Applying Side-chain Flexibility in Motifs for Protein Docking

Hui Liu[1], Feng Lin[1,2], Jian-Li Yang[2], Hong-Rui Wang[2] and Xiu-Ling Liu[2]

[1]School of Computer Engineering, Nanyang Technological University, Singapore, Singapore. [2]School of Electronics and Information Engineering, Hebei University, Baoding, Hebei, China.

ABSTRACT: Conventional rigid docking algorithms have been unsatisfactory in their computational results, largely due to the fact that protein structures are flexible in live environments. In response, we propose to introduce the side-chain flexibility in protein motif into the docking. First, the Morse theory is applied to curvature labeling and surface region growing, for segmentation of the protein surface into smaller patches. Then, the protein is described by an ensemble of conformations that incorporate the flexibility of interface side chains and are sampled using rotamers. Next, a 3D rotation invariant shape descriptor is proposed to deal with the flexible motifs and surface patches; thus, pairwise complementarity matching is needed only between the convex patches of ligand and the concave patches of receptor. The iterative closest point (ICP) algorithm is implemented for geometric alignment of the two 3D protein surface patches. Compared with the fast Fourier transform-based global geometric matching algorithm and other methods, our FlexDock system generates much less false-positive docking results, which benefits identification of the complementary candidates. Our computational experiments show the advantages of the proposed flexible docking algorithm over its counterparts.

KEYWORDS: protein docking, motif, side-chain, flexibility, spherical harmonic descriptor

CITATION: Liu et al. Applying Side-chain Flexibility in Motifs for Protein Docking. *Genomics Insights* 2015:8 1–10 doi:10.4137/GEI.S29821.

TYPE: Original Research

RECEIVED: May 21, 2015. **RESUBMITTED:** August 23, 2015. **ACCEPTED FOR PUBLICATION:** September 1, 2015.

ACADEMIC EDITOR: Gustavo Caetano-Anollés, Editor in Chief

PEER REVIEW: Five peer reviewers contributed to the peer review report. Reviewers' reports totaled 824 words, excluding any confidential comments to the academic editor.

FUNDING: The work is also partially supported by two research grants, MOE2011-T2-2-037 (ARC 4/12) and RG139/14, from Ministry of Education, Singapore. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: asflin@ntu.edu.sg

Paper subject to independent expert blind peer review. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Background and Literature Survey

Conventional prediction uses rigid docking algorithms, but the computational results have been unsatisfactory. This is largely due to the fact that protein structures are flexible in live environments. First, the protein backbone is subject to changes during docking, and second, the involving motifs, side chains, in the bound state can be differently conformed in contrast to the unbound state. Though the conformational deformations of the side chains are not as significant as the backbone movements, they do play an important role in forming the close fit between two docking proteins. Therefore, such flexibility in motifs should be taken into account in the docking algorithms.

In the literature, a divide-and-conquer strategy is adopted for docking algorithms, with an initial-stage of candidates (hits) followed by a refinement step.[1] The refinement step aims to build an effective scoring function[2] to rank the near-native docking candidates in the top of the list. To deal with the side-chain flexibility in motifs, Cherfils et al[3] described the side chains with a crude low-resolution model to account for flexibility. Jackson et al[4] introduced side-chain flexibility into a two-stage process, which is as follows: (i) to apply the self-consistent mean field algorithm to find out the best conformation of each side chain from its rotamer library, taking solvation into account, and (ii) to perform rigid-body minimization of the intermolecular interaction energy on the interface region only, during which the larger protein is fixed. Zacharias[5] introduced side-chain flexibility in ATTRACT by using a rotamer library that contained no more than three pseudo atoms for each amino acid.[6] In RosettaDock, Wang et al[7] used a discrete rotamer library that was complemented by the side-chain conformation of the unbound state and consequently executed continuous optimization of side chains in the vicinity of the rotamers.

Prediction of the side chain can be converted into an optimization problem that aims to discover the combination of rotamers of all residues in the involving motifs to achieve the global minimum.[8] This optimization problem has been found to be NP-hard. Though many algorithms adopt the strategy of restraining the topologies of the residues to simplify and resolve this problem, it is still difficult to get a unique solution. Many factors can affect the stability of 3D structure of a protein complex, to name a few, physiochemical complementarity such as electrostatic complementarity, Coulomb potential, hydrophobicity (HP), Lennard-Jones potential, and so on.[9] Meanwhile, from the perspective of geometry, the two docking proteins usually exhibit geometric complementarity, with the surfaces of the involving motifs matching each other tightly. Therefore, geometric complementarity should be used in this regard.

In this study, we sample the possible side-chain conformations by using the rotamer library to improve local geometric complementarity. Our unique strategy includes ensembles of structures for the flexible patches associated with the motifs of the receptor and ligand considered and a fast docking method for achieving a list of docking candidates based on their ensembles.

Side-chain Flexibility in Motifs and the Algorithm

Applying Morse theory to flexible protein surface segmentation. Structural information of a protein molecule is available in the database Protein Data Bank (PDB), especially the 3D coordinates of all the atoms that constitute the protein. We base the geometric complementarity on atomic representation of the residues lying on the surface. A mathematical model has been developed to describe the protein surface with a sparse distribution of the atoms on the protein surface. We start with the solvent-excluded surface (SES) extraction, as shown in Figure 1.

In the model, the atoms are represented with the spheres of different van der Waals radii. The original protein molecule is assumed as a set S that is made up of a list of overlapping spheres. We adopt the maximal speed molecular surface algorithm[10] to extract the SES, which scrolls a probe sphere of the surrounding solvent molecule over the van der Waals surface of the protein molecule, resulting in the boundary of reachable volume of the probe. A concave patch serves as a connection band to fill the gap between two nearby atoms.

An SES is continuous but may contain singularities. For the segmentation purpose, we simplify the representation to a triangular mesh. This algorithm has been implemented, as shown in Figure 2, where the surface is extracted and triangulated over the macromolecule 4PTI.pdb.

