ResDock: Protein-Protein Docking Using Shape Complementarity of Surface Residues

Sharon Sunny, Jayaraj P B

Abstract—Searching and scoring are imperative in protein-protein docking. An exceptional searching procedure does not generate the entire gamut of structures, but the most likely structures without compromising the accuracy. The proposed method utilizes shape information of protein surfaces to achieve the same. While existing methods work on set of points uniformly spread over the protein surface, the proposed method acts on a point cloud of surface residues. The identified points are treated upon with the second order partial derivative test to find the curved regions. Instead of orienting concave/convex regions of the receptor with convex/concave regions of ligand, our method aligns concave regions of the receptor with convex regions of ligand and thus reducing the number of poses. The proposed algorithm is tested extensively by using the Protein-Protein Docking Benchmark version 5 and PPI4DOCK dataset to get an overall performance of 76.88% and 93.79% respectively. The proposed method made a commendable improvement in ranks of feasible structures. It is observed that, for majority of the targets, at least one of the feasible solutions lies within top 50 category. Also the method could drastically reduce the conformation space size when compared to state of the art methods without compromising the accuracy.

Index Terms—protein-protein docking, second order partial derivative test, force-field based scoring function, shape descriptors, surface variation, eigenvalue decomposition

1 INTRODUCTION

Proteins are essential macromolecules made up of amino acids. They are classified, based on their function, as catalytic, regulatory, protective, storage, transport, toxic, structural, contractile, secretory, and exotic proteins. Since the interaction between proteins are specific, ill-formed interactions result in diseases like Alzheimer’s disease [1], and Huntigton’s disease. It is important to study the functions of proteins and the nature of interactions as no cure is found yet for the diseases mentioned above. Structure prediction techniques gained momentum due to their significant role in unraveling the functions of biological molecules.

Despite being accurate in structure prediction, experimental methods like X-ray crystallography and NMR spectroscopy are expensive and laborious. Thus the number of protein-protein complex structures available in different databases was significantly less than the number of protein structures. There was a need for alternative methods to speed up the structure prediction process. Computational methods attained their form and substance at this juncture. These methods considerably reduce the effort and speed up the process of protein complex structure prediction. Computational methods for structure prediction is generally categorized as template-based and ab-initio methods. While template-based methods rely on the existing templates in different databases, ab-initio methods work on the coordinate information of atoms in individual proteins. Though template-based methods are easy to implement, their effectiveness depends on the presence of a suitable template [2]. Hybrid methods that combine ab initio and template-based methods are gaining popularity [3] [4].

Ab-initio methods for structure prediction, also known as docking techniques, build the protein complex structures from scratch. It accepts two proteins as input; the larger protein is designated as the receptor and the smaller one as the ligand. The results of docking are poses, which are the structures likely to be generated by their interaction. The main steps involved in this process are creating conformation space and ranking the structures using a scoring function. Existing works tackle each of these stages differently. Monte-Carlo simulation [5], FFT [6], spherical harmonics [7], energy-based techniques [8], and a host of other techniques are applied in the docking domain to get improved results. One can choose among force-field based, knowledge-based, empirical, machine learning-based or a combination of these scoring functions to identify the near-native structures.

Critical Assessment of PRediction of Interactions (CAPRI) is a world-renowned competition to bring out the best minds in protein complex structure prediction. It offers a platform for both predictors and scorers to showcase their expertise. In each round, new target sets are brought in to challenge them. Recently, Critical Assessment of Structure Prediction (CASP) and CAPRI started a joint venture, CAPRI/CASP, to predict the complex structure from sequence information of individual proteins. Many tools convincingly established their effectiveness in complex structure prediction by their performance in different rounds. However, the targets provided may not unveil the Achilles heel of the tools. Desta et al. [9] attempted to expose the ability of docking tools in rigid-body docking using the case study of Cluspro. They observe that the availability of
information on the targets can add to the performance of docking tools.

