A Shape Descriptor for Fast Complementarity Matching in Molecular Docking

Apostolos Axenopoulos , Petros Daras, *Member*, IEEE, Georgios Papadopoulos, and Elias Houstis

Abstract— This paper 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.

Index Terms — protein docking, rigid body, geometric complementarity, shape impact descriptor.

———————— Φ ————————

1 INTRODUCTION

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. The study of protein interactions may involve experimental approaches like Co-immunoprecipitation, BiFC, In-vivo crosslinking, DPI, FCS, crystallography, etc. as well as computational approaches like Protein-protein docking, binding site prediction, protein interaction networks, etc. While determining the existence or not of an interaction can be easily carried out experimentally, the same is not possible yet for the accurate prediction of the binding interface, unless crystallography is applied. Therefore, computational approaches of protein-ligand docking are very popular to pharmaceutical companies, providing an important tool in computer-assisted drug design.

The problem of molecular docking involves prediction of a ligand conformation and orientation, also known as pose, within the active site of a receptor. The stability of a pose is a result of the so-called "weak interactions" (Coulomb forces, hydrogen bonds, Van der Waals forces, hydrophobic interactions). However, apart from the physicochemical complementarity, geometric complementarity is not underestimated and it is taken into consideration in several docking algorithms.

A. Axenopoulos and E. Houstis are with the Department of Computer & Communication Engineering, University of Thessaly, Volos, Greece (e-mail: axenop@iti.gr; enh@inf.uth.gr)
P. Daras is with the Informatics & Telematics Institute, Centre for Research & Technology Hellas, Thermi-Thessaloniki, Greece (e-mail: daras@iti.gr)
G. Papadopoulos is with the Department of Biochemistry & Biotechnology, University of Thessaly, Larissa, Greece (e-mail: geopap@med.uth.gr)

1.1 Related Work

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 [1]. A wide spectrum of algorithms including Fast Fourier Transform (FFT) correlations [5], geometric hashing [16], and Monte Carlo (MC) [32] techniques has been utilized in current docking algorithms. The aim in all of the above approaches is to produce a set of candidate docking poses, among which a near-native binding mode is often observed. 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 [19], [33].

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 [2], [3], [4] 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 [5]. 3D FFT has been incorporated in several correlation-based docking algorithms [6], [7], [8]. A recent overview of the principles of grid-based FFT docking approaches is given in [9]. In [10], 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 [19] 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 [20], [21].

Local shape feature matching approaches usually require a representation of the molecular surface, attempting to find regions of interest on the surface. Then, they apply pairwise complementarity matching of these regions between the receptor and the ligand. One of the first docking approaches, based on local shape feature matching, was introduced in 1982 [11]. In [12], a method to match local curvature maxima and minima points was presented. This technique has been extended in [13], [14]. In [15], a method based on geometric hashing [16] is presented. Each protein surface is first preprocessed 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 [17]. These are boolean data structures and correspond to significantly large parts of the protein surface. Complementarity shape matching is achieved using efficient boolean operations. 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 an attempt to deal with the limitations of the above mentioned local shape feature matching methods, the proposed approach provides a fast solution, while being at the same time efficient in terms of complementarity shape matching.

More recent docking approaches aim to combine geometric and physicochemical information in order to produce more accurate solutions. In [35], geometric complementarity matching, achieved by geometric hashing, and several knowledge - based potentials, including electrostatics, desolvation, residue contact preferences and Van-derWaals potential, are efficiently merged, demonstrating remarkable results in a test set of 68 bound and 30 unbound test cases. The most important conclusion that can be drawn by this study is that none of the two aforementioned factors, geometric and physicochemical complementarity, should be underestimated, but the enhancement of a geometric method with additional non-geometric properties, during the soring phase, can successfully discard false positive predictions and improve the results of the algorithm.

1.2 Method Overview and Contributions

The proposed method can be summarized as in the block diagram presented in Fig. 1. The input is the PDB [22] 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.