Upon acquisition of the triangular mesh of the SES, it is segmented into patches in order to do geometric matching between the protein surfaces. Let v_i be a vertex on the triangular mesh, n_i be the normal vector of v_i, $v_{adj} = \{v_1, \ldots, v_m\}$, a set of vertices, be the adjoining vertices of v_i, T_i be the list of triangles that are connected by v_i, and $NT_i = \{nt_{tri_j} | tri_j \in T_i\}$ be a set of normal vectors of triangles in T_i. The Morse theory[11] is regarded as a direct method for analyzing the topology of a manifold by studying the differentiable functions on the manifold.

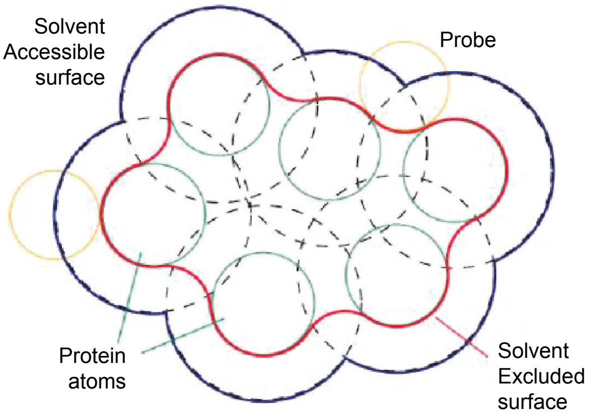

Figure 1. SES extraction using a probe sphere: the atoms are represented with the spheres of different van der Waals radii; the original protein molecule is made up of a list of overlapping spheres.

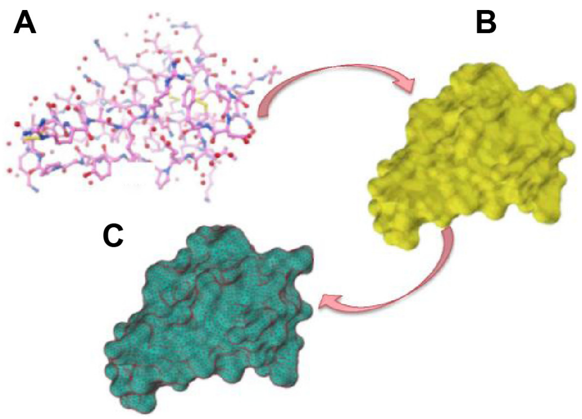

Figure 2. SES is extracted from (**A**) the ball-stick representation of 4PTI.pdb in PDB to (**B**) SES surface and simplified to (**C**) triangular mesh.

To apply the Morse theory, let M^2 be a closed two-manifold surface and f: $M^2 \to R$ a real-valued smooth function. Based on the value of the gradient of f at surface point p, we determine the type of p as follows. If $\nabla f(p) = 0$, p is a critical point; otherwise, it is considered as a regular point. Furthermore, if the Hessian matrix $H(f)$ at the critical point p is nonsingular, p is defined as nondegenerate, else it is called degenerate. The nondegenerate critical points make up the local extreme points and the saddle points. In this way, we segment the triangular mesh and get the local minima, local maxima, and saddle points on it. The mean curvature function serves as the Morse function f, and all the critical points on the triangular mesh can be extracted as a minima if $f(v_i) < f(v_j)$, v_j ε v_{adj}; a maxima if $f(v_{i,}) > f(v_j)$, v_j ε V_{adj}; and regular otherwise. The following procedure describes our region-growing algorithm to decompose the molecular surface:

i. An initial segmentation will decompose the surface coarsely into three types of surface regions, namely, concave, convex, and flat. Both the Gaussian and mean curvatures are calculated for all vertices v_i on the surface mesh to label the surface type. We compute the curvature based on fitting local surfaces.[12]

ii. The three different types of regions are further decomposed into smaller patches. During this procedure, we extract a list of critical points and then around each critical point create a surface patch that contains all the surface points whose geodesic distance from the critical point is less than the experimental value 16 Å.

iii. A local coordinate space for each vertex v_i is constructed. Let v_i be located at the origin and n_i be the z-axis. The x-axis and y-axis are two orthogonal vectors in a plane passing through v_i. We transform all the adjoining vertices to the local coordinate system of v_i, and the fitting surface can be described as a quadric function by applying the least-squares method.

iv. Upon the Gaussian and mean curvatures computed on the mesh, we utilize the criteria proposed by Besl and Jain[13] to label each triangle on the mesh: concave, convex, or flat.[14] The peak, ridge, and saddle ridge are considered as the convex type; the flat and minimal surfaces are contained in the flat type, and finally, the concave type consists of pit, valley, and saddle valley.

v. Finally, we have developed a region-growing algorithm to decompose the entire triangular mesh into convex, concave, and flat regions. Each triangle and its neighboring triangles are segmented into the same surface type. The algorithm is described in the following pseudocodes.

Algorithm: Region Growing in the Three Types of Patch

```
Initializing the segment number seg_id of each triangle in TriArray to -1, assign the current segment number id=0;
Mark the surface type for all the triangles in TriArray;
for (each triangle tri$_j$ in TriArray)
    while (seg_id(tri$_i$)!= -1) {
        seg_id = id;
        Segment(id) = NULL;
        add tri$_i$ into Segment(id);
        for (each triangle tri$_j$ in Segment(id))
            for (each triangle tri$_k$ ε NT$_j$)
                if (the surface type of tri$_k$ is the same as that of tri$_j$
                    && seg_id(tri$_k$) == -1)
                    seg_id(tri$_k$) = id;
                    add tri$_k$ into Segment(id);
        id++;
    }
```

In the procedure, the region is expanded in all directions around a critical point[15] until it reaches the region contour of the surface, resulting in an elementary surface patch. To show the result using the protein 1CGI_ligand.pdb, in Figure 3A, the yellow points represent the critical points on the molecular surface and in Figure 3B, the protein surface regions are segmented into three types of patches: convex (in green), flat (in yellow), and concave (in pink).