Katchalski-Katzir et al. [10] pioneered in applying FFT to grid-based methods, which dramatically increased the speed of exhaustive sampling of search space. The main impediments in FFT based methods are the need for a correlation function for expressing energy and the inability to infer the quality of poses based on the values of energy function [9]. Despite these limitations, FFT based methods received a wide acceptance and, later, several works applying FFT were introduced [11] [12] [13] [14] [15]. An alternative to exhaustive methods can be found by exploring the geometric features of the protein surface. Tools like PatchDock [16], SP-Dock [17] align the proteins on the basis of complementary surface patches. Histogram based descriptors are also popular in finding the complementary regions [18] [19]. LZerD [20] uses Zernike Descriptor (ZD), where the terms in the series expansion of a 3D function can represent the shape of the protein surface.

A recent review on the modeling of biomolecular complexes emphasizes the need for an integrative approach in protein-protein docking. Information extracted from predicted/experimental data can help to circumvent the issue of large conformation space in ab initio techniques [21]. The primary sources of information can be mutagenesis, NMR spectroscopy, distance restraints from cross-linking, cryo-EM, SAXS, sequence alignment, and coevolution analysis. Though tools like ATTRACT [22], Haddock [23], RosettaDock explores the integrative modeling in docking, the technique needs to be further refined. Information-driven Haddock webserver recently incorporated MARTINI coarse-grained force field to smoothen the energy landscape [24]. The new approach not only reduced the computation time but also improved the quality of the result. In this work, the docking stage using the Haddock protocol is followed by refinement and a back mapping stage, which restores the all-atom information. It has an extended capability of docking multiple proteins simultaneously.

Lately, Roel-Touris et al. [25] attempted to incorporate Haddock with LightDock to dock membrane-associated proteins. The docking capacity of LightDock and refinement capability of Haddock yielded a better tool. Proper integration of searching strategy and scoring function results in an enhanced performance of docking tools [26]. The refinement stage is an excellent addition to docking. Also, the possibility of getting superior quality structures increases with more predictions.

Since the number of available protein structures in databases counts to millions and newer proteins are getting added to it, there is ample scope for structure prediction of protein-protein complexes, i.e., the difference in physicochemical properties of newly discovered proteins is constantly challenging the docking tools. As the problem is computationally hard, any advancements in docking techniques are appreciable. Improvements in existing systems can be made to reduce the size of conformation space, develop more advanced scoring functions, or refine the generated structures. A proper docking procedure generates and identifies good poses without overcrowding the conformation space. The potential of shape complementarity to accomplish this goal is underutilized hitherto. A mindful approach can dramatically improve the efficiency of computational docking. The proposed method utilizes information about the curvatures of interacting proteins. It identifies local extrema on the surface of proteins using point cloud data obtained from the surface residues. These local extrema act as a seed point for surface patch generation. Demarcating the curved regions pave the way for pose generation. The results of aligning the curved regions are not just near-native structures, but the infeasible structures too. An appropriate scoring function is used to pick out the singular solutions at this stage.

The remainder of the paper is organized as follows: Section 2 describes the design of the proposed method. Section 3 analyzes the experimental results, and section 4 recapitulates the findings and concludes the work.

## 2 Proposed Method

The docking procedure is usually split into the pose generation phase and the scoring phase. While the pose generation phase should ensure that near-native structures are generated, the scoring phase should be well-equipped to identify the best-bet in the conformation space. The proposed method, which is called ResDock, employs shape complementarity for pose generation and force-field based scoring function for ranking the poses. It works in two levels of granularity: initial coarse-grained modeling for finding the transformation parameters, followed by fine-grained all-atom modeling involving pose generation by translation and rotation of the ligand and scoring of the poses. In the rest of the paper, the term 'point' represents the center of residues/atoms in the protein. Fig. 1 shows the workflow of the proposed method. A detailed explanation of each of the steps is given below:

