distance from the keypoint $\mathbf{q}_i$ to sample point $\mathbf{p}_j$, $j = 1,\ldots,N_S$. The N_S diffusion distances of the set $DD_{\mathbf{q}_i}$ are accumulated into a 1D histogram of $k_H = 100$ bins. Again, the dimension k_h has been experimentally determined [80]. This histogram, which is normalized so that the sum of all values equals 1, constitutes the *HLG descriptor* $\mathbf{D}_i^{HLG}$ of keypoint $\mathbf{q}_i$.

According to the above definition, the HLG descriptor is neither a purely local feature nor a global descriptor. It combines local characteristics – as it is computed for each keypoint – with global characteristics – as it takes into account the set of diffusion distances of the entire molecule. HLG resembles to the Local Distance Feature (LDF) that was proposed in [80].

However, in [80], the distances to all points $\mathbf{p}_j$ are computed using a Manifold Ranking algorithm [110], according to which each keypoint $\mathbf{q}_i$ is used as the source of diffusion of ranking score for the MR. The resulting histogram is created by all ranking scores at sample points $\mathbf{p}_j$. In this work, the distances $dd(\mathbf{q}_i, \mathbf{p}_j)$ are computed using the framework presented in Section 5.4.1. Thus, diffusion distances are computed only once for both the DDMR and the HLG descriptors.

5.5.3 Computation of Bag-of-ALD (BoALD) Descriptors

During this step, the local LDSP descriptors and the hybrid HLG descriptors are integrated into a global histogram. This process is summarized in Figure 5.4. Initially, for each keypoint $\mathbf{q}_i$, its LDSP descriptor $\mathbf{D}_i^{LDSP}$ and HLG descriptor $\mathbf{D}_i^{HLG}$ are concatenated into an ALD descriptor $\mathbf{D}_i^{ALD} = (\mathbf{D}_i^{LDSP}, \mathbf{D}_i^{HLG})$. $\mathbf{D}_i^{ALD}$ is a histogram of dimension $k_A = k_L^4 + k_H = 625 + 100 = 725$. The idea of concatenating histograms of two features into one histogram was inspired by the method in [80], where the Local Distance Feature (LDF) was merged with a local descriptor, the Local Geometrical Feature (LGF), in order to produce a descriptor with improved accuracy. In this work, LGF has been substituted by the LDSP descriptor.

To produce a global descriptor from the N_K local descriptors $\mathbf{D}_i^{ALD}$, the well-known Bag-of-Features approach has been utilized. Let $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{N_V}\}$ be a set of visual words. The dimension of each visual word is equal to k_A i.e. of the ALD histogram. The set V is created by applying *k*-means clustering to a subset (training set) of the ALD descriptors $\mathbf{D}_i^{ALD}$ of the molecular database. The descriptors that constitute the training set are selected randomly (10% of the local features of the database) in order to capture a representative view of the database. Each visual word $\mathbf{v}$ is the center of a cluster. Then, each ALD descriptor $\mathbf{D}_i^{ALD}$ of the 3D molecule is vector quantized into a visual word and a histogram of N_V visual words is produced. This histogram $\mathbf{D}^{BoALD}$ is called *Bag-of-ALD descriptors* or *BoALD*.

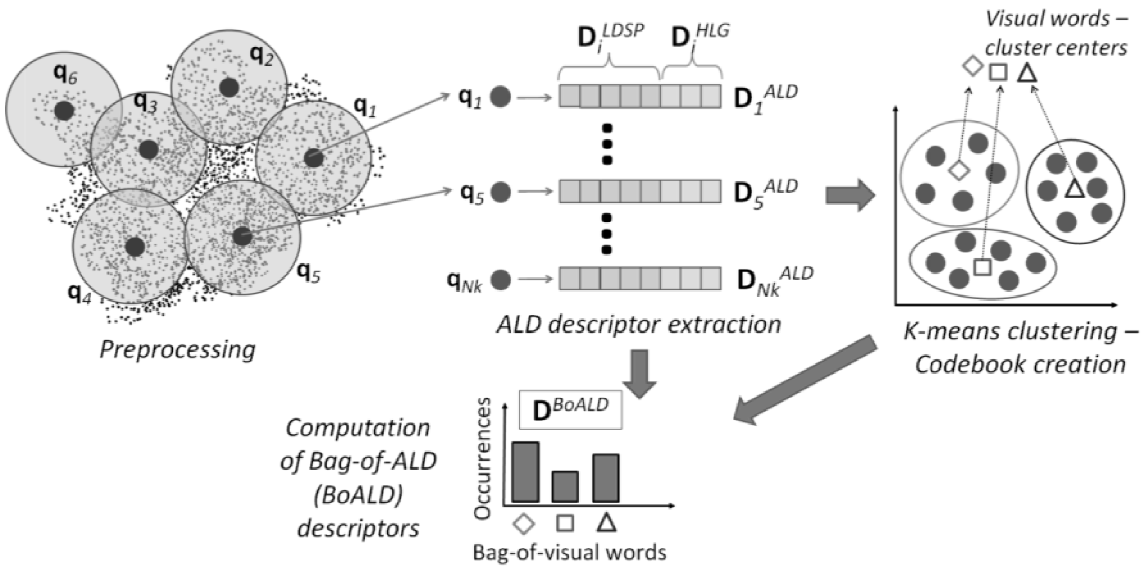

q6
q2
q3
q1
q4
q5
Preprocessing
D_i^LDSP
D_i^HLG
q1
D_1^ALD
q5
D_5^ALD
qNk
D_Nk^ALD
ALD descriptor extraction
Visual words – cluster centers
K-means clustering – Codebook creation
Computation of Bag-of-ALD (BoALD) descriptors
D^BoALD
Occurrences
Bag-of-visual words

Figure 5.4: The process for computing the BoALD descriptors.

The size of vocabulary N_V should be carefully chosen since it affects both retrieval accuracy and computational cost. For large datasets, which imply also a large number of samples to cluster, an increase of size N_V would require high computation times for the k-means clustering. On the other hand, retrieval accuracy is improved as vocabulary size increases, until a specific upper limit is reached, above which no further improvement is observed. Based on the aforementioned criteria, the optimal choice of vocabulary size is $N_V = 1000$, as it has been experimentally found.

5.5.4 A Distance Measure for Shape Similarity Matching

Let $\mathbf{D}^{DDMR}$ and $\mathbf{D}^{BoALD}$ be the DDMR and BoALD descriptor vectors that are extracted using the methods described in Sections 5.4 and 5.5, respectively. The overall shape dissimilarity between two 3D molecules A and B can be calculated as the weighted sum of the dissimilarities of each descriptor separately:

$$dis(A,B) = w^{DDMR} \cdot dis^{DDMR}(A,B) + w^{BoALD} \cdot dis^{BoALD}(A,B), \tag{5.14}$$

where dis^{DDMR} and dis^{BoALD} are the dissimilarities of DDMR and BoALD descriptors, respectively, and w^{DDMR}, w^{BoALD} their corresponding weights. In general, the selection of the optimal distance metric for each descriptor is not trivial. An extensive study on the performance of the most well-known dissimilarity metrics is available in [81]. In the case of the DDMR descriptor, the *X-Distance* (or *normalized Manhattan Distance*) was experimentally proven to be the optimal metric:

$$dis^{DDMR}(A,B)=2\cdot\sum_{i=1}^{N_D}\frac{\left|D_A^{DDMR}(i)-D_B^{DDMR}(i)\right|}{\left|D_A^{DDMR}(i)+D_B^{DDMR}(i)\right|}, \tag{5.15}$$

where $\mathbf{D}_A^{DDMR}$, $\mathbf{D}_B^{DDMR}$ are he descriptors of molecules A and B, respectively and N_D is the dimensionality of the descriptor vector. Similarly, the optimal distance metric for the BoALD descriptor is the *Kullback-Leibler Divergence*:

$$dis^{BoALD}(A,B)=\sum_{i=1}^{N_V}\left(D_A^{BoALD}(i)-D_B^{BoALD}(i)\right)\ln\frac{D_B^{BoALD}(i)}{D_A^{BoALD}(i)}, \tag{5.16}$$

where $\mathbf{D}_A^{BoALD}$, $\mathbf{D}_B^{BoALD}$ are the descriptors of molecules A and B, respectively and N_V is the dimensionality of the descriptor vector.

After selecting the optimal dissimilarity metrics, the weights w^{DDMR}, w^{BoALD} need to be determined. In our case, we followed the Particle Swarm Optimization (PSO) strategy [81] for the weight optimization. PSO is an algorithm for global optimization. It is motivated by the social behavior of organisms such as bird flocking and fish schooling. PSO optimizes a problem in which a best solution can be represented as a point or surface in an n-dimensional space. It iteratively tries to improve a candidate solution based on a given quality measure (fitness function). PSO establishes a population (swarm) of candidate solutions, known as particles that move around in the search space, and are guided by the best found positions, updated while better positions are found by the particles.

The population of candidate solutions, in our case, is the weights w^{DDMR}, w^{BoALD}, which can take arbitrary values between $[0,1]$. The fitness function to be optimized is the *average Tier-1 precision*, which is calculated on a train dataset. More specifically, each 3D molecule of the dataset is used as query to retrieve similar objects, using (14) as dissimilarity metric. The retrieved results are ranked in ascending order. The Tier-1 precision is given by the following equation:

$$P_{T1} = \frac{R^C(K)}{K}, K = |C| - 1 \tag{5.17}$$

where K is the number of first retrieved objects, $R^C(K)$ is the number of retrieved objects within the K-first, which are of the same class C with the query, and $|C|$ is the number of objects that belong to class C. PSO resulted in the following weights: $w^{DDMR} = 0.62$, $w^{BoALD} = 0.38$.