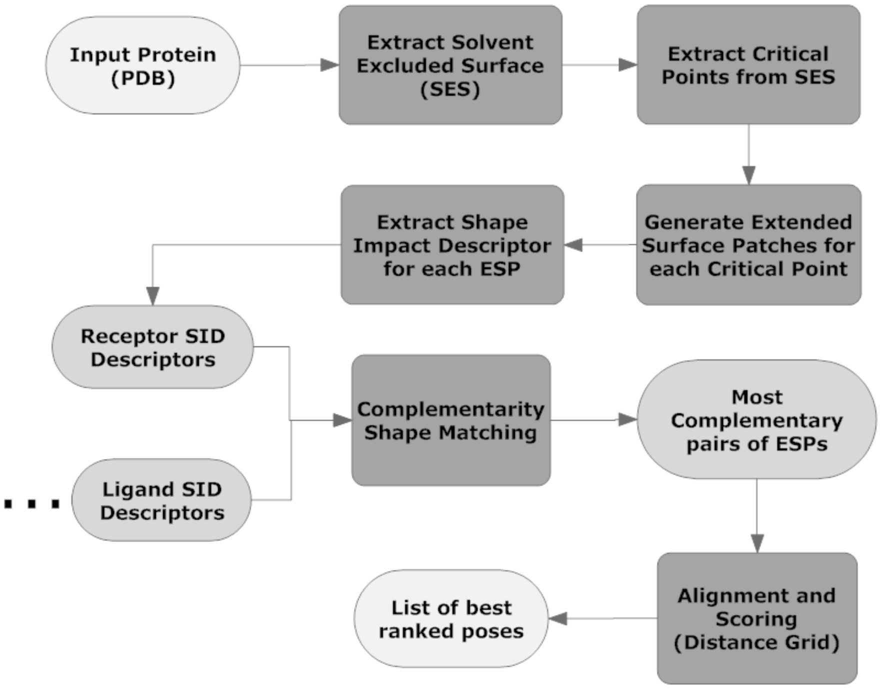
Input Protein (PDB)
Extract Solvent Excluded Surface (SES)
Extract Critical Points from SES
Generate Extended Surface Patches for each Critical Point
Extract Shape Impact Descriptor for each ESP
Receptor SID Descriptors
Ligand SID Descriptors
...
Complementarity Shape Matching
Most Complementary pairs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

Fig. 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 [36], 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 [37], differs from the proposed method in the following: the method presented in [37] 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 [17]. The concept of pairwise complementarity matching of equally sized surface patches is common to both approaches; however, the method presented in this paper introduces several innovative features. First of all, in [17], the authors adopt the method in [14] in order to generate an initial set of sparse critical points, while, in this paper, 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 [17], 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 [15], in order to produce more accurate results.

The rest of the paper is organized as follows: in Section 2 a new approach for extraction of critical points from the molecular surface is introduced. In Section 3, a description of the Shape Impact Descriptor is given, while in Section 4, the final step of the algorithm, which includes alignment and geometric scoring is presented. Then, in Section 5, the experimental results are presented, where the proposed method is compared with other state-of-the-art approaches. Finally, conclusions are drawn in Section 6.

2 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) [23] 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) [24] 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 [25]. 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.

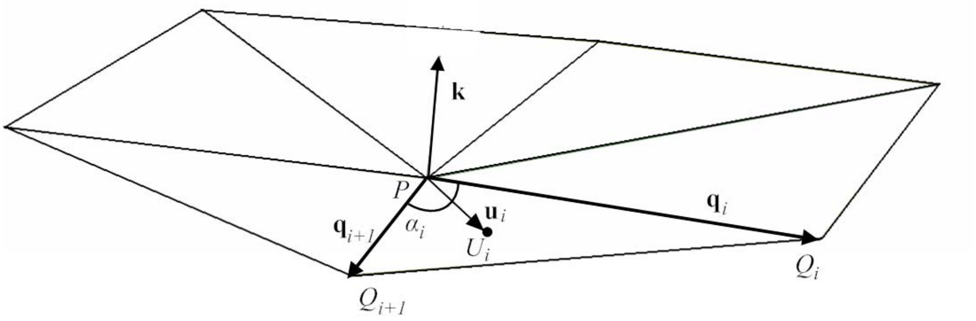

Fig. 2. Estimation of the local curvature around a point *P*

In this paper, a method for generating critical points based on the local curvature of the surface is introduced. The reason for not adopting the sparse surface [25] 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 [25] 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 (Fig. 2):

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

where N is the total number of neighboring points Q_i of P, $\mathbf{u}_i$ 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:

$$\alpha_i = \arccos\left(\frac{\mathbf{q}_i \cdot \mathbf{q}_{i+1}}{\left|\mathbf{q}_i\right|\left|\mathbf{q}_{i+1}\right|}\right) \tag{2}$$

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

For surface points P that belong to convex areas, their corresponding vectors $\mathbf{k}$ 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 (Fig. 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 Fig. 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.