### 2.1 Dataset

ResDock is mainly tested using two datasets. The well-known benchmark set for checking the capability of a protein-protein docking algorithm is Protein-Protein Docking Benchmark set [27] [28]. Different versions of this benchmark set contain experimental structures. According to the difficulty level, the structures in it are classified as rigid-body, medium, and difficult cases. An addition of 55 new cases to the benchmark version 4 yielded benchmark version 5. Moreover, structures available in the PP14DOCK dataset [29] are also used to analyze the proposed method. The structures, which count to 1417, in this dataset are obtained by homology modeling and are classified as very easy, easy, hard, very hard, and super hard based on difficulty level. In the case of both datasets, the unbound structures of individual proteins only are taken for analysis. The algorithm works on cleaned input data and hence adequate preprocessing steps should be carried out before running the program to get good results. PDB files downloaded from RCSB PDB website contain all the data relevant to proteins. Ab initio docking tools prefer ATOM records only and hence other details must be removed from the input files. Make sure that the input files contain only the required and all the required chains and amino acids. All the ‘H’ and ‘HOH’ atoms, and alternate locations of residues
should be removed. The missing atom information must also be fixed before docking.

2.2 Extraction of surface residues

The docking procedure starts by utilizing surface information. The proposed method uses a function, `min_depth()`, in MSMS [30] to identify the surface residues. Given the coordinates of centers of residues, the function identifies those residues having a depth less than a threshold value as surface residues. Further processing on the protein surface is now confined to the resultant set of coordinates. Fig. 2 shows the surface points extracted by existing methods and proposed method for a sample protein. The drastic reduction in the size of the point cloud owes to the residue-wise representation, which in turn reduces the size of conformation space.

2.3 Identification of seed points and segmentation of surface patches

The proposed method aligns complementary regions for generating different poses, i.e., align a hole on the surface of one protein with a knob in the other protein. Now the problem is twofold; getting the surface patches, which are either holes/knobs, and identifying points on these patches, which can usher the alignment procedure.

Holes can be interpreted as concave regions on the surface, while knobs are identified as convex regions. Each of these curved regions contains a point, the local maxima or local minima, around which surface patches can be generated. So the idea is to grow the surface patches from the local extrema. The geometric problem of identifying the local extrema from a point cloud is similar to the mathematical problem of finding the optimum of a function. The second order partial derivative test finds its application in
the docking problem at this point.

Second order partial derivative test and FDM

Second order partial derivatives in multivariable calculus is used in optimization problems to find local extrema. Let \( f(x, y, z) \) be a function of three variables. Then \( f_{xx}, f_{yy}, \) and \( f_{zz} \) represent the second order derivatives of \( f(x, y, z) \) with respect to \( x, y, \) and \( z \) respectively. Let \( D_1 = f_{xx}, D_2 = \frac{f_{xy}}{f_{yx}} \) and \( D_3 = \frac{f_{xx} f_{yx} f_{yy} f_{yz} f_{xz} f_{zy}}{f_{xx} f_{xy} f_{yx} f_{yy} f_{yz} f_{zy}}. \)

Let \( (x_0, y_0, z_0) \) be a point on the surface. Then second derivatives test can be stated as follows [31]:

1) If \( D_1(x_0, y_0, z_0) > 0, D_2(x_0, y_0, z_0) > 0, \) and \( D_3(x_0, y_0, z_0) > 0, \) then \( (x_0, y_0, z_0) \) will be a local minimum of \( f. \)

2) If \( D_1(x_0, y_0, z_0) < 0, D_2(x_0, y_0, z_0) > 0, \) and \( D_3(x_0, y_0, z_0) < 0, \) then \( (x_0, y_0, z_0) \) will be a local maximum of \( f. \)

Since the algorithm works on point cloud data, an approximation method, Finite Differences Method (FD), is used to calculate the derivatives with the available data. If \( f(x) \) is the function to be approximated, then \( \hat{f}'(x) \), the derivative, at a point \( a \) using the standard formula for FD is written as:

\[
\hat{f}'(a) = \lim_{h \to 0} \frac{f(a + h) - f(a)}{h} \tag{1}
\]

The Taylor series expansion supports the mathematical correctness of this approach.