5.6 Experimental Results

For the experimental evaluation of the proposed method, four different datasets have been selected. The first dataset is part of the Database of Macromolecular Movements (MolMovDB) [106], which comprises molecules with large conformational changes (http://www.molmovdb.org/), also including the intermediate morphs. It consists of 2695 PDB files classified into 214 categories [111]. This dataset is used for parameter selection and for comparison with existing flexible molecular shape matching approaches [72], [73]. The second dataset consists of 2631 3D protein structures. It is a subset of the FSSP database [112] and was created by us to demonstrate the performance of the Spherical Trace Transform (STT) in [62]. The 2631 proteins are classified into 27 categories according to the DALI algorithm [113] and the proteins that belong to the same class demonstrate rigid shape similarity. This dataset is used to evaluate the performance of the proposed method in rigid shape matching of 3D protein structures and is available for download at vcl.iti.gr/protein_retrieval/PDB_FSSP.zip. Finally, the third and fourth dataset are used to demonstrate the performance of our framework in large-scale virtual screening of ligands. Experiments have been performed on a PC with i5 2.8GHz processor, 4GB RAM, running Windows 7.

5.6.1 Parameter Selection for the DDMR Descriptor

For the implementation of the DDMR descriptor, the Matlab Toolbox for Dimensionality Reduction[3] (v0.8.1) has been selected, using the default parameters $h = 1$ and $t = 1$. The discriminative power of DDMR mainly depends on two parameters: a) the number N_S of sample points $\mathbf{p}_i$ on the molecular surface, and b) the dimensionality of DDMR descriptor vector, i.e. the number n of first singular values $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ of SVD (5.4). By increasing the number of sample points N_S, a higher-quality representation of the surface is achieved and accuracy is improved, however, this results in higher descriptor extraction times. Additionally, an increase of n may also improve the accuracy. We run several sets of experiments using different values of N_S and n. As a performance metric, the average *Tier-1 Precision* has been selected (5.17).

In Figure 5.5, the average Tier-1 Precision for different values of N_S and n, in MolMovDB is presented. It is obvious that as the number of sample points N_S increases a higher precision is achieved. Using a mesh resolution higher than 2000 points, though, the improvement in accuracy is negligible. Similar conclusions are drawn regarding the number n of first singular values. For values n higher than 50-60, there is no significant improvement in precision.

A critical factor for the parameter selection is the descriptor extraction time. Since the process of extracting the DDMR descriptor involves computations on $N_S \times N_S$ matrices, the processing time may increase prohibitively as the number of sample points N_S increases. This is highlighted in Table 5.1, where it is obvious that for meshes consisting of 4000 points it takes approximately one minute for descriptor extraction, while for meshes of 1000 points the extraction time is less than 2 seconds. For the experiments that will be presented in the following subsections the values $N_S = 2000$ and $n = 50$ have been selected for DDMR.

[3] http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html

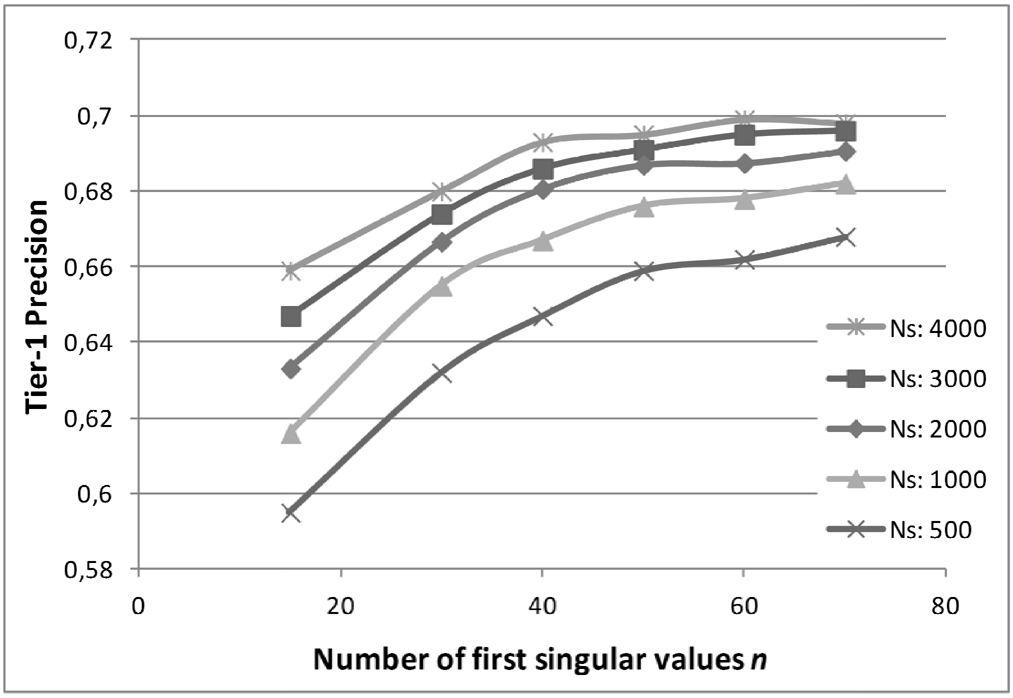


Figure 5.5: Parameter selection for DDMR descriptor: the average Tier-1 Precision in MolMovDB for different values of n and N_S.

Table 5.1: Average extraction times of the DDMR descriptor for different numbers of sample points.

Number of sample	DDMR Descriptor
500	0.47
1000	1.34
2000	4.69
3000	14.08
4000	46.52

5.6.2 Parameter Selection for the BoALD Descriptor

The BoALD descriptor has been implemented by us in C++ based on the works presented in [80] and [105]. The performance of BoALD is affected by several parameters: a) the radius R of the local descriptor LDSP; b) the number of sample surface points N_S; c) the number of local points N_K, and d) the vocabulary size N_V of the codebook. The number of surface points N_S is related to radius R as follows: a small R provides sufficient locality to the descriptor but it requires a high N_S so that the local histograms (5.9)-(5.12) are well populated. The number of local points N_K affects the selection of the vocabulary size N_V: for a given N_K, an increase of

N_V improves the accuracy until a specific upper limit is reached. Beyond that limit a further increase of N_V has no effect in accuracy. If we increase N_K, then we can achieve a higher upper limit for N_V resulting in a more discriminative descriptor.

In Figure 5.6, the average Tier-1 Precision of the BoLDSP descriptor in MolMovDB for different values of radius R and number of sample points N_S is depicted. BoLDSP is produced by applying bag-of-features to the local LDSP histograms only (while the full BoALD descriptor is produced by the concatenation of LDSP and HLG histograms). For lower resolution meshes (1000-2000 points) an increase of R improves precision, while for higher resolution meshes (3000-4000 points) increasing R over $0.4 \cdot R_A$ may have an opposite effect.

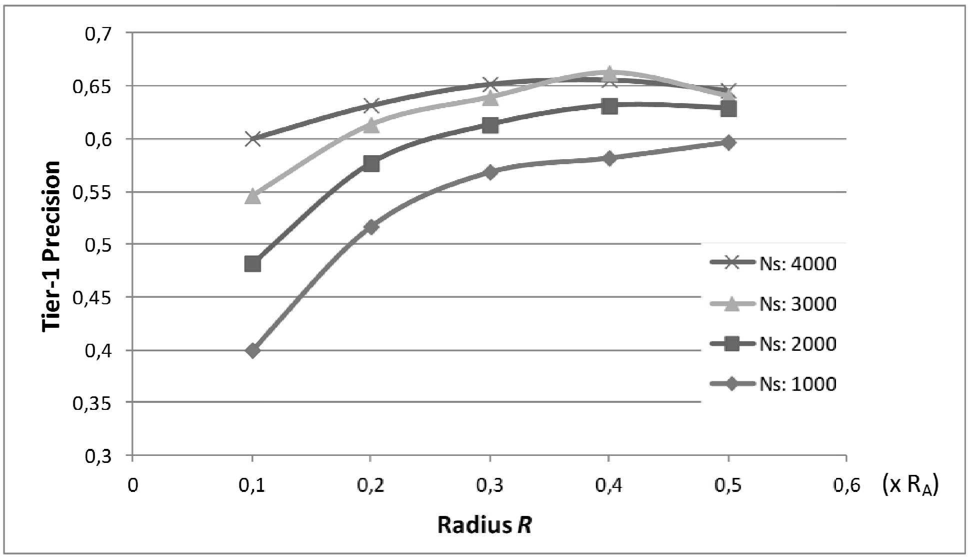


Figure 5.6: Parameter selection for BoLDSP descriptor: the average Tier-1 Precision in MolMovDB for different values of radius R and N_S.

In Figure 5.7, the average Tier-1 Precision of the BoALD descriptor in MolMovDB for different values of vocabulary size N_V and number of local points N_K is depicted. For $N_K = 250$, an increase of N_V does not affect the average precision. Similarly, for $N_K = 500$ the precision is not improved for $N_V \geq 1000$. Finally, for $N_K = 1000$, the improvement in accuracy comparing to $N_K = 500$ is negligible. It is also worth mentioning that the dimensionality N_V of BoALD should be kept relatively low to achieve faster matching times. For the experiments that will be presented in the sequel the values $N_S = 3000$, $R = 0.4 \cdot R_A$, $N_K = 500$ and $N_V = 1000$ have been selected for BoALD.