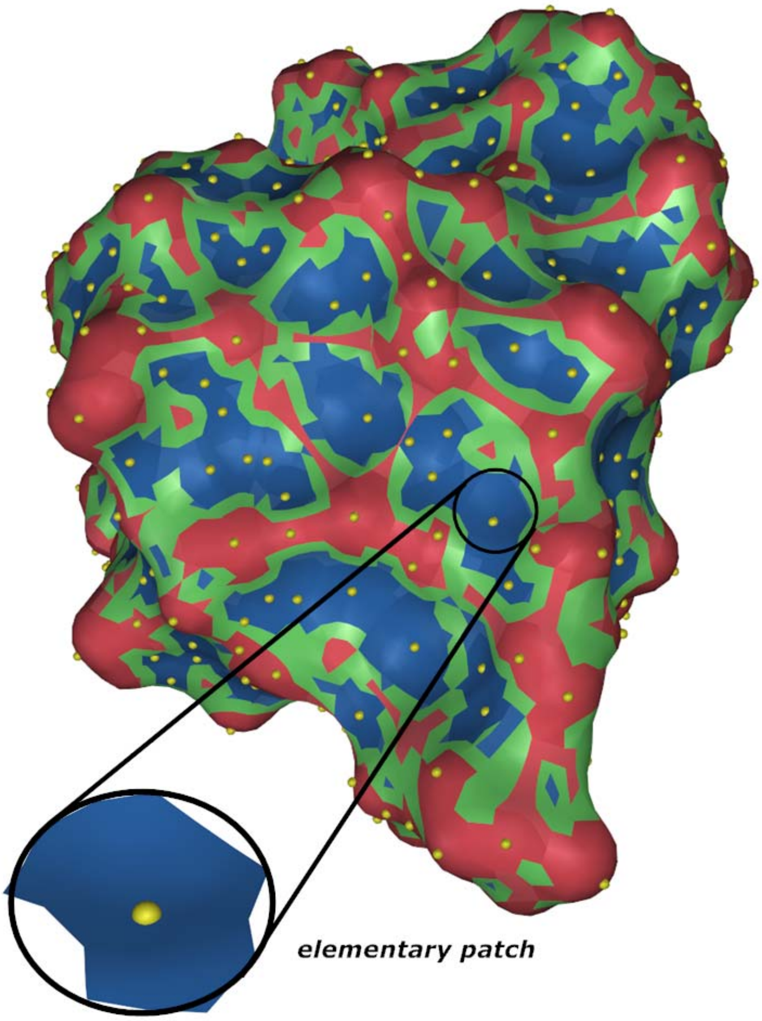

Fig. 3. Segmentation of SES into convex, concave and flat regions. The critical points are represented by yellow dots.

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 (Fig. 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 Fig. 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 Fig. 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 (Fig. 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.

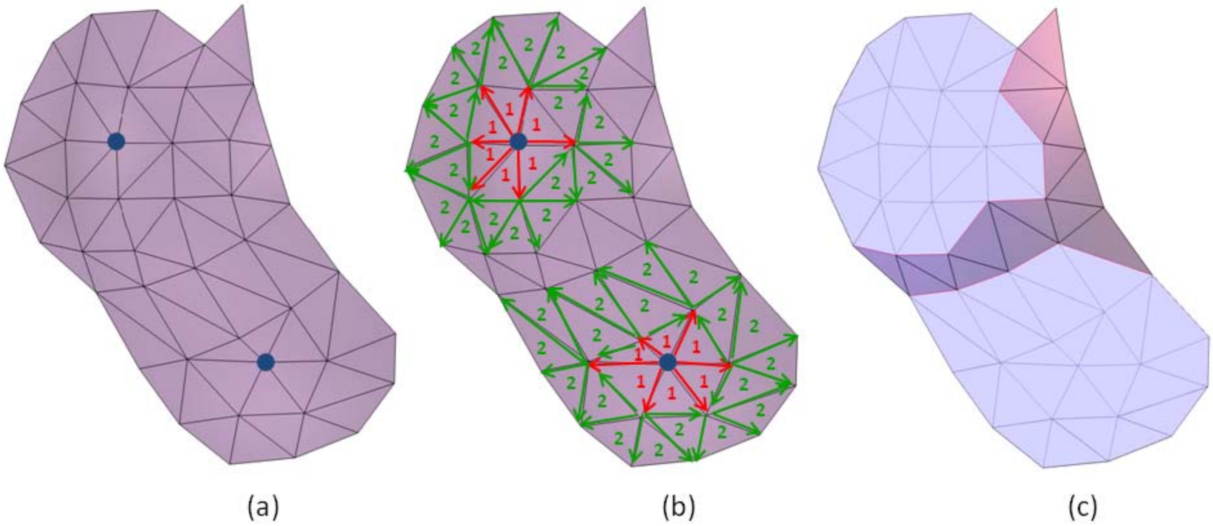

Fig. 4. The steps for segmenting a continuous region of surface points of the same type: (a) select the most distant points from the region contour as seed points (b) expand uniformly to all directions until the region contour is reached; the numbers represent the level of expansion around the seed point (c) group all surface points covered by the expansion around each seed point; these sets of points constitute the elementary patches.