The second-order partial derivative test is utilized in the proposed method to categorically identify the seed points of alignment. Seed points are a set of points on the protein surface around which surface patches are developed. The method discussed in this paper identifies local minima/local maxima as seed points. A local minimum/local maxima, along with its neighbors, defines a concave/convex surface patch on the protein surface.

The detailed procedure for identifying seed points and extraction of surface patches is given in Algorithm 1. The algorithm adopts the interpretation of vector field as fluid flow, and thus the protein’s surface is treated as a flow of points having a velocity and speed. These velocity and speed are utilized to get the normal vector at a point.

The normal vector represents the direction of flow at a point, i.e., a downward-pointing normal vector represents a downhill movement of flow, while an upward-pointing normal vector represents an upward motion. Our method accomplishes the segmentation of points on the surface into different patches by utilizing this information. The algorithm is applied to all points, which are not yet included in any surface patch. Thus the disjoint patches in both receptor and ligand identified using the algorithm can now be the lead for the docking procedure.

2.4 Surface descriptor

Any processing that utilizes the geometric characteristics of a protein needs an appropriate invariant representation of its shape. A surface descriptor can be used for this purpose. Mean curvature, Gaussian curvature, and distance histogram are examples of commonly used descriptors. Some of the preferable features of shape descriptors are explained below:

**Algorithm 1:** Extracting the convex and concave regions on surface

**Input:** Matrix \( M \) containing coordinates of the point and its neighbors within a threshold distance.

**Output:** Seed point together with the set of points in its convex (concave) curve.

1. Find approximate derivatives, \( f_M, f_{M_x}, \) and \( f_{M_z} \) of each point in \( M \) using FDM;
2. Calculate the values of the following terms:\
   3. Velocity, \( f_M = [f_M, f_x, f_z]^T; \)
   4. Speed, \( S = \|f_M\|; \)
   5. Tangent, \( T = \frac{f_M}{S}; \)
   6. \( f_T = [f_T, f_{T_x}, f_{T_z}]^T; \)
   7. Normal \( = \frac{f_T}{\|f_T\|}; \)
8. Differentiate \( f_M \) to get \( f_{M,M_x}, f_{M,M_y}, f_{M,M_z}, \) the second order derivatives of \( M; \)
9. if \( f_{M,M_x} \cdot f_{M,M_y} \cdot f_{M,M_z} \neq 0 \) then
   10. Add the point to the seed point set of concave curvature
11. end
12. else
13. if \( f_{M,M_x} < 0 \) and \( f_{M,M_y} \neq 0 \) and \( f_{M,M_z} \neq 0 \) then
14. Add the point to the seed point set of convex curvature
15. end
16. end
17. Add those points having Normal vector in the direction of seed point to its curvature set;
Translation invariance: Receptor and ligand-protein may exist physically separated. When translating the ligand to orient its surface patch with its receptor counterpart, the corresponding shape descriptors should not change regardless of the coordinates.

Rotation invariance: Alignment of the receptor with the ligand may involve rotational motion of ligand with respect to the receptor. A good shape descriptor should keep its value unchanged during this rotation.

Easiness in computation: A descriptor requiring heavy computation may not fit for the job as proteins may contain tens of thousands of atoms. In this work, it may not be a bottleneck as the center of residues on the surface only is considered for calculations.