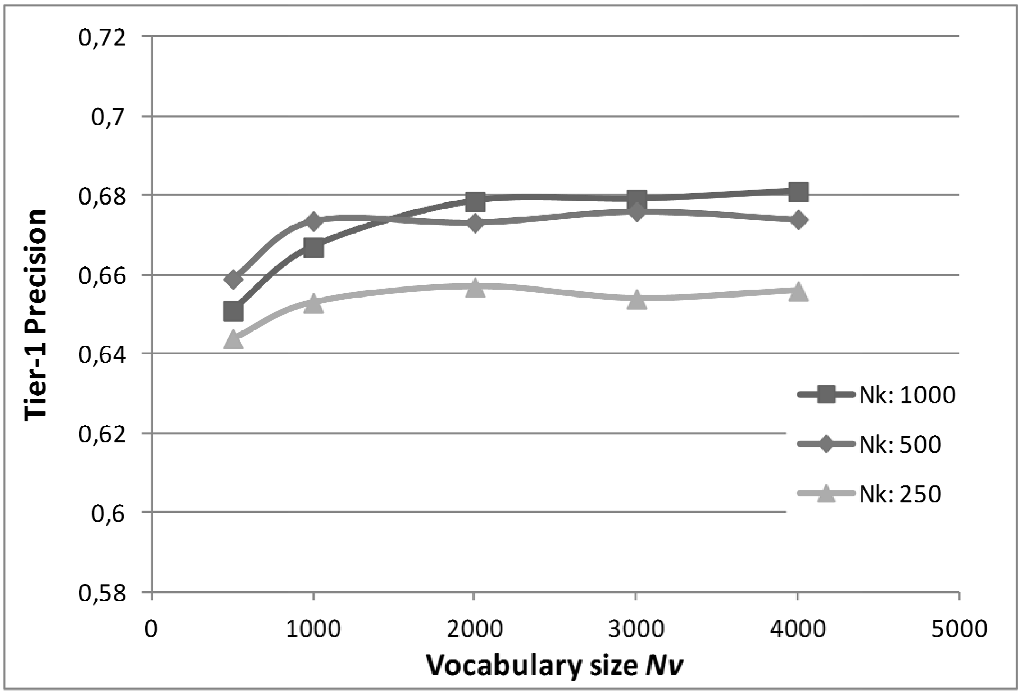


Figure 5.7: Parameter selection for BoALD descriptor: the average Tier-1 Precision in MolMovDB for different values of vocabulary size N_V and N_K.

The processing times for extraction of local features LDSP and HLG and for the BoALD bag-of-feature integration are given in Table 5.2. The codebook learning via *k*-means clustering is a computationally expensive process. For the MolMovDB dataset with 2695 molecules and $N_K = 500$ local features per molecule, the total number of training samples (10% of the dataset) is 134750 local features. The *k*-means clustering of 134750 features with vocabulary size $N_V = 1000$ took about 1700s (28 minutes). Then, the bag-of-features integration time for each molecule is 2.34s, thus, 6300s (105 minutes) for the entire database. These computations need to be performed only once, during the pre-processing stage.

Table 5.2: Average extraction times of the BoALD descriptor.

Number of	LDSP Descriptor	HLG Descriptor	BoALD bag-of-feature
1000	0.28	0.34	2.34
2000	0.42	1.66	
3000	0.75	3.08	
4000	0.95	8.74	

5.6.3 Performance Evaluation in MolMovDB – Flexible Similarity Matching

For performance evaluation in MolMovDB the *precision-recall* curve has been used, where precision is the proportion of the retrieved molecules that are relevant to the query and recall is the proportion of relevant molecules in the entire database that are retrieved. As it is described in Section 5.5.3, ALD is generated by concatenating a histogram of the LDSP descriptor with a histogram of the HLG descriptor. It is also stated that the substitution of the local LGF feature of [80] with the LDSP feature, which is based on the Surflet-Pair relations [105], produces an augmented descriptor for non-rigid shape matching with higher discriminative power. This combination of LDSP with HLG into the ALD descriptor is proposed here for the first time.

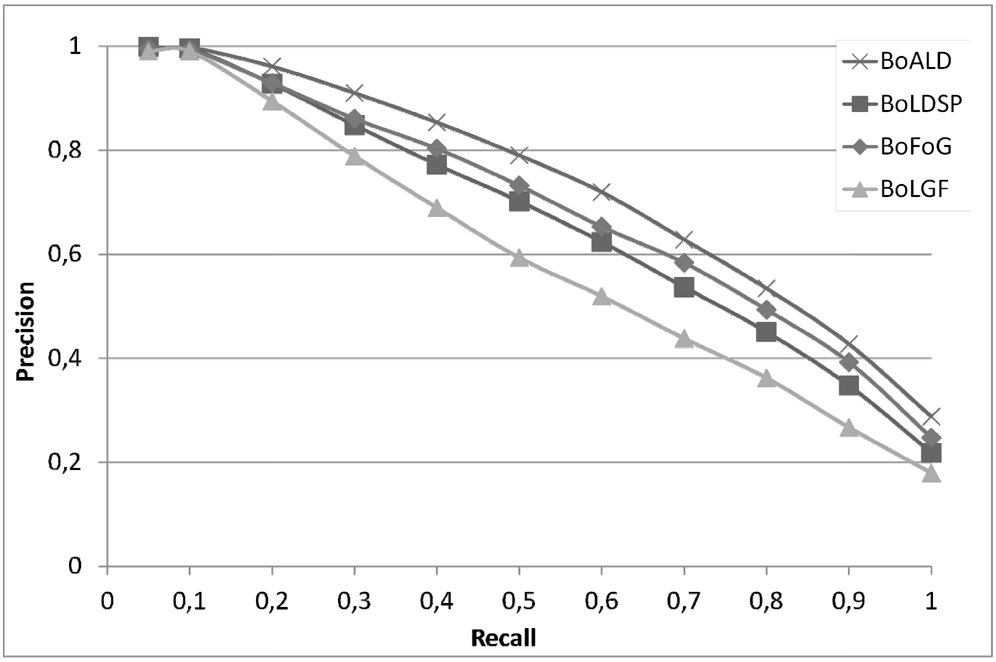


Figure 5.8: Comparison of BoLGF, BoFoG, BoLDSP and BoALD in MolMovDB.

The improvement in terms of accuracy is depicted in Figure 5.8. BoLGF is the bag-of-feature integration of the Local Geometrical Feature (LGF) that was used in [80], while its combination with the hybrid local-global feature is the BoFoG descriptor. In our case, BoLDSP is the bag-of-feature integration of the local-only feature (LDSP), while the combination of LSDP with the hybrid local-global feature (HLG) is our BoALD descriptor. It is clear that LDSP is a more accurate geometric feature than LGF, which in turn improves the precision of BoALD over BoFoG.

Another innovative feature of the proposed work is the modal representation of the diffusion distance matrix, which results in the DDMR descriptor. In Figure 5.9, DDMR is compared against the method of [73], which accumulates the pairwise diffusion distances into a histogram (DD-Hist). The proposed DDMR descriptor outperforms DD-Hist especially for higher values of recall. Finally, the combination of DDMR with BoALD, using the weighted sum of dissimilarities (5.14), is presented in Figure 5.9. DDMR-BoALD clearly outperforms the rest of descriptors, which confirms our assumption that the combination of a global feature (DDMR) with a local feature (BoALD) achieves higher retrieval accuracy than each descriptor separately.

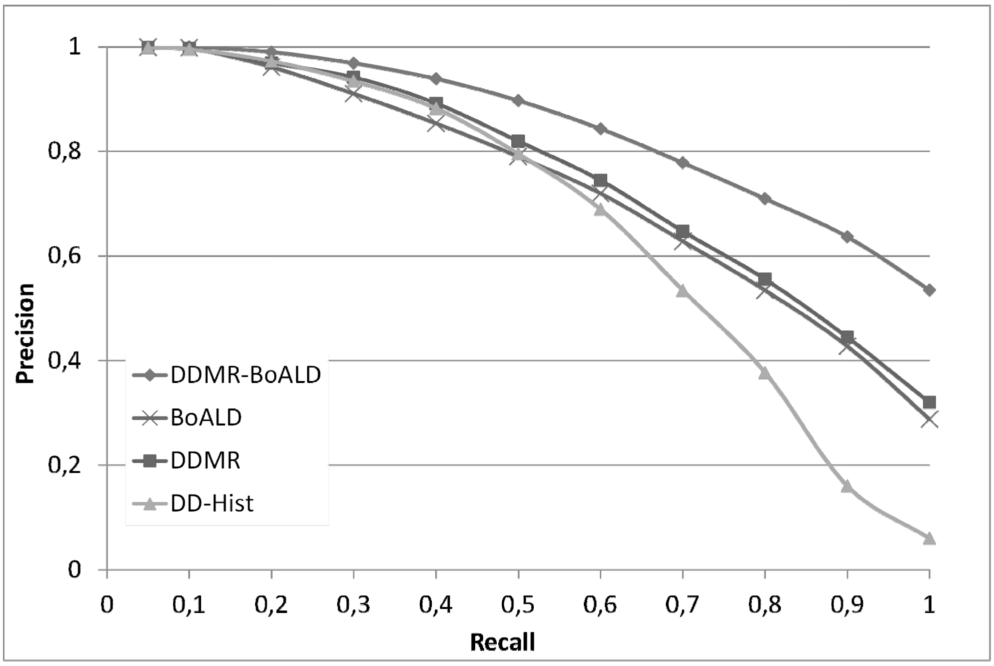


Figure 5.9: Comparison of DD-Hist, DDMR, BoALD and DDMR-BoALD in MolMovDB.

5.6.4 Evaluation of Rigid Similarity Matching

A shape descriptor that performs well in non-rigid similarity matching may not achieve high accuracy in rigid shape retrieval problems. As an example, the BoFoG method [80] outperforms rigid-shape-based approaches in a non-rigid database (MSB), while it has inferior performance in a rigid dataset (PSB). In this section, we prove that the proposed method is robust even to rigid shape retrieval problems.