In Fig. 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.

3 The Shape Impact Descriptor

The idea of local shape complementarity matching in this paper is similar to the one presented in [17]. 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.

3.1 Preprocessing

Given the SES of a protein along with the set of critical points described in Section 2, 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 Fig. 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 IV.

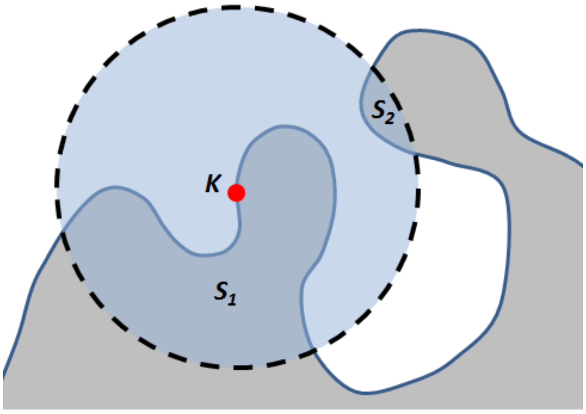

Fig. 5. Removal of unconnected surface parts using geodesic distance: taking into account only Euclidean distances from the center *K* of the ESP, both S_1 and S_2 surface parts are enclosed. However, points of S_2 have geodesic distances greater than the predefined threshold G_{max}, thus, they are discarded.

In Fig. 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 Fig. 6a and 6b, the outer parts of the surface patches are shown. In Fig. 6c, the inner part of the ligand ESP is depicted. It is obvious that the latter patch has similar shape with the receptor ESP (Fig. 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.

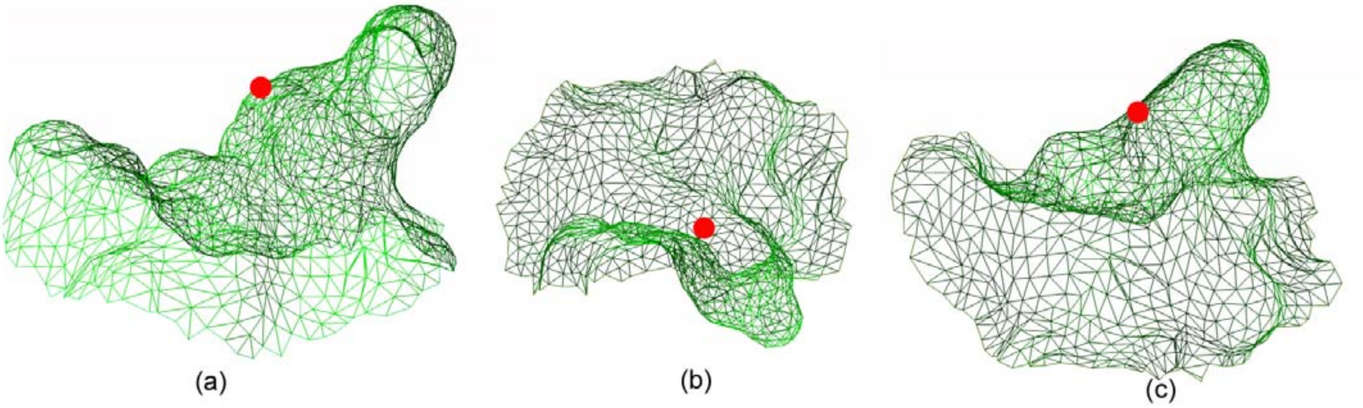

Fig. 6. a) an ESP of the receptor of the 1CGI complex (large protrusion); b) the ESP of the ligand (deep cavity) centered at a critical point which is a point of actual contact with the ESP in a); c) the ESP of b) turned upside down so that the inner surface is visible. The patches in a) and c) have approximately similar shapes.

The Shape Impact Descriptor was first introduced in [26] 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 (Fig. 7). The binary 3D function *f(i,j,k)* for each voxel [*i,j,k*] of the cubic grid is given as:

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

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.