Pauly et al. [32] defines geometric properties – normal vector, and surface variation - of point cloud using covariance analysis. Our method calculates shape descriptor of a point as a function of surface variance. The procedure is similar to principal component analysis. Calculate the covariance matrix, $C$, of a point, $p$ as

$$ C = \begin{bmatrix} y_1 - p \top & \cdots & y_1 - p \\ \vdots & \ddots & \vdots \\ y_n - p & \cdots & y_n - p \end{bmatrix} $$ (2)

where $y_1...y_n$ are the $n$ neighbors of $p$. Now perform eigenvalue decomposition of $C$. Let $\lambda_0$, $\lambda_1$, and $\lambda_2$ are the eigenvalues of $C$, where $\lambda_0 < \lambda_1 < \lambda_2$. The surface variation [32], $\sigma_n(p)$, of a point $p$ with $n$ neighbors is defined as follows:

$$ \sigma_n(p) = \frac{\lambda_0}{\lambda_0 + \lambda_1 + \lambda_2} $$ (3)

It is the smallest eigenvalue of the covariance matrix, i.e., the eigenvector corresponding to the smallest eigenvalue is chosen to represent a point in the point-cloud. Eigenvector, being invariant to transformations, can solve the problems with the spatial vector representation. Another advantage of this scheme is that the covariance matrix needs no parameterization. Thus the eigenvector representing the surface variation is chosen as the shape descriptor and is utilized in the generation of conformation space.

2.5 Aligning the surface patches for pose generation

The proposed method uses a simple alignment step involving a pair of surface patches, one from the ligand and the other from the receptor. It needs two pairs of critical points on each patch to make appropriate transformations to the ligand. The seed points themselves can be taken as one pair of critical points. For fixing the other pair, the physical properties of residues that can guide the interaction at the interfaces are examined. Features like polarity and hydrophobicity of the interface residues determine the stability of the interaction between proteins. Since CYS-CYS contacts are significant in protein folding and have a dominant role in interaction, surface patches are first searched for CYS residue’s presence. The absence of such a pair ends up searching for polar residue pairs/hydrogen bond acceptor-donor pairs. If none of these properties are satisfied, points near to the seed points are considered for determining the angle of rotation.

Rotation about a point in 3D space is an isometric transformation involving rotation followed by the translation. This transformation is isometric as it preserves the relative distances between points. The ligand is first rotated around the origin and then translated back to the pivot point. The transformations are given below:

$$ \begin{pmatrix} s_x \\ s_y \\ s_z \end{pmatrix} + R \ast \begin{pmatrix} p_x - s_x \\ p_y - s_y \\ p_z - s_z \end{pmatrix} $$ (4)

where $s$ represent the coordinates of receptor seed point which is the pivot point in this case, $R$ represents the rotation matrix, and $p$ represent the coordinates of points in ligand. Since the rotation is rooted at a seed point, Euler’s rotation theorem is applicable here. According to the theorem, the application of a single rotation matrix along some axis that passes through the fixed point suffice to perform the intended transformation to a rigid body fixed at some point.

To elaborate on $R$, axis-angle representation, which requires a unit vector $v = (v_x,v_y,v_z)$ that represent the axis and an angle $\theta$, is used here. $\theta$ is calculated by utilizing the shape descriptors of points chosen as per the aforementioned physical properties. Let $a$ and $b$ be the shape descriptors of points identified in the neighborhood of seed points of receptor and ligand, respectively. The angle between them is calculated using the following equation:

$$ \cos \theta = \frac{a \ast b}{||a|| \ast ||b||} $$ (5)

Now the operation involving $R$ can be written as

$$ R = \cos \theta \ast I + \sin \theta \ast K + (1 - \cos \theta) \ast K^2 $$ (6)

where $I$ is the identity matrix and $K$ is the cross-product matrix of unit vector $v$. The rotation matrix can be further elaborated as follows:

$$ R = \begin{bmatrix} A & B & C \\ D & E & F \\ G & H & I \end{bmatrix} $$ (7)

where, $A = \cos \theta + v_x^2(1 - \cos \theta)$

$B = v_x v_y (1 - \cos \theta) - v_z \sin \theta$

$C = v_x v_z (1 - \cos \theta) + v_y \sin \theta$

$D = v_y v_z (1 - \cos \theta) + v_x \sin \theta$

$E = \cos \theta + v_y^2(1 - \cos \theta)$

$F = v_y v_z (1 - \cos \theta) - v_x \sin \theta$

$G = v_z v_x (1 - \cos \theta) + v_y \sin \theta$

$H = v_z v_y (1 - \cos \theta) - v_x \sin \theta$

$I = \cos \theta + v_z^2(1 - \cos \theta)$

The end result of this transformation is a structure of protein-protein complex called a pose.