Extraction of involving motifs associated flexible interaction sites. Based on the segmented surface, we can now describe the protein with an ensemble of conformations that incorporate the flexibility of interface side chains and are sampled using rotamers. As will be further discussed in the next section, the highest geometric matching score is assumed to correspond to the closest conformation in the actual bound state. Therefore, we focused on reduction of the number of possible side-chain conformations so as to get a sample of rational size. The main processes are as follows:

i. As protein docking is dominated by short sequences of amino acids, we use these sequences to control the conformational deformations during binding. The atoms in the short sequences constitute the contact surfaces between the two docking proteins. Our algorithm identifies the flexible contact surface as well as the amino acids on the contact surface, referred as interaction sites.

ii. In contrast to the other amino acids in the protein surfaces, the interaction sites have some unique properties that facilitate binding the two proteins together. In our algorithm, the priority in identification of the interaction sites is for the physicochemical and geometric complementarity, such as the cavities of the surface. Hence, instead of a full computation of the conformation of the whole molecular surface, we mainly compute the complementarity of the side chains on the interface sites between the two docking proteins.

iii. Next, as the interface region is small, so it is sufficient to describe the conformational deformations within the local surface regions. We divide the interface sites into small molecular surface patches that contain eight or

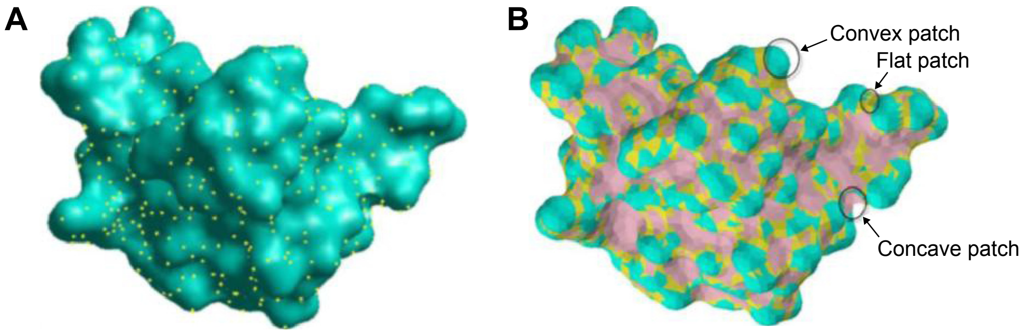

Figure 3. Elementary surface patch using the protein 1CGI_ligand.pdb: (**A**) the yellow points represent the critical points on the molecular surface and (**B**) the protein surface regions are segmented into three types of patches: convex (in green), flat (in yellow), and concave (in pink).

nine residues in practices. During the docking process, a set of patches is selected to lay over the entire molecular surface. We compare our sampled conformations with the side-chain conformations in the bound complex.

iv. To incorporate the flexibility into the interaction sites, various conformations of the side chains of the residues contained in the patch are generated. To estimate the conformations of side chains, we also consider the conformational space of the side chains as discrete, and assume each discrete conformation as a rotamer. The rotamers are then grouped into the rotamer library.

v. Furthermore, the algorithm identifies the best combination of rotamers that corresponds to the lowest-energy state. A through listing of all possible rotamer combinations of the amino acids included in one surface patch eventually results in a few flexible conformations.

vi. Finally, to avoid only adopting the 3D conformations with the lowest energy in the combination, we also consider a better sampling of the conformational space approachable by the patch. Our experiments show that conformations with similar low energies show few differences around the same local minimum. Hence, in order to achieve a broader sampling, we subdivide the conformational space of each residue into three parts based on its three common c1 torsion angles (60°, 180°, and −60°).

Figure 4 presents the key steps in extraction of the flexible interaction site for the 1CGI complex obtained from PDB. Figure 4A gives an overview of the conformation, and Figure 4B reveals the interaction sites (ribbon rendering) in the midst of the running algorithm. Figure 4C shows the extracted ligand patch (surface shading), and Figure 4D gives the identified interaction site from the ligand. In this example, we assume that the patch includes one tryptophan that has 6 rotamers, two aspartic acids that each has 3 rotamers, one glutamine that has 28 rotamers, one lysine that has 8 rotamers, one arginine that has 6 rotamers, and two serines that each has 3 rotamers. Therefore, the flexible surface has a total of $6 \times 3 \times 3 \times 28 \times 8 \times 6 \times 3 \times 3 = 653{,}184$ possible conformations.

In practice, we can sample configuration of every possible torsion angle cl of the patch residues. As mentioned in the algorithm descriptions, we assume that there are about eight or nine residues on each patch. For example, take the patch that has nine residues. Each residue has no more than three cl angle possibilities. The amount of cl configurations reaches to $3^9 \approx 20{,}000$. Rather than fetching the lowest-energy samples out from all the conformations, we pick out the lowest-energy