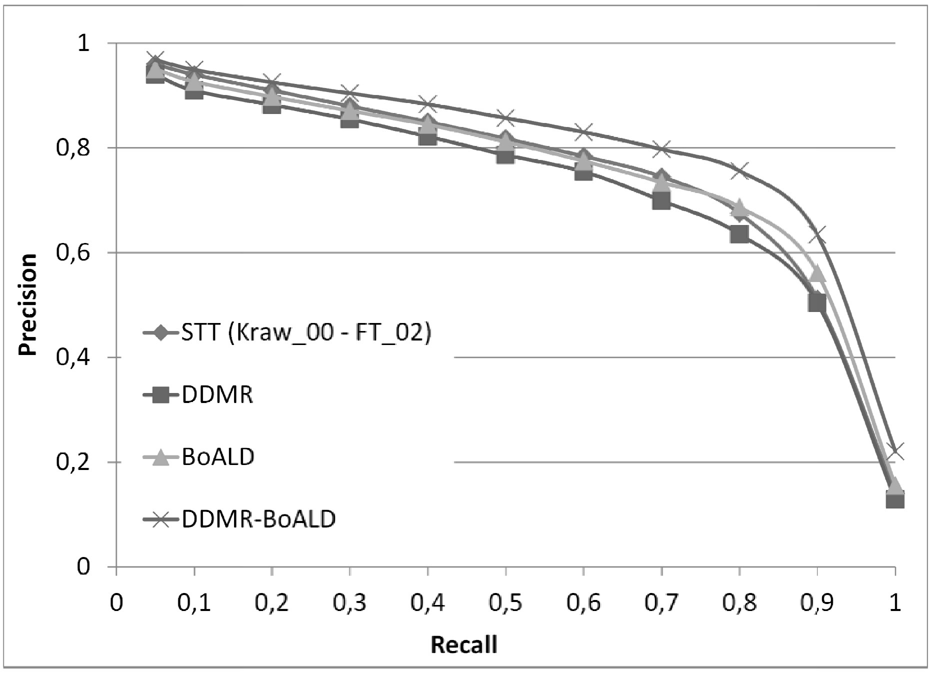


Figure 5.10: Comparison of the proposed method with STT in the subset of FSSP database that was used in [62].

In Figure 5.10, the precision-recall curves for the second dataset (subset of FSSP) are depicted. Our DDMR-BoALD descriptor is compared with STT [62], which is a rigid shape matching method. It is obvious that DDMR-BoALD outperforms STT in a rigid-shape dataset as well. This is mainly due to the fact that the combination of intrinsically different features (a global, a local and a hybrid local-global) increases the robustness of the resulting descriptor.

5.6.5 Virtual Screening of Ligands

The proposed method has been also evaluated in large-scale virtual screening of ligand molecules, where the investigation of an accurate algorithm for rapid shape matching is a major scientific challenge. Two benchmark datasets have been used in our tests. The first is called the "*Directory of Useful Decoys*" (DUD) [114]. DUD is derived from the ZINC database of commercially available compounds for virtual screening [115]. A subset of DUD[4] was downloaded, which consists of 13 targets and has been already used in recent studies [69]. The dataset is presented in Table 5.3. More specifically, each of the 13 *targets* is used as query to retrieve similar molecules from its corresponding set of *actives+decoys* (e.g. *ace* is used as query in the set of 46 actives and 1796 decoys and so on). The difference between decoys and actives is that the

[4] http://dud.docking.org/

former are presumed to be inactive against a target. The more *actives* are included among the first retrieved results the better the accuracy of the search algorithm is.

Table 5.3: The subset of DUD dataset that was used in our experiments.

Target	PDB	Actives	Decoys	Decoys per Active
angiotensin-converting enzyme (ace)	1o86	46	1796	39.04
acetylcholinesterase (ache)	1eve	100	3859	38.59
cyclin-dependent kinase 2(cdk2)	1ckp	47	2070	44.04
cyclooxygenase-2(cox2)	1cx2	212	12606	59.46
epidermal growth factor receptor(egfr)	1m17	365	15560	42.63
factor Xa(fxa)	1f0r	64	2092	32.69
HIV reverse transcriptase(hivrt)	1rt1	34	1494	43.94
enoyl ACP reductase(inha)	1p44	57	2707	47.49
P38 mitogen activated protein(p38)	1kv2	137	6779	49.48
phosphodiesterase(pde5)	1xp0	26	1698	65.31
platelet derived growth factor receptor kinase (pdgfrb)	1t46	124	5603	45.19
tyrosine kinase SRC(src)	2src	98	5679	57.95
vascular endothelial growth factor (vegfr2)	1fgi	48	2712	56.5

The second benchmark is the anti-HIV dataset derived from the National Cancer Institute[5] (NCI) and is employed to simulate a typical virtual screening experiment. It consists of 42687 compounds [116], which are split into 423 confirmed actives, 1081 moderately actives and 41185 confirmed inactives. The structures are available for download[6] in SDF format. The objective of the virtual screening experiment in this dataset is to use the 1081 moderately actives as queries and search into the database of actives and inactives. The more confirmed actives are retrieved among the first ranked results the higher the accuracy of the search algorithm is.

Three different metrics have been used to evaluate the performance of the proposed method in these datasets. The first is the *Enrichment Factor (EF) [117]*, which describes the ratio of actives retrieved relative to the percentage of the database scanned:

$$EF^{x} = \frac{N_a / N_x}{T_A / T_D} \tag{5.18}$$

[5] http://dtp.nci.nih.gov/docs/aids/aids_data.html

[6] http://ligand.info

where T_A is the total number of actives in the database of size T_D and N_a is the number of actives in the top x percent N_x of the database.

Another metric is the *Boltzmann Enhanced Discrimination of Receiver Operating Characteristic (BEDROC)* [118], calculated as:

$$BEDROC = \frac{\sum_{i=1}^{n} e^{-a \cdot \frac{r_i}{N}}}{\frac{n}{N}\left(\frac{1-e^{-a}}{e^{a/N}-1}\right)} \times \frac{R_a \sinh(a/2)}{\cosh(a/2)-\cosh(a/2-aR_a)} + \frac{1}{1-e^{a(1-R_a)}} \tag{5.19}$$

where n is the number of actives among N compounds, $R_a = n/N$, r_i is the rank of the i^{th} active and a is a weighting parameter. In our experiments, $a = 32.2$ is selected, which corresponds to $x = 5\%$ of the relative rank. Similarly, $x = 5\%$ is also selected for the EF metric (5.18).

Finally, the *Area Under Curve for Receiver Operator Characteristic (ROCAUC)* [69] is computed by:

$$AUCROC = 1 - \frac{1}{N_a}\sum_{i}^{N_a}\frac{N^i_{decoys}}{N_d} \tag{20}$$

where N_a and N_d is the number of actives and decoys, respectively, and N^i_{decoys} is the number of decoys ranked above the i^{th} active.

The proposed DDMR-BoALD descriptor is compared with two approaches for fast virtual screening, which are also based on shape similarity matching. The first one is the 3D Zernike Descriptor (3DZD) [69], which is based on a series expansion of a given 3D function. The second one is the Ultrafast Shape Recognition (USR) scheme [57], which represents the molecular shape as a set of statistical moments generated from all-atom distance distributions that are calculated with respect to preselected reference locations. Both aforementioned methods are rotation-invariant, i.e. are able to capture the shape information independent of orientation.

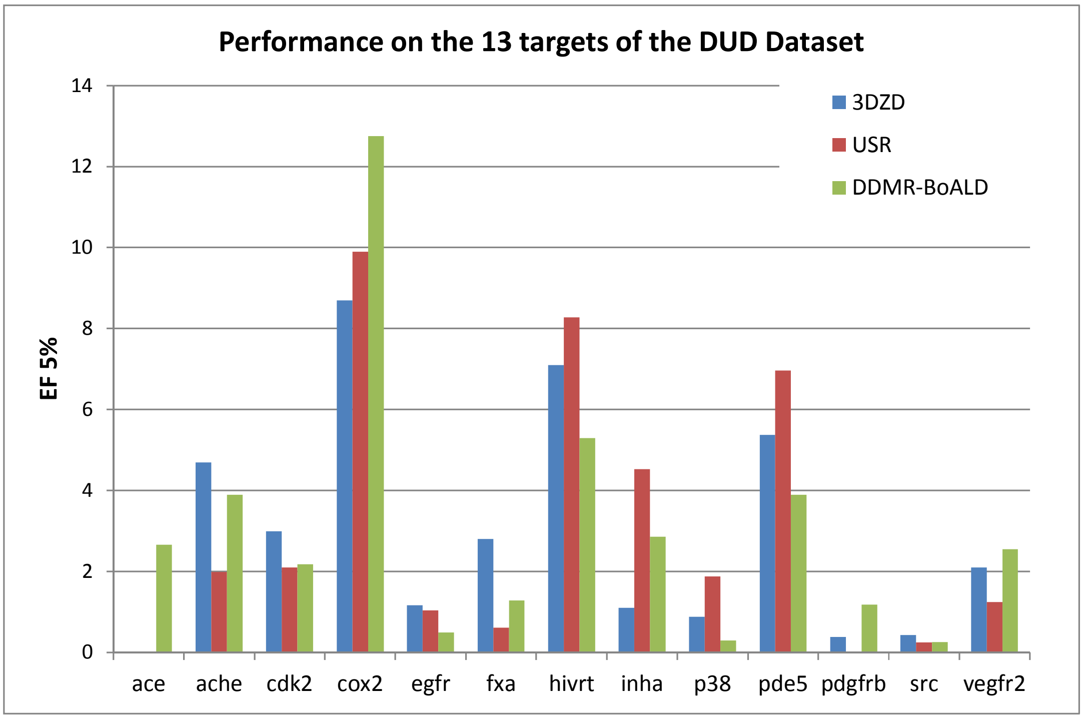


Figure 5.11: Performance of the 3DZD, USR and the proposed method on the 13 targets of the DUD dataset, using the Enrichment Factor metric.

In Figure 5.11, Figure 5.12 and Figure 5.13, the performance of 3DZD, USR and DDMR-BoALD on the 13 targets of the DUD dataset is given for the metrics EF ($x = 5\%$), AUCROC and BEDROC ($a = 32.2$), respectively. For 3DZD, the descriptor of order-12 using Correlation Coefficient as distance metric is reported, while for USR, the descriptor of order-16 using Correlation Coefficient as distance metric is reported [69].