2.6 Scoring

Scoring is the final phase involved in predicting the structure of the complex formed by interaction. The choice of the scoring function is crucial as it must identify the best structure from the set of generated conformations. Several research works on scoring functions are existing; each method is tailor-made for different scenarios.
In the proposed method, major forces driving interaction between proteins – van der Waals and electrostatic potential – are considered at residue-level. The values of these potentials are calculated using meetdock [33]. Van der Waals potential is calculated using the following equation, which is widely known as Lennard Jones potential:

\[ E_{\text{vdw}} = 4 \varepsilon \left( \frac{\sigma}{d} \right)^{12} - \left( \frac{\sigma}{d} \right)^{6} \]  

(8)

where \( \sigma \) is the van der Waals radius, \( \varepsilon \) is the potential well depth, and \( d \) is the distance between residues. Electrostatic is a stronger and longer range potential which is calculated using Coulomb’s equation.

\[ E_{\text{elec}} = \frac{1}{4 \pi \varepsilon_0} \sum q_1 q_2 \frac{1}{\varepsilon(r)d} \]  

(9)

where \( q_1 \) and \( q_2 \) are partial charges of interacting residues, \( \varepsilon_0 \) is the absolute dielectric constant, \( d \) is the distance between the residues, and \( \varepsilon(r) \) is the dielectric function. Total potential is the sum of electrostatic and van der Waals potentials.

\[ E_{\text{total}} = E_{\text{vdw}} + E_{\text{elec}} \]  

(10)

At this stage, some of the poses are filtered-out based on the total potential. Structures having negative potential only are retained as results of execution.

3 RESULTS AND ANALYSIS

As mentioned before, input data for testing the efficiency of the proposed algorithm is taken from Protein-Protein Docking Benchmark version 4 and 5 and the PPI4DOCK dataset. The unbound docking samples in the dataset are more challenging than bound samples as it does not contain any information about conformational changes that can happen during docking [34]. The docking benchmark version 5 contains 55 new complexes in addition to the 123 rigid-body cases, 29 medium difficulty cases, and 24 difficult cases in benchmark version 4. PPI4DOCK benchmark set comprises 261 very_easy structures, 800 easy cases, 222 hard cases, 84 very_hard cases, and 50 super_hard structures.

In the rest of the paper, top1, top5, and top10 predictions may be denoted using T1, T5, and T10 respectively.

3.1 Perfromance measure

In this work, CAPRI criteria is adopted to analyze the correctness and efficiency of the docking algorithm. The main advantage of this criteria is that it rarely overlook good poses by cleverly employing superimposition. Though it makes the calculations complicated, it can deal with any conformational changes occurring in the non-interface region. According to the CAPRI criteria detailed in [35], docking results can be categorized into high, medium, acceptable, and incorrect. This classification is based on the value of three terms: \( f_{\text{nat}} \), \( L-\text{rmsd} \), and \( I-\text{rmsd} \) [36]. \( L-\text{rmsd} \) is the ligand RMSD, which measures the difference in backbone atoms of ligands in a predicted and actual structure after superimposing receptor proteins. \( I-\text{rmsd} \) estimates the difference in backbone atoms present in the interface. A residue in a protein is considered as interface residue if it is within a distance of 10 Å from any atom of its partner protein. \( f_{\text{nat}} \) is the fraction of native residues present in the native structure, reproduced in the predicted structure. Table 1 summarize the CAPRI quality criteria, which classifies structures into different quality classes.