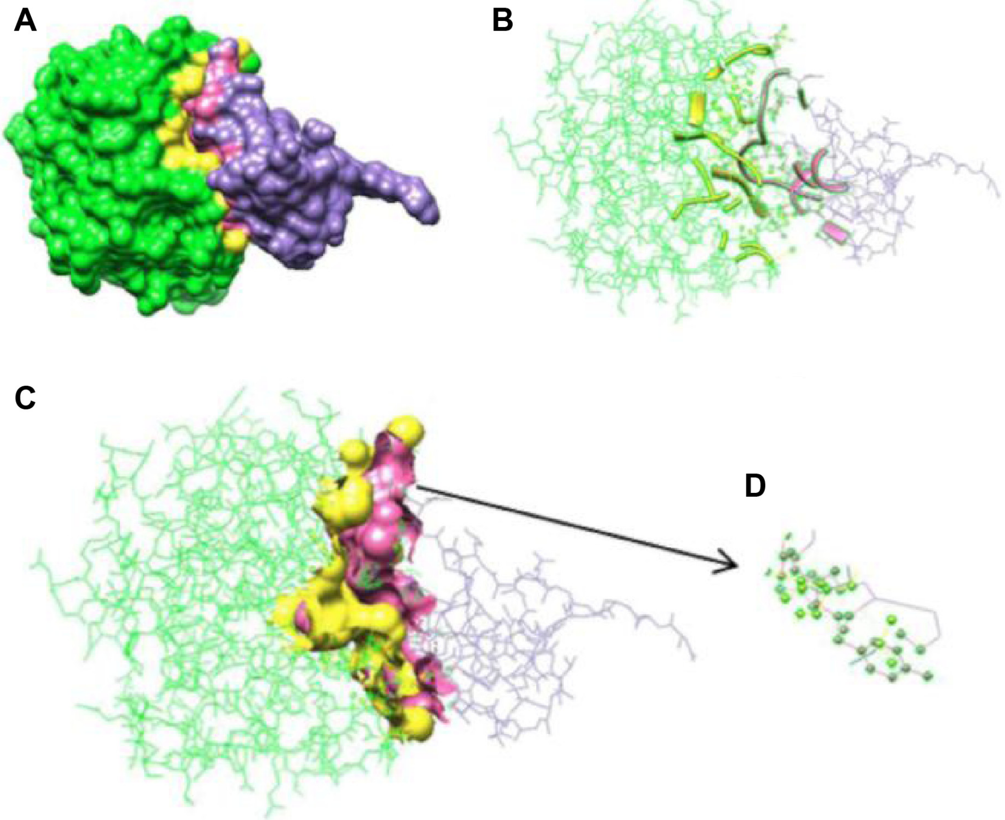

Figure 4. Extraction of the interface sites for the 1CGI complex: (**A**) shows an overview of the conformation, (**B**) reveals the interaction sites (ribbon rendering) in the midst of the algorithm running, (**C**) shows the extracted ligand patch (surface shading), and (**D**) shows the identified interaction site from the ligand.

structure of each of the about 20,000 cl configurations as follows:

i. Fix the cl configurations and choose the torsion angles c2 and c3 in the patch by energy minimization. The energy function involves the van der Waals energy, Coulomb energy, and so on.
ii. Apply the clustering methods, such as *k*-means, according to the root mean square deviation (RMSD) to group the 20,000 configurations into a much smaller number of clusters.
iii. Select one configuration that has the lowest-energy conformation from each cluster, and put it into the final sample for the patch.

Spherical harmonic descriptor (SHD)-based surface matching and alignment. With the flexible surface representation, we are able to get all the possible relative positions of the two docking monomers. The 3D conformations of protein complexes indicate a close geometric match on the interface sites. Hence, geometric matching plays a significant role in the docking. As described in the previous sections, our flexible docking method is based on local-shape feature matching. Surface complementarity is largely the geometric similarity matching. The key is how to extract and describe the shape features for the similarity matching, considering that proteins can have the same representation with their conformation after a geometric transformation.

Two methods can be considered: 3D protein structures are normalized by applying a canonical transformation, or the 3D structures are described by a transformation invariant descriptor. In general, protein structure can be normalized by moving its mass center to the origin of the coordinate system for translation and by setting main axes for rotation. Existing methods are robust for translation normalization but not rotation normalization. Therefore, docking is better to be based on local-shape feature matching by a 3D shape descriptor. Based on the concept of the SHD,[16] we propose a rotation invariant protein structure descriptor. The main processes are as follows.

First, the surface patch is represented with a three-dimensional matrix M and then rasterized into a $2R \times 2R \times 2R$ voxel grid, where R is the radius of the bounding sphere. The grid cell is assigned a scalar value 1 if it is located in the voxel of a protein surface patch, otherwise 0. Before getting the protein descriptor for each surface patch, the centroid should be determined by the following equations:

$$\overline{x} = \frac{M_{100}}{M_{000}}, \overline{y} = \frac{M_{010}}{M_{000}}, \overline{z} = \frac{M_{001}}{M_{000}}$$

where $M_{lmn} = \sum_{i=0}^{L} \sum_{j=0}^{M} \sum_{k=0}^{N} x_i^l y_j^m z_k^n$.

The surface patch is translated so that the centroid is located at the point (0, 0, 0), and thus, the radius of the bounding sphere is R.

Then, the Cartesian coordinates (x, y, z) in the 3D space are transformed to the corresponding spherical coordinates (r, θ, φ); thus, the voxel grid can be defined as a binary-valued function:

$$f(r, \theta, \varphi) = Voxel\,(r\, sin\theta\, cos\varphi,\ r\, cos\theta,\ r\, sin\theta\, sin\ \varphi)$$

where $r\varepsilon[0, R]$, $\theta\varepsilon[0, \pi]$ and $\varphi\varepsilon[0, 2\pi]$.

With different values of radii, a set of spherical functions $\{f_0, f_1, \ldots f_R\}$ can be generated: $f_r(\theta, \varphi) = f(r, \theta, \varphi)$.

Assuming $f(r, \theta, \varphi) = R(r)\, Y(\theta, \varphi)$ and using the Laplace's equation in spherical coordinates, we have

$$\nabla^2 f = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial f}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial f}{\partial \theta}\right) + \frac{1}{r^2 (\sin\theta)^2}\frac{\partial^2 f}{\partial \varphi^2} = 0$$

By separating the variables, two differential equations can be calculated by imposing Laplace's equation:

$$\frac{1}{R}\frac{\partial}{\partial r}\left(r^2 \frac{\partial R}{\partial r}\right) = \lambda$$

$$\frac{1}{Y}\frac{1}{\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial \varphi}{\partial \theta}\right) + \frac{1}{Y}\frac{1}{(\sin\theta)^2}\frac{\partial^2 Y}{\partial \varphi^2} = -\lambda$$

We can now obtain the spherical harmonic function by further separating the variables:

$$Y_l^m(\theta, \varphi) = \sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}}\, p_l^m(\cos\theta) e^{im\varphi}$$

where $\lambda = l(l + 1)$; $l = 0, 1, 2, \ldots$; $m = -l, -1 + 1, \ldots, l - 1, l$; and $p_l^m(\cos\theta)$ are the associated Legendre polynomials.