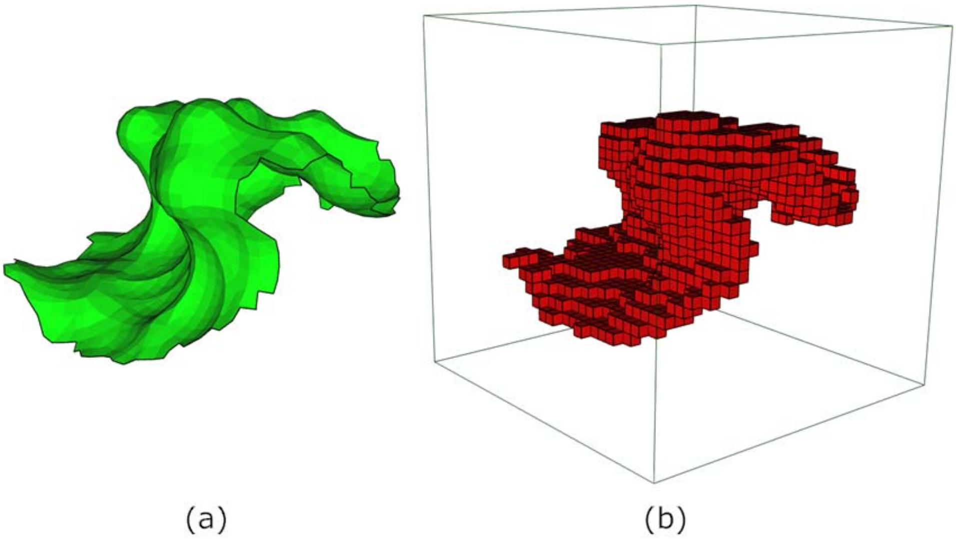

Fig. 7. a) An ESP of the receptor of the 1AY7 complex (triangulated mesh); b) the same ESP represented as a binary 3D function *f*. Here only the voxels (red boxes) where *f* has non-zero values are depicted.

3.2 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, presented in this paper, 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 [34].

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 **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:

$$E(x) = C\sum_{i=1}^{N}\frac{1}{|x-x_i|^{r+1}}(x-x_i) \qquad (3)$$

$$\phi(x) = C\sum_{i=1}^{N}\frac{1}{|x-x_i|^{r-1}} \qquad (4)$$

where r = 1,2, . . . 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) and (4) are applied to all points **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) and (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 $\left|x - x_i\right|^{r-1}$ in the denominator of (3) and (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.

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) and (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 (Fig. 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.

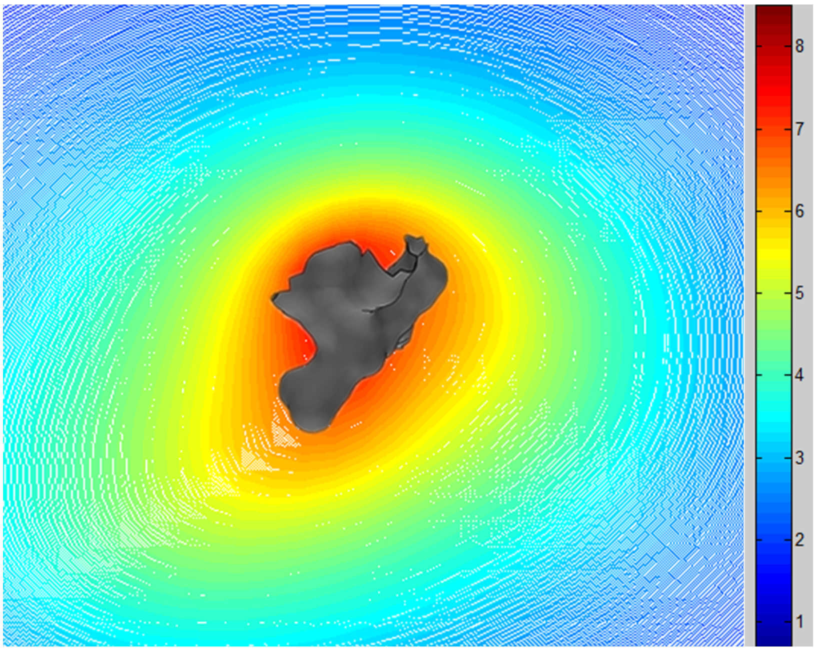

Fig. 8. The field’s potential f(x) produced from the surface of an ESP

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 **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, described in Section 3.1, is used, where points **x** lie at the centers of zero valued voxels.

$$\left\{\phi(x) : x \in R^3, \min(x - x_i) = d\right\} \qquad (5)$$

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

$$\left\{\left|E(x)\right| : x \in R^3, \min(x - x_i) = d\right\} \qquad (6)$$

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

$$\left\{E(x) \cdot n_r(x) : x \in R^3, \min(x - x_i) = d\right\} \qquad (7)$$

Where $n_r(x) = \dfrac{x - x_c}{\left|x - x_c\right|}$ 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, see description page 14, paragraph 4), 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 x M x 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 x 4 x 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 [34], which describes the extraction of the SID descriptor in detail.