| Quality       | Conditions                                      |
|---------------|-------------------------------------------------|
| High          | \( f_{\text{nat}} \geq 0.5 \) AND \( (L-\text{rmsd} \leq 1.0 \) OR \( I-\text{rmsd} \leq 1.0 \) |
| Medium        | \( f_{\text{nat}} \geq 0.3 \) AND \( f_{\text{nat}} < 0.5 \ ) AND \( (L-\text{rmsd} \leq 5.0 \) OR \( I-\text{rmsd} \leq 2.0 \) \) OR \( f_{\text{nat}} \geq 0.5 \) AND \( L-\text{rmsd} > 1.0 \) AND \( I-\text{rmsd} > 1.0 \) |
| Acceptable    | \( f_{\text{nat}} \geq 0.1 \) AND \( f_{\text{nat}} < 0.3 \) AND \( (L-\text{rmsd} \leq 10.0 \) OR \( I-\text{rmsd} < 4.0 \) OR \( f_{\text{nat}} \geq 0.3 \) AND \( L-\text{rmsd} > 5.0 \) AND \( I-\text{rmsd} > 2.0 \) |
| Incorrect     | \( f_{\text{nat}} < 0.1 \) OR \( (L-\text{rmsd} > 10.0 \) AND \( I-\text{rmsd} > 4.0 \) |

Note: Reprinted from “Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures”, by Méndez, Raúl and Leplae, Raphael and Lemaire, Marc F and Wodak, Shoshana J, 2005, Proteins: Structure, Function, and Bioinformatics, 60, p150-169.
Fig. 3. Analysis of the quality of structures in T1, T5, T10, T20, and T50.

Fig. 4. Analysis of the quality of structures in T10 based on the category of input proteins.

Fig. 5. Analysis of the quality of structures in T30 based on the category of input proteins.

Fig. 6. Performance analysis of ResDock considering the category and difficulty level of targets.

performance of the proposed method in different ranges of ranks. A better analysis, considering the type of proteins and quality of structures generated in top10, is shown in Fig. 4. For enzymes, totaling 87, ResDock generated 8.05%, 22.98%, and 57.47% of high, medium, and acceptable quality structures, respectively. In the total 39 antibodies in the dataset, the rate of total high, medium, and acceptable quality structures obtained are 12.82%, 25.64%, and 38.46%, respectively.

As illustrated in Fig. 5, top30 results are better with 21.83% high quality, 48.27% medium quality, and 70.11% acceptable quality solutions in the case of enzymes. The same trend is followed in the case of antibodies and other targets. Fig. 6 analyses the performance of ResDock considering the category of the complex and its difficulty levels. Among the total 147 rigid-body cases, 117 complexes (79.59%) have acceptable or better solutions. ResDock could obtain CAPRI quality structures for 34 (77.27%) of 44 medium difficulty cases and 22 (64.70%) of 34 difficult cases. It is observed that ResDock could generate a feasible structure for all antibodies in medium difficulty category. At the same time, it succeeded for only 78.15% of antibodies in rigid-body category. A possible reason for the same could be the tight fit between receptor and ligand in rigid-body cases [28].

A more rigorous analysis to confirm the effectiveness of ResDock is carried out using targets in Protein-Protein Docking Benchmark version 2.4. Predicted structures having I-rmsd < 2.5 Å only are considered as feasible in this analysis. Comparison of the proposed method with S-Dock, SPDock, PatchDock, ZDock, LzerD, shDock, F²Dock(S), and F²Dock(S-E) is shown in Fig. 7. Though the success rate of SP-Dock is slightly better than the proposed method, the later beats the former when considering the ranks of feasible solutions. I-rmsd and ranks of the other tools are taken from [17]. Due to the unavailability of Benchmark version 2.4, ligand, receptor, and reference structures for testing the proposed method are taken from Benchmark v 4.0. It is to be noted that no drastic changes are made to these target structures while updating the dataset.