Based on the spherical harmonics, the spherical function can be represented with the amount of energy it has at different frequencies. We first decompose the spherical function into its spherical harmonics, then summarize the harmonics within each frequency, and finally, compute the norm of each frequency component as follows:

$$f_r(\theta, \varphi) = \sum_m f_r^l(\theta, \varphi)$$

$$f_r(\theta, \varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{m=l} a_{lm} Y_l^m(\theta, \varphi)$$

Let v_1 be the subspace of the spherical harmonics. Then, we have

$$v_l = span\left(Y_l^{-l}, Y_l^{-l+1}, \ldots, Y_l^{l-1}, Y_l^l\right)$$

For any $f\, \varepsilon\, V_l$ and spherical rotation rotation $R(f)\ \varepsilon\ V_l$,

$$f_r(\theta, \varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{m=l} a_{lm} Y_l^m(\theta, \varphi)$$

$$R\left(f_r(\theta, \varphi)\right) = \sum_{l=0}^{\infty} \sum_{m=-l}^{m=l} R\left(a_{lm}\right) Y_l^m(\theta, \varphi)$$

For a spherical rotation, we have the property:

$$\Sigma_{m=-l}^{m=l} \mid a_{lm} \mid^2 = \Sigma_{m=-l}^{m=l} \mid R(a_{ml}) \mid^2$$

The energy at each frequency I is rotation invariant:

$$\mid f_r^l(\theta,\varphi) \mid = \Sigma_{m=-l}^{m=l} \mid a_{lm} \mid^2$$

From the properties of the spherical harmonics, we know that the L_2-norm of the spherical function will not change when it is rotated, so the energy function can be represented by

$$SH(f_r) = \left\{ \| f_r^0(\theta,\varphi) \|, \| f_r^1(\theta,\varphi) \|, \ldots \right\}$$

where f_r^l are the frequency components of f_r and can be obtained by

$$f_r^l(\theta,\varphi) = \pi_l(f_r) = \sum_{m=-l}^{m=l} a_{lm} Y_l^m(\theta,\varphi)$$

The above expression is independent of the orientation of the spherical function, so we get

$$\begin{aligned} SH\left(R(f_r)\right) &= \left\{ \| \pi_0\left(R(f_r)\right) \|, \| \pi_1\left(R(f_r)\right) \|, \ldots \right\} \\ &= \left\{ \| R\left(\pi_0(f_r)\right) \|, \| R\left(\pi_1(f_r)\right) \|, \ldots \right\} \\ &= \left\{ \| \pi_0(f_r)) \|, \| \pi_1(f_r) \|, \ldots \right\} = SH(f_r) \end{aligned}$$

With the above derivation, we get the rotation invariant feature for each surface patch, which is given as

$$F(r,l) = \sqrt{\Sigma_{m=-1}^{l} \mid a_{ml}^r \mid^2}$$

where (r, l) corresponds to the length of the lth frequency of the restriction of f to the sphere with radius r.

In this way, to measure the similarity between two surface patches, we calculate the Euclidean distance between the two corresponding SHDs, which is as follows:

$$D = \sqrt{sum \mid F_1(r,l) - F_2(r,l) \mid^2}$$

In our study, a protein surface is segmented into three types of patches: convex, concave, and flat. We use the derived rotation invariant 3D shape descriptor to get the signature for each patch. By comparing the signatures, we are able to filter out the patch pairs with the most similarities. As illustrated in Figure 5, each convex patch from the receptor will be matched with all the concave patches of the ligand and vice versa.

During the process of alignment, a transformation matrix for the ligand is computed by using the iterative closest point (ICP) algorithm. As shown in Figure 6, the red points and blue points, respectively, belong to the point sets A and B for registration. As each surface patch is composed

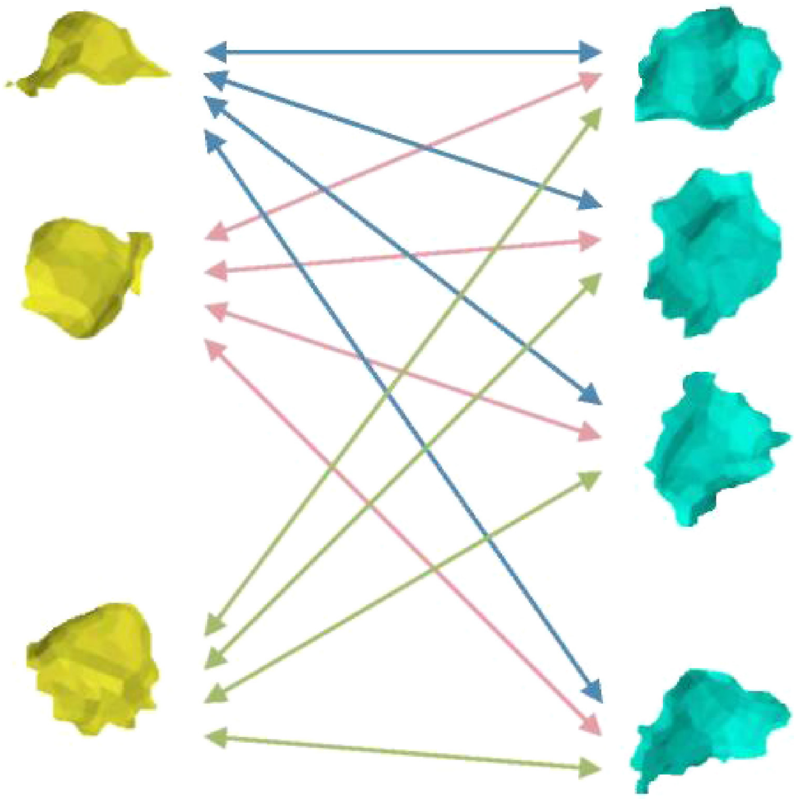

Figure 5. Pairwise similarity between convex patches of ligand with concave patches of receptor and vice versa: the derived rotation invariant 3D shape descriptor is used to get the signature for each patch.