Regarding the EF metric, the proposed method outperforms the other two in 4 out of 13 targets of the DUD Dataset, while 3DZD and USR are better in 5 and 4 targets, respectively. For the AUCROC metric, DDMR-BoALD is better in 5 targets, 3DZD in 3 and USR in 5. Finally, regarding the BEDROC metric, the proposed method outperforms others in 6 targets, 3DZD in 6 and USR in 1 target. The average scores are given in Table 5.4. The results derived using the 3 different metrics are not fully consistent, since e.g. USR is better than 3DZD in EF and AUCROC but it is worse in BEDROC. Overall, the proposed method is more robust since it outperforms the other two approaches in all metrics.

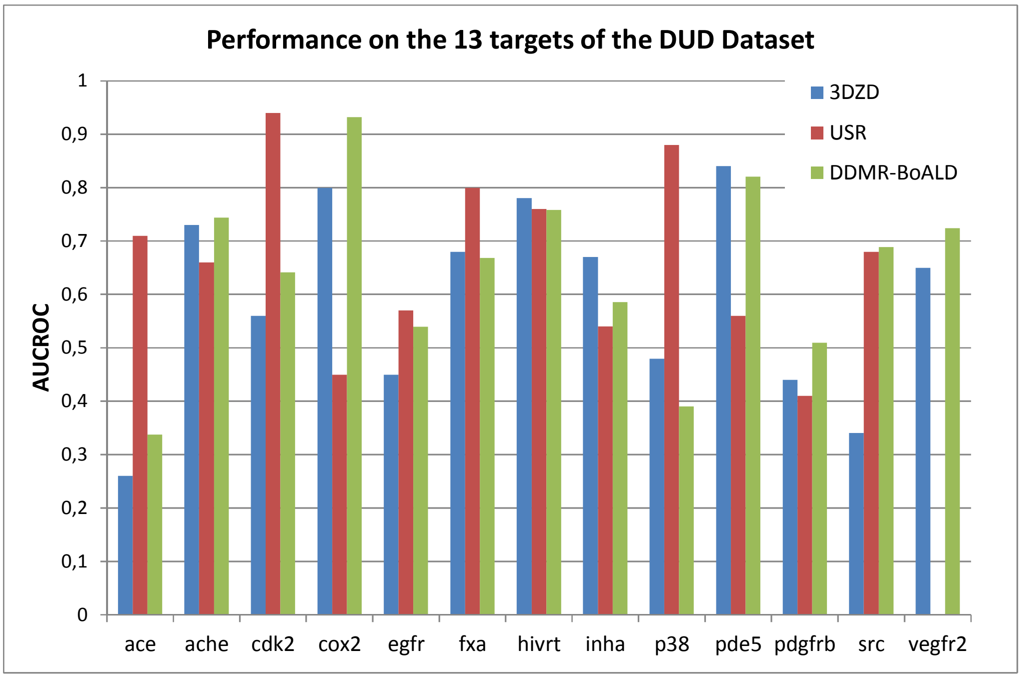


Figure 5.12: Performance of the 3DZD, USR and the proposed method on the 13 targets of the DUD dataset, using the AUCROC metric.

Table 5.4: Average values of EF, AUCROC and BEDROC in the DUD dataset for 3DZD, USR and the proposed method.

Descriptors	Metric	Order	EF 5%	AUCROC	BEDROC 32.2
3DZD	Correlation	12	2.90	0.59	0.14
USR	Correlation	16	2.99	0.62	0.12
DDMR-BoALD	-	-	**3.05**	**0.64**	**0.16**

The performance of 3DZD, USR and DDMR-BoALD is also compared in the anti-HIV dataset. In Table 5.5, the average values of EF ($x = 5\%$), AUCROC and BEDROC ($\alpha = 32.2$), for the three methods, are presented. Several results are available for both 3DZD and USR depending on the order of expansion of descriptor and the distance metric used. Again, the proposed method outperforms others in all three evaluation metrics.

Table 5.5: Average values of EF, AUCROC and BEDROC in the anti-HIV dataset for 3DZD, USR and the proposed method.

Descriptors	Metric	Order	EF 5%	AUC ROC	BEDROC
3DZD	Correlation coefficient	4	1.298	0.421	0.0485
		6	1.334	0.423	0.0500
		8	1.297	0.430	0.0490
		10	1.208	0.435	0.0461
		12	1.297	0.430	0.0490

		14	1.146	0.444	0.0440
	Euclidean (DE)	4	1.307	0.411	0.0471
		6	1.292	0.416	0.0473
		8	1.301	0.427	0.0477
		10	1.255	0.435	0.0464
		12	1.301	0.427	0.0477
		14	1.263	0.455	0.0470
	Manhattan (DM)	4	1.281	0.412	0.0466
		6	1.267	0.418	0.0463
		8	1.250	0.431	0.0462
		10	1.201	0.442	0.0448
		12	1.251	0.431	0.0462
		14	1.222	0.463	0.0461
USR	Correlation coefficient	12	1.248	0.417	0.0461
		16	1.357	0.422	0.0480
	Euclidean (DE)	12	1.301	0.392	0.0486
		16	1.296	0.386	0.0485
	Manhattan (DM)	12	1.403	0.395	0.0515
		16	1.335	0.386	0.0497
DDMR-BoALD	-	-	**1.923**	**0.479**	**0.0521**

A critical parameter that should be taken into account in virtual screening, especially in large databases, is the similarity matching time. In the anti-HIV dataset, which consists of more than 40000 molecules, the search times for USR are approximately 0.74-0.76s, while for 3DZD are 2.62-2.70s. These methods are significantly faster than non-shape-based approaches, which may take several hours for the same virtual screening task. The reason is that the shape-based descriptor vectors constitute a very compact representation of the molecular structure, thus, similarity matching using a common distance metric is rapid. The proposed DDMR-BoALD descriptor takes about 2.83s for a one-to-all matching in the anti-HIV dataset, thus, it is comparable to 3DZD. Consequently, since DDMR-BoALD outperforms 3DZD and USR in terms of retrieval accuracy, it can provide a better solution for rapid geometric virtual screening.

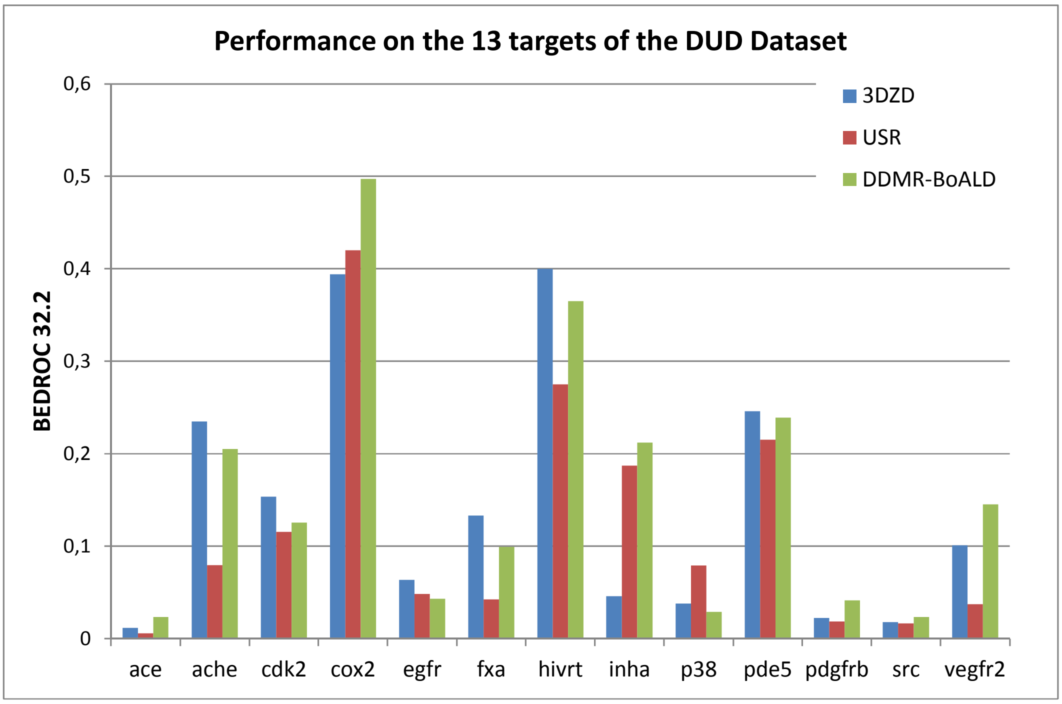


Figure 5.13: Performance of the 3DZD, USR and the proposed method on the 13 targets of the DUD dataset, using the BEDROC metric.

5.7 Summary

We have presented a framework for similarity search of flexible molecules, which exploits both local and global geometric features. The global feature is based on pairwise computations of diffusion distances over the points of the surface and a singular value decomposition of the resulting diffusion distance matrix. The local feature is computed on each keypoint of the surface by accumulating pairwise relations among oriented surface points into a local histogram. Finally, the hybrid local-global feature is computed for each keypoint, taking into account the diffusion distances from the keypoint to all surface points, thus, enhancing the local keypoint with spatial context. The local and the hybrid features are concatenated into a joint histogram per keypoint and the multiple histograms are integrated into a global descriptor using the bag-of-features approach. The global and local features are combined to produce a geometric descriptor that achieves higher retrieval accuracy than each feature does separately.

The proposed method achieves high retrieval accuracy in similarity search of flexible molecules. Experiments in the MolMovDB dataset, which consists of molecules with large conformational changes, demonstrate the superiority of the framework over existing approaches. At the same time, the proposed DDMR-BoALD descriptor achieves high retrieval performance in datasets of rigid molecules. Additionally, DDMR-BoALD provides a compact representation of the 3D molecular structure; therefore, it is appropriate for large-scale search tasks such as the virtual screening in large ligand databases. It is worth mentioning that DDMR-BoALD outperforms existing state-of-the-art approaches in two benchmarks for virtual screening.