3.3 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 [27], has been utilized:

$$dis(H_1, H_2) = \sum_{i=0}^{K} \frac{2\left|H_1(i) - H_2(i)\right|}{H_1(i) + H_2(i)} \tag{8}$$

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 [28] 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 summation 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 $ESP_L(j)$, as well as the Shape Impact Descriptors $SID_R(i)$ and $SID_L(j)$ for each critical point, where i=1,...,N_R and j=1,...,N_L. All pairwise dissimilarities $\mathrm{dis}_{\mathrm{ij}}$ between convex (or concave) critical points i of the receptor and concave (or convex) critical points j of the ligand, are computed:

$$dis_{ij} = dis\left(SID_R(i), SID_L(j)\right) \tag{9}$$

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 [29]. The results are shown in Table I. 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:

$$dis_{EUCL}(C_R, C_L) < \varepsilon \qquad (10)$$

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 I: 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

4 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.

4.1 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* [17] 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 **v** is the vector from *P* to *M*, as shown in Fig. 9. For the alignment of two superimposed ESPs with respect to rotation, their corresponding solid vectors (**v** and **v′**) are placed such that their angle ω is 180 degrees.

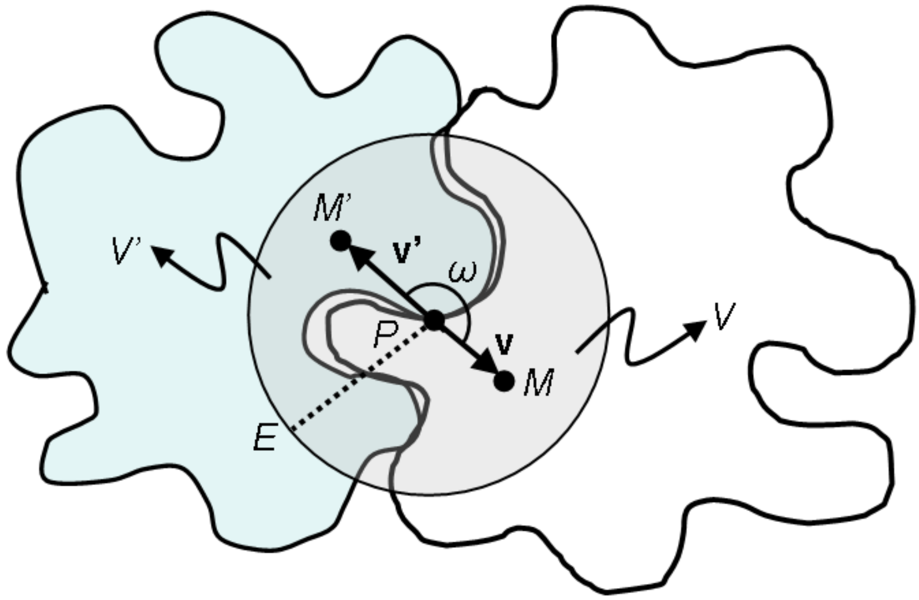

Fig. 9. Alignment of two ESPs based on their solid vectors v and v′. The angle ω between the two solid vectors is 180 degrees.

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 (Fig. 10). This results in a set of 360/ φ 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/ φ more poses. The procedure is repeated several times, keeping the direction of the solid vector within a region of solid angle Ω (Fig. 10). Eventually, a set of N_θ uniformly sampled positions of the solid vector are retained, resulting in a total of (N_θ x (360/ φ)) 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_θ x (360/ φ)) 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.