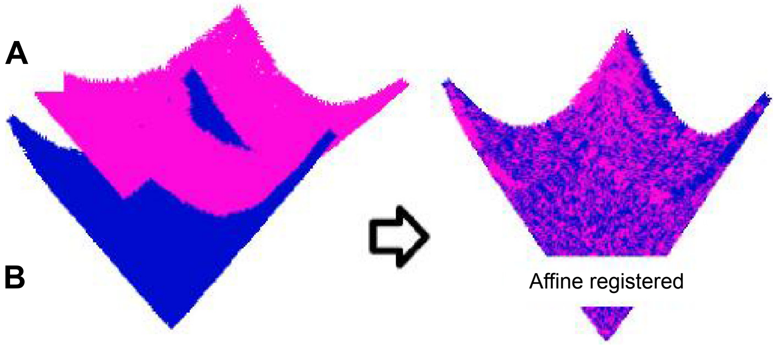

Figure 6. Registration of two different point sets using the ICP algorithm: the red points and blue points, respectively, belong to the point sets A and B for registration. ICP is used for geometric alignment of the two 3D protein surface patches, one from the receptor that is fixed and the other from the ligand that is transformed.

of a set of points, ICP is used for geometric alignment of the two 3D protein surface patches, one from the receptor that is fixed and the other from the ligand that is transformed. In our experiments, as the area of the surface patch is fairly small, we set the iterative parameter as 10. The final transformation matrix is applied to transform the ligand protein, which, in turn, binds the receptor protein to form the candidate structure.

Scoring of complementarity. Upon complementarity matching, a list of candidate complex conformations will be generated. We have to select the near-native complex conformations from the docking candidates, known as scoring or ranking in protein–protein docking.

Owing to the fact that most of the energy obtained upon complex formation is derived from the hydrophobic effect and the hydrophobic forces are short ranged, the two docking proteins should have short distances between interaction sites. This is in correspondence with the premise that the two docking monomers should exhibit corresponding radii of curvature on macroscopic and microscopic scales on their surfaces. We can accordingly build the complementarity of the confrontation of concavities on one side and convexities on the other side in the interaction sites.

The scoring method used in our work is to partition the receptor into several shells based on the distance from the protein surface. Each shell is determined by a range of distances in the 3D distance grid. Each conformation of the protein–protein docking candidates is scored by a weighted function of the number of ligand surface points in each shell. Briefly, the scoring function is computed after transformation and alignment, while the SES of the ligand enters into the 3D distance grid of the SES of the receptor. Each surface point of the ligand is assigned with a value according to its distance from the surface of the receptor. The scoring function is defined as

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

where N_i denotes the number of the ligand surface points that are located in the ith shell of the receptor and w_i signifies the weight of shell i. Furthermore, as the SESs of both ligand and receptor are represented by the triangular mesh, the above function can thus be described precisely as follows:

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

In this equation, N_i denotes the number of triangles on the triangular mesh of the ligand whose centroids are located in the ith shell. s_{ij} denotes the surface area (in Å) of the triangle j in the ith shell.

While this method can give an accurate geometric scoring for the 3D structure of the docking candidates, its computational time grows fast with the increment of size and resolution of the surface of the ligand. Multiresolution can be considered. In our scoring system, a low-level mesh resolution with the point density of one point per angstrom and a high-level mesh resolution with the point density of four points per angstrom have been utilized in the scoring function. First, the low-resolution mesh surface is applied to all the docking candidates, and only the 3D conformations with the highest scores are extracted to be further filtered by using their high-resolution mesh surface.

In the final step, we calculate the RMSD between each candidate structure and native structure of the corresponding complex. The results are filtered into high accuracy (RMSD $<$ 2.5 Å) and medium accuracy (RMSD $<$ 5 Å). In our experiments, there are two thresholds 2.5 Å and 5 Å. If the distance is less than the threshold, we call the candidate structure a hit.

Computational Experiment Results

To assess the performance of our flexible docking system, FlexDock, a public dataset *protein–protein docking benchmark*[17] is used. In this dataset, for each complex, the receptor and ligand are separated from each other. The 3D position of the receptor is fixed as it is in the complex, whereas that of the ligand is transformed. Different docking systems are applied to get a list of candidate docking poses of the ligand. If the RMSD between the candidate pose and the ground-truth pose in the interaction sites is within the predefined threshold, the candidate pose is defined as a hit.

The experimental results of our FlexDock are compared with those of the other two docking systems: PatchDock is local patch-based searching method where the SES is generated,[18] while ZDock is a fast Fourier transform-based global searching method where the surface residues are extracted.[19] As in the ZDock system, when the sampling angular is set to 15°, the total amount of conformations predicted is limited to 3600. For a fair comparison, only the top 3600 docking candidates were considered for all the three systems.

In the experiments, we set the threshold to 2.5 Å. For each system, the rank of the first hit within the top 3600 candidates is described for all the test cases in the Benchmark v2.4. Owing to the space constraint, only the top 84 hits of this large table are given in Table 1. These hits do not necessarily correspond to the smallest RMSD value. We found that our FlexDock identified 70 out of 84 hits in the top list of the database, which is much more than 45 and 57, respectively, from the other two systems.