Chapter 6

Conclusions and Future Work

In this dissertation, we aimed to address the problems of protein similarity search and protein-protein docking by exploiting 3D shape matching approaches. Regarding protein-protein docking, we took into account the fact that when two proteins interact, their surfaces at the binding site demonstrate geometric complementarity (apart from physicochemical complementarity). Two methods for protein docking have been introduced. The first is based on geometric complementarity only and is appropriate for bound docking problems. The second method advances the previous one in several aspects: a) it offers a more approximate similarity matching algorithm, which makes it appropriate even for unbound docking, and b) it combines physicochemical complementarity with geometric complementarity into a unified scoring function, which produces a more accurate scoring of the candidate poses. Protein similarity search is achieved by introducing a novel descriptor for shape matching of the global molecular shape. The descriptor is compact and it can be applied to both rigid and non-rigid molecular shape similarity tasks.

More specifically, the first protein docking method reveals various innovative features. The most significant one is that it introduces a shape similarity descriptor to measure surface complementarity. Since there is a wide variety of algorithms for similarity shape matching, it is easier to develop a method for partial surface complementarity by appropriately modifying a shape matching technique. Another interesting feature is the rotation invariance of SID descriptor. This obviates the need for an exhaustive search of relative orientations, during the pairwise complementarity matching of ESPs.

Although it outperforms the other geometric docking approaches, in several cases, the proposed method failed to return a hit within the first ranked positions, for two reasons: the implementation of the final scoring step was based on the notion that the bigger the area of the interface between two proteins the more probable is to be the actual docking area. The second reason is that no consideration of non-geometric factors (electrostatics, hydrogen bonds, residue interface propensity, etc.) was taken into account. An efficient scoring function able to integrate all non-geometric factors with geometric complementarity is of significant importance.

The second protein docking method that is proposed in this dissertation aims to overcome some of the drawbacks of the previous method. The so-called SP-Dock advances the state of the art mainly in the parts of local surface complementarity matching and alignment. The local spectral descriptor provides a more robust measure for shape complementarity of local patches, while the new grouping algorithm enhances the certainty of a wider surface region of receptor to be complementary to a wider surface region of the ligand. Additionally, instead of superimposing the sparse points of the ligand on the matching points of the receptor, as it is the case with most of the existing local-patch-based docking approaches, the ICP algorithm used by SP-Dock achieves alignment of the two proteins by taking into account the overall shape of the complementary regions. While this feature provides less accurate poses (i.e., with higher iRMSD) in the bound case, it significantly improves the unbound case. The reason is that the surfaces of the two proteins at their binding interfaces have approximate complementarity in the unbound case. Thus, a method based on exact matching and alignment would probably fail to retrieve a near-native pose within the list of predicted poses, while a more approximate method, such as SP-Dock, is more likely to achieve a correct prediction. This is an interesting conclusion and could assist in further research in protein-protein docking by proposing ideas on how to deal with unbound docking and slight side-chain flexibility.

Another advancement of SP-Dock is the new scoring process based on geometric and physicochemical factors. Several works have been presented so far dealing with the assessment of physicochemical factors but only few of them address both geometric and physicochemical complementarity. The proposed scoring of SP-Dock can be used as a starting point for further research, where the effect of additional factors, apart from Atom Desolvation Energy, Electrostatic Complementarity, Hydrophobicity, Coulomb Potential and van der Waals Potential, could be investigated.

Results performed on the 84 complexes of the Docking Benchmark 2.4 demonstrate the superiority of the proposed SP-Dock method over five similar docking approaches. While in the case of bound complexes, our method performs slightly better than the best docking methods reported so far, in the unbound case, our approach clearly outperforms them. This confirms the assumption that shape complementarity should be approximate (not exact) to take into account small side-chain conformations on the protein surface. Additionally, when several physicochemical factors are introduced (SP-Dock), the shape-only docking predictions are improved in both bound and unbound cases. Despite the improvements of the proposed SP-Dock method presented above and the interesting conclusions regarding the protein-protein docking problem, there is still a lot of work to be done in this direction. In terms of accuracy, research should focus on the following two goals: 1) to appropriately model the contribution of each factor (geometric or nongeometric) to protein interactions; 2) to appropriately model the flexibility (both side-chain and backbone) of the interacting proteins. Existing methods have reached an acceptable level in terms of computation time, though not adequately modeling the flexibility. If a deeper analysis of the flexibility takes place, then the computational time increases prohibitively. This tradeoff between accuracy and computation time should be considered, until a method that will address both problems is proposed.

Regarding our last method that has been introduced for molecular shape matching, it has been proven that it is quite efficient in both rigid and flexible shape comparison and, due to its compactness, it can provide a useful tool for virtual screening in large ligand databases. Nevertheless, the retrieval accuracy especially in virtual screening can be further improved, by enhancing the geometric features with non-geometric ones, such as physicochemical properties. At the moment, the latter are exploited by approaches that are extremely time-consuming, which, in combination with the rapid increase in size of the molecular databases, leads to prohibitively large search times. The effective integration of non-geometric information into a compact representation along with the shape-based features still remains a challenge for future research.

Bibliography

[1] L. Kavraki, "Geometric Methods in Structural Computational Biology," *Connexions*, June 11, 2007, http://cnx.org/content/col10344/1.6/.

[2] G. C. Diamantidis, “Introduction to Biochemistry”, *University Studio Press*, 3rd edition, 2007, ISBN 978-960-12-1624-9.

[3] E. Fischer, "Einfluss der Configuration auf die Wirkung der Enzyme". *Ber. Dt. Chem.* Ges. 27 (3): 2985–93. doi:10.1002/cber.18940270364, 1894.

[4] A. Fahmy, G. Wagner, “TreeDock: A Tool for Protein Docking Based on Minimizing van der Waals Energies”, *J. Am. Chem. Soc.*, 124, 1241-1250, 2002.

[5] C. Dominguez, R. Boelens, A. M. Bonvin, “HADDOCK: a protein-protein docking approach based on biochemical or biophysical information”, *J Am Chem Soc.* Feb 19;125(7):1731-7, 2003.

[6] I. Moreira, S. Pedro, A. Fernandes, and M. J. Ramos. "Protein–protein docking dealing with the unknown." *Journal of computational chemistry* 31.2 (2010): 317-342.

[7] D. W. Richie, “Recent Progress and Future Directions in Protein-Protein Docking”, *Current Protein and Peptide Science*, 2008, 9, 1-15.

[8] A. Bender and R. C. Glen. "Molecular similarity: a key technique in molecular informatics." *Organic & biomolecular chemistry 2*, no. 22 (2004): 3204-3218.

[9] M.L. Connolly. “Solvent-accessible surfaces of proteins and nucleic acids”. *Science*, 221:709–713, 1983.

[10] M.F. Sanner, A.J. Olson, and J.-C. Spehner. “Fast and robust computation of molecular surfaces”. *In 11th ACM Symposium on Computational Geometry*, 1995.

[11] J. Liang, H. Edelsbrunner, P. Fu, P.V. Sudhakar, and S. Subramaniam “Analytical shape computation of macromolecules: I. molecular area and volume through alpha shape”, *Proteins: Structure, Function, and Genetics*, 33:1-17, 1998.

[12] C. M. Roth, B. L. Neal, A. M. Lenhoff, “Van der Waals interactions involving proteins”, *Biophys. J.* 1996, 70, 977-987.

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

[14] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A.A. Friesem, C. Aflalo, and I.A. Vakser. “Molecular Surface Recognition: Determination of Geometric Fit between Protein and their Ligands by Correlation Techniques”. *Proc. Natl. Acad. Sci*. USA, 89:2195–2199, 1992.

[15] H.J. Wolfson and I. Rigoutsos. "Geometric hashing: An overview". *IEEE Computational Science and Eng*., 11:263–278, 1997.

[16] J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl and D. Baker. "Protein–Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations", *Journal of Molecular Biology* Volume 331, Issue 1, 1 August 2003, Pages 281-299

[17] J.C. Camacho, D.W. Gatchell, S.R. Kimura, and S. Vajda. "Scoring docked conformations generated by rigid body protein–protein docking". *PROTEINS: Structure, Function and Genetics*, 40:525–537, 2000.

[18] R. Chen and Z Weng. "Docking unbound proteins using shape complementarity, desolvation, and electrostatics". *PROTEINS: Structure, Function and Genetics*, 47:281–294, 2002.

[19] H.A. Gabb, R.M. Jackson, and J.E. Sternberg. Modelling protein docking using shape complementarity, electrostatics, and biochemical information. J. Mol. Biol., 272:106–120, 1997.

[20] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A.A. Friesem, C. Aflalo, and I.A. Vakser. "Molecular Surface Recognition: Determination of Geometric Fit between Protein and their Ligands by Correlation Techniques". *Proc. Natl. Acad. Sci*. USA, 89:2195–2199, 1992.

[21] I.A. Vakser. "Protein docking for low resolution structures". *Protein Engineering*, 8:371–377, 1995.

[22] Carter, P., Lesk, V.A., Islam, S.A. and Sternberg, M.J.E. (2005) *Proteins: Struct. Func. Bioinf.*, 60, 281-288.

[23] Kozakov, D., Brenke, R., Comeau, S.R. and Vajda, S. (2006) *Proteins: Struct. Func. Bioinf.,* 65, 392-406.

[24] Eisenstein, M. and Katchalski-Katzir, E. (2004) *Comptes Rendus Biologies*, 327, 409-420.

[25] Ritchie, D.W. and Kemp, G.J.L. (2000) *Proteins: Struct. Func. Genet.,* 39(2) 178-194.

[26] R. Chen and Z. Weng. "ZDOCK: An initial-stage protein-docking algorithm". *Proteins: Structure, Function and Genetics*, 52:80–87, 2003.

[27] E.J. Gardiner, P. Willett, and P.J. Artymiuk. "Protein docking using a genetic algorithm". *PROTEINS: Structure, Function and Genetics*, 44:44–56, 2001.

[28] G. Jones, P. Willet, R. Glen, and Leach. A.R. "Development and validation of a genetic algorithm for flexible docking". *J. Mol. Biol*., 267:727–748, 1997.