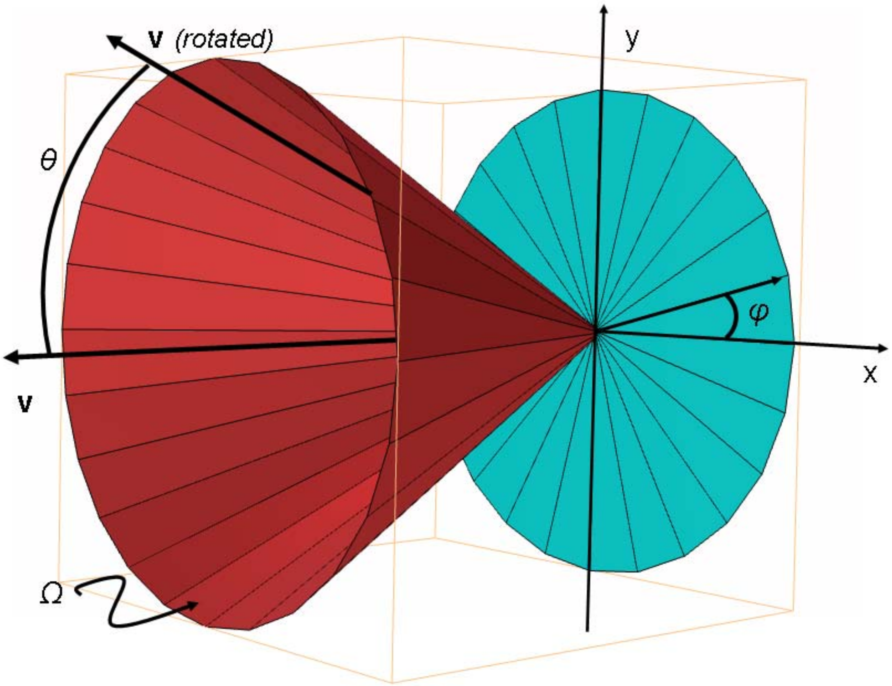

Fig. 10. Rotations of the ligand ESP, after first alignment based on solid vector: angle φ corresponds to rotations about the solid vector v. Angle θ corresponds to rotations of the solid vector from its initial position. The direction of the solid vector is kept within a region of solid angle Ω.

4.2 Geometric Scoring

For the geometric scoring of each pose, a method based on a 3D distance grid [15] 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:

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

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 II. The ranges of the shells have been experimentally determined (Section 5.1).

Table II: 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:

$$Score = \sum_{i=1}^{5} w_i N_i \qquad (11)$$

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 II). The above equation can be modified to better represent the surface of the ligand *L* in each shell, as follows:

$$Score = \sum_{i=1}^{5} w_i \left(\sum_{j=1}^{N_i} s_{ij} \right) \qquad (12)$$

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 [24], 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:

$$\frac{\sum_{j=1}^{N_1} s_{1j}}{\sum_{j=1}^{N_{Total}} s_j} > 0.5 \qquad (13)$$

$$\frac{\sum_{j=1}^{N_2} s_{2j}}{\sum_{j=1}^{N_{Total}} s_j} < 0.1 \qquad (14)$$

$$N_3 = 0, N_4 = 0 \qquad (15)$$

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.

5 Results and Discussion

The proposed method was experimentally evaluated using the protein-protein docking benchmark v2.4 [29]. 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.

5.1 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) [17], b) ZDOCK (PSC) [19] and c) PatchDock [30]. 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 [31]. This dataset was selected so as not to include complexes common in benchmark v2.4. The dataset is depicted in Table III.

Table III: 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)

The set of parameters that required optimization is given in Table IV:

Table IV: 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)

Each of the above 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.

In Table V, 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 V.

Table V: 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 V, 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 VI, 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 VI: 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 VII, 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 VII: R-bound/L-bound: The 7 cases where all three methods fail.

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

In Table VIII, 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 VIII: 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 IX, 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 IX: 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 X, 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.

Table X: 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

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 paper, 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. A discussion about how to deal with flexible docking is given in Section 5.3. 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.

In Table XI, 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 XI: 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 ≤ 2.5Å				
46	43	33	22	31
RMSD ≤ 5Å				
54	52	52	50	18

Similar conclusions can be drawn in the win-tie-loss-failure records (Table XII). 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 XII: 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

5.2 Performance Analysis of the proposed method

In Table XIIIII, 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.

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

	Number of Atoms		Number of ESPs				Number of ESP pairs
PDB	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

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 Fig. 11, 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.