To assess the execution performance, we looked into the different stages of the docking tasks. In the first process, the molecular surface of the receptor and ligand is first extracted and segmented into convex and concave patches, respectively. We can see that the number of atoms and the corresponding segmented surface patches for each complex in the Benchmark v2.4 increase proportionally to the increment in the atom number of the two docking proteins. The average running time for preprocessing and other steps during docking is described in Table 2. The experiments are conducted using a CPU with a quad-core 3.20 GHz processor and 4 GB RAM.

The first process consists of extracting and segmenting the protein surface. The average time cost in this step is 29.06 seconds for each pair of receptor and ligand. This task can be completed offline.

The time to calculate the descriptor for each patch is 0.49 seconds. In the experiment, the average patch number for each pair of the interacting proteins is 2673. Therefore, the average descriptor extraction time for the two docking proteins is 1309 seconds.

The comparison of descriptors is to calculate the distance between the two signatures of each patch pair as a value in the range [0.0, 2.0], with small values corresponding to two similar patches. The patch pairs are sorted based on the distance value from smallest to largest. Only the first several patch pairs will be used for aligning the receptor and ligand. The average time for comparison of descriptors between each pair

Table 1. Comparisons: rank gives the first hit within the top 3600 predictions with a threshold of 2.5°Å (only top 84 are shown).

	PatchDock	ZDock	FlexDock
PDB	Rank	Rank	Rank
1A2K	300	570	**12**
1ACB	**10**	**6**	184
1AHW	**40**	56	96
1AK4	–	3471	**3**
1AKJ	–	448	**6**
1ATN	–	558	**2**
1AVX	43	1	**1**
1AY7	24	46	**1**
1B6C	40	24	**3**
1BGX	–	–	–
1BJ1	–	3	–
1BUH	83	393	**10**
1BVK	131	1087	**96**
1BVN	**1**	10	101
1CGI	1	1	1
1D6R	–	35	**3**
1DE4	–	452	–
1DFJ	–	–	–
1DQJ	83	**19**	269
1E6E	**2**	58	132
1E6J	1706	699	**23**
1E96	1767	–	**312**
1EAW	**1**	**1**	**1**
1EER	1	–	**1**
1EWY	139	–	**104**
1EZU	**1**	–	**1**
1F34	**1**	–	**1**
1F51	**1**	–	**1**
1FAK	–	–	**182**
1FC2	49	55	**5**
1FQ1	–	–	–
1FQJ	248	120	**102**
1FSK	218	**19**	197
1GCQ	–	382	**350**
1GHQ	–	–	–
1GP2	–	–	**411**
1GRN	3	7	**2**
1H1V	–	1510	**1182**
1HE1	**1**	7	**1**
1HE8	–	–	–
1HIA	14	**1**	17
1I2M	–	14	**1**
1I4D	**167**	793	385
1I9R	–	1271	**31**
1IB1	–	–	**1**

Table 1. (*Continued*)

	PatchDock	ZDock	FlexDock
1IBR	–	–	**1**
1IQD	–	55	**10**
1IJK	–	–	–
1JPS	96	**23**	217
1K4C	337	**30**	396
1K5D	–	**10**	554
1KAC	–	381	**332**
1KKL	–	–	**127**
1KLU	–	–	116
1KTZ	–	–	76
1KXP	–	–	1031
1KXQ	29	30	**6**
1M10	–	33	**18**
1MAH	**1**	**1**	**1**
1ML0	7	75	**3**
1MLC	516	1205	11
1N2C	–	–	–
1NCA	–	20	76
1NSN	–	–	–
1PPE	**1**	2	**1**
1QA9	–	–	**2**
1QFW	–	**16**	160
1RLB	3143	–	1045
1SBB	–	–	**135**
1TMQ	**1**	8	**1**
1UDI	**1**	**1**	**1**
1VFB	–	–	1249
1WEJ	–	1120	**126**
1WQ1	**1**	4	**1**
2BTF	137	21	**1**
2JEL	282	532	**40**
2MTA	**115**	1447	175
2PCC	–	–	–
2SIC	–	9	**1**
2SNI	13	4	**1**
7CEI	–	5	**1**

of patches is 0.001 seconds, which is nearly 100 times faster than the scoring task. Therefore, the comparison of descriptors can be used as a prescoring and filtering tool. This can help to accelerate the docking procedure.

Overall, the average running time of the three systems for all the test cases in Benchmark v2.4 is given in Table 3. While our FlexDock is able to identify a hit for more cases than the other two docking systems, it runs faster than ZDock but slower than PatchDock. This is mainly because, for flexible surface docking, the number of segmented patches for each pair of receptor and ligand is much larger than those of the other systems.

Table 2. The average running time for each step.

PROCESSING STEP	AVERAGE COMPUTING TIME
Extracting and segmenting each pair of proteins	29.06s
Descriptor calculation/patch	0.490s
Descriptor comparing/each pair of patches	0.001s
Alignment/each pair of patches	0.028s
Scoring/each ligand pose	0.135s

Discussions and Conclusions

The influence of the side-chain flexibility in motifs has been analyzed, and it has been added into the local surface patches for a soft surface representation. The flexibility of the side chains has also been implemented by the rotamers. It is helpful to improve the accuracy of the docking algorithm. We summarize the procedure in our proposed algorithm as follows:

i. Applying the Morse theory to the flexible protein surface segmentation: The region is expanded in all directions around a critical point until it reaches the region contour of the surface, resulting in an elementary surface patch. The surface regions are segmented into three types of patches: convex, flat, and concave.
ii. Extraction of the involving motifs: Extraction of the flexible interaction site includes first the conformation, the progressive interaction sites, the extracted ligand patch, and the identified interaction site from the ligand.
iii. Surface matching and alignment: All the possible relative positions of the two docking monomers are acquired. Based on the SHD, a rotation invariant protein structure descriptor is utilized. During the process of alignment, a transformation matrix for the ligand is computed by using the ICP algorithm. The final transformation matrix is applied to transform the ligand protein and to bind the receptor protein to form the candidate structure.
iv. Scoring of complementarity: A list of candidate complex conformations is generated to select the near-native complex conformations from the docking candidates. A low-level mesh resolution with the point density of one

Table 3. The average running time of each system.