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

[30] M.L. Connolly. "Shape complementarity at the hemoglobin α1β1 subunit interface." *Biopolymers*, 25:1229–1247, 1986.

[31] R. Norel, S. L. Lin, H.J. Wolfson, and R. Nussinov. “Shape complementarity at protein-protein interfaces”. *Biopolymers*, 34:933–940, 1994.

[32] R. Norel, S. L. Lin, H.J. Wolfson, and R. Nussinov. “Molecular surface complementarity at protein-protein interfaces: The critical role played by surface normals at well placed, sparse points in docking”. *J. Mol. Biol*., 252:263–273, 1995.

[33] D. Duhovny, R. Nussinov, and H. J. Wolfson. “Efficient unbound docking of rigid molecules”. *In 2’nd Workshop on Algorithms in Bioinformatics*, pages 185–200, 2002.

[34] Zujun Shentu, Mohammad Al Hasan, Chris Bystroff and Mohammad J. Zaki, “Context Shapes: Efficient Complementary Shape Matching for Protein-Protein Docking”. *Proteins: Structure, Function and Bioinformatics*, 70(3):1056-1073. February 2008.

[35] C. Bajaj, R. Chowdhury, V. Siddavanahalli, “F^{2}Dock: Fast Fourier Protein-Protein Docking”, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 8, No. 1, Jan/Feb 2011.

[36] V. Venkatraman, Y. Yang, L. Sael, D. Kihara, “Protein-protein docking using region-based 3D Zernike descriptors”, *BMC Bioinformatics*, 2009;10:407.

[37] S. Gu, P. Koehl, J. Hass, N. Amenta, “Surface-histogram: A new shape descriptor for protein-protein docking”, *Proteins 2012,* 80:221–238.

[38] Brian Pierce and ZhipingWeng. “ZRANK: Reranking Protein Docking Predictions With an Optimized Energy Function”, *PROTEINS: Structure, Function, and Bioinformatics*, 67:1078–1086 (2007).

[39] Tim Geppert, Ewgenij Proschak, Gisbert Schneider, “Protein-protein docking by shape-complementarity and property matching”, *Journal of Computational Chemistry,* Volume 31 Issue 9, Pages 1919 – 1928, 2010.

[40] S. Liang, C. Zhang, S. Liu and Y. Zhou, “Protein binding site prediction using an empirical scoring function”, *Nucleic Acids Research*, Vol. 34, No. 13, 3698–3707, 2006.

[41] J. Janin, K. Henrick, J. Moult, LT. Eyck, MJ. Sternberg, S. Vajda, I. Vakser, SJ. Wodak, “CAPRI: a Critical Assessment of PRedicted Interactions”, *Proteins* 2003;52:2–9.

[42] E. Mashiach, D. Schneidman-Duhovny, A. Peri, Y. Shavit, R. Nussinov and H. J. Wolfson, “An Integrated Suite of Fast Docking Algorithms”, *Proteins*. 2010 November 15; 78(15): 3197–3204.

[43] H. Hwang, T. Vreven, B. G. Pierce, J. Hung, and Z. Weng, “Performance of ZDOCK and ZRANK in CAPRI rounds 13–19”, *Proteins, Structure Function Bioinformatics*, Vol. 78, Issue 15, pages 3104–3110, 15 November 2010.

[44] D. Kihara and J. Skolnick, “The PDB is a covering set of small protein structures”, *Journal of Molecular Biology*, 334, 793-802, 2003.

[45] R. Kolodny, D. Petrey, and B. Honig, “Protein structure comparison: implications for the nature of ‘fold space’, and structure and function prediction”, *Curr. Opin. Struct. Biol.* 2006;16:393–398.

[46] I. N. Shindyalov, P. E. Bourne, “Protein structure alignment by incremental combinatorial extension (CE) of the optimal path”, *Protein Eng.* 1998;11:739–747.

[47] K. Mizuguchi and N. Gö, "Comparison of spatial arrangements of secondary structural elements in proteins", *Protein Engineering*, 8.4 (1995): 353-362.

[48] V. Venkatraman, L. Sael and D. Kihara. "Potential for protein surface shape analysis using spherical harmonics and 3D Zernike descriptors." Cell biochemistry and biophysics 54.1-3 (2009): 23-32.

[49] L. Holm and C. Sander. "Protein structure comparison by alignment of distance matrices." *Journal of molecular biology* 233.1 (1993): 123-138.

[50] M. R. Betancourt and J. Skolnick, "Local propensities and statistical potentials of backbone dihedral angles in proteins", *Journal of molecular biology* 342.2 (2004): 635-649.

[51] Y-S Liu, Q. Li, G-Q. Zheng, K. Ramani, W. Benjamin, “Using diffusion distances for flexible molecular shape comparison”, *BMC Bioinformatics 2010*, 11:480.

[52] K. Kinoshita and H. Nakamura, "Identification of protein biochemical functions by similarity search using the molecular surface database eF-site", *Protein Science* 12.8 (2003): 1589-1595.

[53] M. E. Bock, G. M. Cortelazzo, C. Ferrari and C. Guerra, “Identifying similar surface patches on proteins using a spin-image surface representation”, *In Combinatorial Pattern Matching*. Springer Berlin Heidelberg, 2005. p. 417-428.

[54] M. Ankerst, G. Kastenmüller, H. P. Kriegel and T. Seidl, "3D shape histograms for similarity search and classification in spatial databases." *Advances in Spatial Databases*. Springer Berlin Heidelberg, 1999.

[55] J. S. Yeh, D.Y. Chen, B.Y. Chen, M. Ouhyoung, “A web-based three-dimensional protein retrieval system by matching visual similarity”, *Bioinformatics 2005*, 21(13):3056-3057.

[56] P. Daras, A. Axenopoulos, "A Compact Multi-View Descriptor for 3D Object Retrieval" *IEEE 7th International Workshop on Content-Based Multimedia Indexing (CBMI 2009)*, Chania, Greece, Jun 2009.

[57] P. J. Ballester and W. G. Richards, "Ultrafast shape recognition to search compound databases for similar molecular shapes", *Journal of Computational Chemistry* 28.10 (2007): 1711-1723.

[58] P. J. Ballester and W. G. Richards, "Ultrafast shape recognition for similarity search in molecular databases", *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science* 463.2081 (2007): 1307-1321.

[59] M-K. Hu, "Visual pattern recognition by moment invariants", *IRE Transactions on Information Theory*, 8.2 (1962): 179-187.

[60] M. R. Teague, "Image analysis via the general theory of moments", *J. Opt. Soc.* Am 70.8 (1980): 920-930.

[61] C-H. Teh and R. T. Chin, "On image analysis by the methods of moments", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10.4 (1988): 496-513.

[62] P. Daras, D. Zarpalas, A. Axenopoulos, D. Tzovaras and M. G. Strintzis, "Three-dimensional shape-structure comparison method for protein classification", *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2006, 3(3), 193-207.

[63] W.Cai, J. Xu, X. Shao, V. Leroux, A. Beautrait and B. Maigret, "SHEF: a vHTS geometrical filter using coefficients of spherical harmonic molecular surfaces", Journal of molecular modeling, 2008, 14(5), 393-401.

[64] R. J. Morris, R. J. Najmanovich, A. Kahraman and J. M. Thornton, "Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons" Bioinformatics, 2005, 21(10), 2347-2355.

[65] D. W. Ritchie and G. J. Kemp. "Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces." Journal of Computational Chemistry 20.4: 383-395, 1999.

[66] D. W. Ritchie and G. J. Kemp, "Protein docking using spherical polar Fourier correlations." Proteins: Structure, Function, and Bioinformatics, 39.2 (2000): 178-194.

[67] L. Mak, S. Grandison and R. J. Morris, "An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison", Journal of Molecular Graphics and Modelling 26.7 (2008): 1035-1045.

[68] L. Sael, B. Li, D. La, Y. Fang, K. Ramani, R. Rustamov and D. Kihara, (2008). Fast protein tertiary structure retrieval based on global surface shape similarity. Proteins: Structure, Function, and Bioinformatics, 72(4), 1259-1273.

[69] V. Venkatraman, P. R. Chakravarthy and D. Kihara, "Application of 3D Zernike descriptors to shape-based ligand similarity searching", J Cheminform 17.1 (2009): 19.

[70] Y. Fang, Y-S Liu and K. Ramani, "Three dimensional shape comparison of flexible proteins using the local-diameter descriptor", BMC Structural Biology 2009, 9:29 doi:10.1186/1472-6807-9-29.

[71] Y-S Liu, Y. Fang and K. Ramani, “IDSS: deformation invariant signatures for molecular shape comparison”, BMC Bioinformatics 2009, 10:157 doi: 10.1186/1471-2105-10-157.

[72] Y-S Liu, K. Ramani and M. Liu, “Computing the Inner Distances of Volumetric Models for Articulated Shape Description with a Visibility Graph”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 12, December 2011.

[73] Y-S Liu, Q. Li, G-Q. Zheng, K. Ramani, W. Benjamin, “Using diffusion distances for flexible molecular shape comparison”, BMC Bioinformatics 2010, 11:480.

[74] S. Yin, E. A. Proctor, A. A. Lugovskoy, and N. V. Dokholyan, “Fast screening of protein surfaces using geometric invariant fingerprints”, PNAS 2009, vol. 106, no. 39, pp. 16622–16626, September 29, 2009.

[75] P. Heider, A. Pierre-Pierre, R. Li and C. Grimm, "Local shape descriptors, a survey and evaluation", In Proceedings of the 4th Eurographics conference on 3D Object Retrieval, pp. 49-56. Eurographics Association, 2011.

[76] Z. Lian, A. Godil, X. Sun, H. Zhang, “Non-rigid 3D shape retrieval using multidimensional scaling and bag-of-features”, in Proceedings of the International Conference on Image Processing (ICIP2010), 2010, pp.3181–3184.