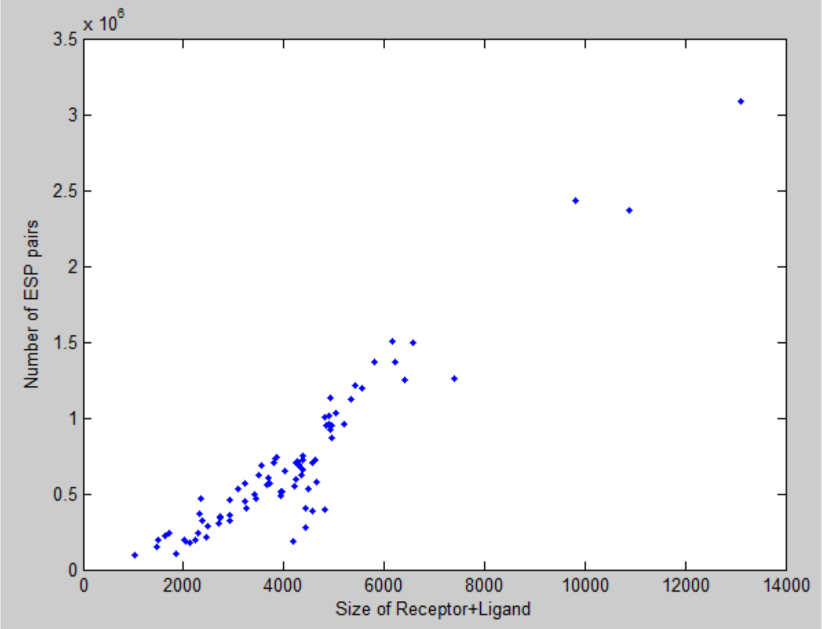

Fig. 11. Scatter-plot of receptor plus ligand size versus the total number of ESP pairs for each complex. When the total number of receptor and ligand atoms increases, then the number of ESP pairs increases as well.

In Table XIV, 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.

Table XIV: 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 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 XV, 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.

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 still faster than the method presented in this paper.

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).

Table XV: 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

5.3 Extensions

In some complexes of the Docking Benchark, described in the previous section, all methods fail to return a hit within the first ranking positions. This is due to the fact that geometric complementarity alone is insufficient for scoring protein-protein complexes as it results in multiple possible complementing solutions. It is, therefore, common to include additional physicochemical features for scoring, such as electrostatic complementarity, desolvation energy, amino acid contact preferences and Van-der-Waals potentials, by using a weighted potential function. Several approaches have been presented so far towards this direction [35]. The Shape Impact Descriptor has been introduced in this paper as a geometric method, but it could fit to non-geometric docking as well. The concept of resulting fields around a surface patch can be extended in order to describe electrostatic potentials, desolvation, Van-der-Waals potentials, etc., thus, provide an efficient descriptor incorporating additional non-geometric information. These extensions, which are planned for future research, are expected to improve the performance of the existing geometric-only solution.

In this paper, it is assumed that proteins are static shapes, i.e., rigid-bodies. In fact, protein-protein docking involves conformational changes of side chains and backbone atoms. In order to model flexibility in a docking algorithm, a solution would be to consider conformational changes as refinements after the initial rigid-body docking. This approach is known as an induced-fit process [38]. As an alternative, one can model flexibility as an ensemble of possible conformations of each protein and treat each conformation as a rigid-body [39]. The latter, also known as selected-fit process, seems to fit better to our proposed method, since geometric-only approaches, such as SID, are based on complementarity matching of rigid shapes. What would be of great interest is to combine the extensions of SID described in the previous paragraph with a selected-fit approach for flexible docking. If SID is extended so as to incorporate physicochemical features, it is expected that the multiple conformations of the selected-fit process might be avoided, since the influence of geometric complementarity will not be so strong. These thoughts are also included to our plans for further research.

6 Conclusions

In this paper, 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 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.

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.

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 and provides a challenging task for future work.

REFERENCES

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

[2] 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.

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

[4] 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.

[5] 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.

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

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

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

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

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

[11] 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.

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

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

[14] 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.

[15] 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.

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

[17] 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.

[18] R. Chen and Z. Weng. “A novel shape complementarity scoring function for protein-protein docking”. *Proteins: Structure, Function and Genetics*, 51(3):397–408, May 2003.

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

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

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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

[27] 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

[28] LING H., OKADA K.: “Diffusion distance for histogram comparison”. *In 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.

[29] 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.

[30] D. Schneidman-Duhovny, Y. Inbar, V. Polak, M. Shatsky, I. Halperin, H. Benyamini, A. Barzilai, O. Dror, N. Haspel, R. Nussinov, and H. J. Wolfson. “Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking.” *Proteins: Structure, Function and Genetics*, 52(1):107–12, 2003.

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

[32] 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

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

[34] 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.

[35] 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.

[36] 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

[37] 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

[38] C.J. Camacho and S. Vajda. “Protein-protein association kinetics and protein docking”. *Curr. Opin. Struct. Biol*., 12:36–40, 2002.

[39] G.R. Smith, M.J.E. Sternberg, and P.A. Bates. “The relationship between the flexibility of proteins and their conformational states on forming protein–protein complexes with an application to protein–protein docking”. *J. Mol. Biol*., 347(5):1077–101, 2005.