SYSTEM	AVERAGE RUNNING TIME FOR THE BENCHMARK v2.4
ZDock	3192s
PatchDock	1098s
FlexDock	2735s

point per angstrom and a high-level mesh resolution with the point density of four points per angstrom are utilized in the scoring function. If the distance is less than the threshold, the candidate structure is a hit.

Innovation in our algorithm designs benefits protein docking. First, in our model, we extract the SES of the protein molecules and segment it into concave, convex, and flat patches. The atoms are represented with the spheres of different van der Waals radii, and the original protein molecule can be represented in a list of overlapping spheres. The Morse theory is applied for analyzing the topology of the protein molecule surface by studying the differentiable functions on the surface. We are able to describe the protein with an ensemble of conformations. The flexibility of interface side chains is incorporated by sampling their conformations using rotamers, in which the matching scores correspond to the conformation in the actual bound state. Eventually, we are able to reduce the number of possible side-chain conformations so as to get a sample of rational size.

For the surface patch alignment, we propose a transformation matrix for computing the ligand. With each surface patch being composed of a set of points, ICP shows its usefulness in geometric alignment of the two 3D protein surface patches. The final transformation matrix can be applied to transform the ligand protein, which, in turn, binds to the receptor protein to form the candidate structure.

In candidate complex conformations, because the energy is most derived from the hydrophobic effects and the forces are short ranged, we are able to build the complementarity with concavity on one side and convexity on the other side.

While our proposed flexible docking method exhibits its advantages in identification of complementary candidates, we have not taken into considerations physicochemical factors during the scoring process. In the future, we will combine the geometrical and physiochemical factors, such as sequence conservation, HP, interface residue propensity, electrostatic potential and so on, to filter out the surface patches with binding sites.[20,21]

Acknowledgments

We acknowledge all the valuable inputs from the clinical departments of Hebei University Affiliated Hospital.

Author Contributions

Conceived and designed the experiments: HL and FL. Analyzed the data: HL and XLL. Wrote the first draft of the manuscript: HL and FL. Contributed to the writing of the manuscript: HL, FL, JLY, HRW, and XLL. Agreed with manuscript results and conclusions: HL, FL, JLY, HRW, and XLL. Jointly developed the structure and arguments for the paper: HL, FL, and XLL. Made critical revisions and approved the final version: FL. All authors reviewed and approved the final manuscript.

REFERENCES

1. Tovchigrechko A, Vakser IA. GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res*. 2006;34(suppl 2):W310–W314.
2. Wang CX, Chang S, Gong XQ, Yang F, Li CH, Chen WZ. Progress in the scoring functions of protein-protein docking. *Acta Phys Chim Sin*. 2012;28(4):751–758.
3. Cherfils J, Bizebard T, Marcel Knossow JJ. Rigid-body docking with mutant constraints of influenza hemagglutinin with antibody HC19. *Proteins*. 1994;18:8–18.
4. Jackson RM, Gabb HA, Sternberg MJE. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J Mol Biol*. 1998;276:265–285.
5. Zacharias M. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci*. 2003;12:1271–1282.
6. May A, Zacharias M. Protein-protein docking in CAPRI using ATTRACT to account for global and local flexibility. *Proteins*. 2007;69:774–780.
7. Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein-protein docking. *Protein Sci*. 2005;14:1328–1339.
8. Lasters I, De Maeyer M, Desmet J. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. *Protein Eng*. 1995;8(8):815–822.
9. Lin L, Guo D, Huang Y, Liu S, Xiao Y. ASPDock: protein-protein docking algorithm using atomic solvation parameters model. *BMC Bioinformatics*. 2011;12(1):36.
10. Sanner MF, Olson AJ, Spehner JC. Fast and robust computation of molecular surfaces. In: Proceedings of 11th ACM Annual Symposium on Computational Geometry, Vancouver, B.C., Canada, 1995.
11. Bestvina M, Brady N. Morse theory and finiteness properties of groups. *Invent Math*. 1997;129(3):445–470.
12. Rusinkiewicz S. Estimating curvatures and their derivatives on triangle meshes, 3D Data processing, visualization and transmission, 2004, 3DPVT 2004. In: Proceedings 2nd International Symposium on. IEEE, Thessaloniki, Greece, 2004.
13. Besl PJ, Jain R. Segmentation through variable-order surface fitting. *IEEE PAMI*. 1988;10:167–192.
14. Zhu H, Domingues F, Sommer I, Lengauer T. NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*. 2006;7(1):27.
15. Banchoff TF. Critical points and curvature for embedded polyhedral surfaces. *Am Math Monthly*. 1970;77:475–485.
16. Kazhdan M, Funkhouser T, Rusinkiewicz S. Rotation invariant spherical harmonic representation of 3D shape descriptors. In: Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing. Eurographics Association, 2003.
17. Mintseris J, Wiehe K, Pierce B, et al. Protein-protein docking benchmark 2.0: an update. *Proteins*. 2005;60(2):214–216.
18. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res*. 2005;33(suppl 2):W363–W367.
19. Ritchie DW, Kozakov D, Vajda S. Accelerating and focusing protein-protein docking correlations using multi-dimensional rotational FFT generating functions. *Bioinformatics*. 2008;24(17):1865–1873.
20. Liu H, Lin F, Lee YT, Qian K, Seah HS. Fast geometric complementarity matching in protein-protein docking. In: International Workshop on Advanced Image Technology (IWAIT'14), Bangkok, Thailand, 6–8 January, 2014.
21. Liu H, Lin F, Lee YT, Qian K, Seah HS. Visual analysis with dynamic geometric complementarity and physiochemical match in protein docking. In: 18th International Conference on Information Visualization (iV'14)/11th International Conference on BioMedical Visualization (MediViz'14), Paris, France, 15–18 July, 2014.