[77] D. Smeets, T. Fabry, J. Hermans, D. Vandermeulen, P. Suetens, “Isometric deformation modeling for object recognition”, in Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns (CAIP’09), 2009, pp.757–765.

[78] G. Lavoue, “Bag of words and local spectral descriptor for 3D partial shape retrieval”, in Proceedings of the Eurographics Workshop on 3D Object Retrieval (3DOR’11), 2011, pp.41–48.

[79] R. Ohbuchi, K. Osada, T. Furuya, T. Banno, “Salient Local Visual Features for Shape-Based 3D Model Retrieval”, Proc. IEEE International Conference on Shape Modeling and Applications (SMI’08), Stony Brook University, June 4 - 6, 2008.

[80] S. Kawamura, K. Usui, T. Furuya, and R. Ohbuchi. "Local goemetrical feature with spatial context for shape-based 3D model retrieval." In Proceedings of the 5th Eurographics conference on 3D Object Retrieval, pp. 55-58. Eurographics Association, 2012.

[81] P. Daras, A. Axenopoulos, G. Litos, "Investigating the Effects of Multiple Factors towards more Accurate 3D Object Retrieval", IEEE Transactions on Multimedia, Vol. 14, No. 2, Page(s): 374 – 388, April 2012.

[82] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, “The Protein Data Bank,” *Nucleic Acids Research*, vol. 28, pp. 235-242, 2000.

[83] Renee L. DesJarlais, Robert P. Sheridan, George L. Seibel, J. Scott Dixon, Irwin D. Kuntz,and R. Venkataraghavan, “Using Shape Complementarity as an Initial Screen in Designing Ligands

for a Receptor Binding Site of Known Three-Dimensional Structure", *J. Med. Chem.* 1988,31, 722-729.

[84] Renee L. DesJarlais, Robert P. Sheridan, J. Scott Dixon, Irwin D. Kuntz,and R. Venkataraghavan, "Docking Flexible Ligands to Macromolecular Receptors by Molecular Shape", *J. Med. Chem.* 1986,29, 2149-2153.

[85] D. Fischer, S.L. Lin, H.L. Wolfson, and R. Nussinov. "A geometry-based suite of molecular docking processes". *J. Mol. Bio*., 248:459–477, 1995.

[86] A.Mademlis, P.Daras, D.Tzovaras and M.G.Strintzis, "3D Object Retrieval based on Resulting Fields" *29th International conference on EUROGRAPHICS 2008, workshop on 3D object retrieval*, Crete, Greece, Apr 2008.

[87] A.Mademlis, P.Daras, D.Tzovaras and M.G.Strintzis, "3D Object Retrieval using the 3D Shape Impact Descriptor" *ELSEVIER, Pattern Recognition*, Volume 42 , Issue 11, pp. 2447-2459, Nov 2009.

[88] Daras P., Zarpalas D., Tzovaras D., Strintzis M. G.: "Efficient 3-d model search and retrieval using generalized 3-d radon transforms". *IEEE Transactions on Multimedia* 8, 1 (2006), 101–114.

[89] H. Ling, K. Okada, "Diffusion distance for histogram comparison", *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 2006), IEEE Computer Society, pp. 246–253.

[90] J. Mintseris, K. Wiehe, B. Pierce, R. Anderson, R. Chen, J. Janin, and Z. Weng. "Protein-protein docking benchmark 2.0: An update". *Proteins: Structure, Function and Genetics*, 60(2):214–216, 2005.

[91] R. Chen, J. Mintseris, J. Janin, and Z. Weng. "A protein-protein docking benchmark". *Proteins: Structure, Function and Genetics*, 52(1):88–91, 2003.

[92] Z. Lian, A. Godil, B. Bustos, M. Daoudi, J. Hermans, S. Kawamura, Y. Kurita, G. Lavoué, H.V. Nguyen, R. Ohbuchi, Y. Ohkita, Y. Ohishi, F. Porikli, M. Reuter, I. Sipiran, D. Smeets, P. Suetens, H. Tabia, D. Vandermeulen, "SHREC'11 Track: Shape Retrieval on Non-rigid 3D Watertight Meshes", *Eurographics Workshop on 3D Object Retrieval (2011)*, April 10, 2011, Llandudno (UK).

[93] G. Lavoué, "Bag of Words and Local Spectral Descriptor for 3D Partial Shape Retrieval", *Eurographics Workshop on 3D Object Retrieval (2011)*, April 10, 2011, Llandudno (UK).

[94] H. Edelsbrunner, E. P. Miicke, "Three-dimensional alpha shapes", *Proceedings of the 1992 workshop on Volume visualization (VVS '92),*pp 75-82, NY, USA, 1992.

[95] B. Vallet, B. Levy, "Spectral geometry processing with manifold harmonics", *Computer Graphics Forum* 27, 2 (2008), 251–260.

[96] M. Reuter, F.-E. Wolter and N. Peinecke, “Laplace-Beltrami spectra as "Shape-DNA" of surfaces and solids”, *Computer-Aided Design* 38 (4), pp.342-366, 2006.

[97] P.J. Besl and N.D. McKay, Reconstruction of Real-World Objects via Simultaneous Registration and Robust Combination of Multiple Range Images, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 14:2(239-256), 1992.

[98] C. Zhang, G. Vasmatzis, J. L. Cornette, C. DeLisi, “Determination of atomic desolvation energies from the structures of crystallized proteins”, J. Mol Biol, 1997, 267, 707.

[99] F. Glaser, D.M. Steinberg, I.A. Vakser, N. Ben-Tal, “Residue frequencies and pairing preferences at protein-protein interfaces”, Proteins, 2001 May 1;43(2):89-102.

[100] M. Berrera, H. Molinari, F. Fogolari, “Amino acid empirical contact energy definitions for fold recognition in the space of contact maps”, BMC Bioinformatics 2003, 4:8.

[101] G. Ausiello, G. Cesareni, M. Helmer-Citterich, “ESCHER: a new docking procedure applied to the reconstruction of protein tertiary structure”, Proteins 1997, 28(4), 556-567.

[102] N. P. Palma, L. Krippahl, J. E. Wampler, J. J. G. Moura, “BiGGER: A new (soft) docking algorithm for predicting protein interactions”, Proteins: Structure, Function, and Bioinformatics, 2000, 39, 372.

[103] R. Norel, D. Petrey, H.J. Wolfson, R. Nussinov, “Examination of shape complementarity in docking of unbound proteins”, Proteins 1999, 36, 307.

[104] Particle Swarm Optimization, Available Online: http://www.swarmintelligence.org/tutorials.php.

[105] E. Wahl, U. Hillenbrand and G. Hirzinger, "Surflet-pair-relation histograms: a statistical 3D-shape representation for rapid classification", *IEEE Fourth International Conference on 3-D Digital Imaging and Modeling*, 3DIM 2003.

[106] N. Echols, D. Milburn and M. Gerstein, “MolMovDB: Analysis and visualization of conformational change and structural flexibility”, *Nucleic Acids Research 2003*, 31:478-482.

[107] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, “Numerical recipes in C+: the art of scientific computing”, *Cambridge: Cambridge University Press*, Vol. 994, 2009.

[108] R. Osada, T. Funkhouser, B. Chazelle and D. Dobkin, “Shape distributions”, *ACM Transactions on Graphics,* 2002, 21(4):807-832.

[109] B. Nadler, S. Lafon, R. R. Coifman, I. G. Kevrekidis, ”Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators”, *in Advances in Neural Information Processing Systems* 18, 2005.

[110] D. Zhou, O. Bousquet, T. N. Lal, J. Weston and B. Schölkopf, “Learning with Local and Global Consistency”, *NIPS 2003*.

[111] S. C. Flores, L. J. Lu, J. Yang, N. Carriero, and M. B. Gerstein, “Hinge Atlas: relating protein sequence to sites of structural flexibility”, *BMC Bioinformatics* 8: 167, 2007.

[112] L. Holm and C. Sander, “The FSSP Database: Fold Classification Based on Structure-Structure Alignment of Proteins,” Nucleic Acids Research, vol. 24, pp. 206-210, 1996.

[113] L. Holm and C. Sander, “Touring Protein Fold Space with Dali/FSSP,” Nucleic Acids Research, vol. 26, pp. 316-319, 1998.

[114] N. Huang, B. K. Shoichet, J. J. Irwin, “Benchmarking sets for molecular docking”, J. Med. Chem. 2006, 49:6789-6801.

[115] J.J. Irwin, B. K. Shoichet, “ZINC–a free database of commercially available compounds for virtual screening”, J. Chem. Inf. Model 2005, 45:177-182.

[116] O.S. Weislow, R. Kiser, D. L. Fine, J. Bader, R. H. Shoemaker, M. R. Boyd, “New soluble-formazan assay for HIV-1 cytopathic effects: application to highflux screening of synthetic and natural products for AIDS-antiviral activity”, J. Natl. Cancer. Inst. 1989, 81:577-586.

[117] A. Bender, R. C. Glen, “A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication”, J. Chem. Inf. Model 2005, 45:1369-1375.

[118] J. F. Truchon, C. I. Bayly, “Evaluating virtual screening methods: good and bad metrics for the “early recognition” problem”, J. Chem. Inf. Model 2007, 47:488-508.

[119] S. Malassiotis, M. G. Strintzis, "*Snapshots: A Novel Local Surface Descriptor and Matching Algorithm for Robust 3D Surface Alignment*" IEEE PAMI, vol.29, no. 7, pp. 1285-1290, Jul 2007.

[120] R. T. Kroemer, "Structure-based drug design: docking and scoring", *Curr. Protein Pept. Sci.* 8 (4): 312–28. doi:10.2174/138920307781369382, 2007.

[121] P. J. Ballester, I. Westwood, N. Laurieri, E. Sim, W. G. Richards, "Prospective virtual screening with Ultrafast Shape Recognition: the identification of novel inhibitors of arylamine N-acetyltransferases". *Journal of The Royal Society Interface* 7: 335–342, 2010.