Haddock-CG [24] reported to be highly successful in docking large complexes in Protein-Protein Docking Benchmark version 5. All targets, but antibody-antigen cases, having at least 5000 heavy atoms only, are chosen for analysis. Fig. 8 shows a comparison of Haddock-CG, Cluspro, and ResDock. The integrative model has undeniably outperformed ResDock and Cluspro. In top 1 category itself,
Haddock-CG displayed splendid performance with 51.85% success rate. In the top 10 and above, ResDock showed better performance than Cluspro. Cluspro webserver returns around 30 structures only, and this must be considered in conjunction with its overall performance. Comparison of the same tools based on the complexity class of the targets is given in Supplementary Fig. S3. Haddock-CG performed better in all complexity classes. When Haddock-CG solved the structure of three hard targets, Cluspro and ResDock could do it for two and one hard targets respectively. The overall performance of Haddock in ab initio mode with Cluspro and ResDock is given in Fig. 9. It must be noted that Haddock-CG in ab initio mode returns 10000 structures and all of them are considered in this analysis. While Haddock has a 62.96% (17 out of 27 targets) success rate, Cluspro has 29.62% (8 out of 27 targets), and ResDock has 55.55% (15 out of 27 targets) success rate. The data for this analysis are taken from [24] and [9].

An analysis of the ranks of first acceptable or better structures of HawkDock [38], ATTRACT, and ZDOCK 3.0.2 with ResDock using 52 newly added targets in Protein-Protein Docking Benchmark version 4 reveals that both ATTRACT and HawkDock are successful in predicting near-native structures for 98.07% of targets when ZDOCK and ResDock could generate results for 88.46% and 84.61% of the targets respectively. Nevertheless, ResDock excel in the top 1 to top 100 categories. A success rate of 84.61% in the top 100 is much better than 42.3% of ZDOCK, 51.92% of HawkDock, and 44.23% of ATTRACT. Fig. 10 illustrates these results.

Folding quality of the generated structures are examined using the online tool, RAMPAGE [39]. Ramachandran plot of High quality structure obtained for the target 1AY7 is shown in Fig. 11. It shows that fold quality is excellent as 97% of the residues are in favorable regions.

### 3.3 Comparison with InterEvDock2, and ZDOCK using PPI4DOCK

The outcome of the proposed method is compared with the results of ZDOCK [40], and InterEvDock2 [41] using PPI4DOCK dataset. While ZDOCK is an FFT based techniques, InterEvDock2 utilizes the co-evolutionary information. It is to be noted that structures generated by different tools vary in number. In the analysis, we are interested to check whether the tool can generate a near-native structure or not. The higher the rank of an acceptable structure, the better the tool is. PPI4DOCK supply the result of ZDOCK on the benchmark set in a file named ‘Zdock_results_CAPRI_ranks.txt’. This data is also utilized in analyzing the efficiency of the proposed method. The performance analysis of InterEvDock2 for 812 cases, which
Fig. 10. Analysis of HawkDock, ZDOCK3.0.2, ATTRACT, and ResDock using 52 newly added targets in Protein-Protein Docking Benchmark version 4. All refers to percentage of targets in dataset for which at least one feasible structure is obtained in predictions.

Fig. 11. Ramachandran plot of predicted structure of 1AY7 exclude super-hard and antibody targets in PPI4DOCK, is given in [42]. The same cases are considered in this analysis. Figure 12 shows the graphical analysis of ranks of first acceptable or better solutions for ResDock, ZDOCK, and InterEvDock2 with 3 different scoring functions - InterEvScore, SOAP-PP, and Frodock 2.1. ResDock successfully generated results for 801 of 812 targets (98.64%). While ZDOCK generated feasible results for 772 structures (95.07%), all three variants of InterEvDock2 were successful for 752 targets (92.61%). ResDock outperformed the other two tools when considering top 200 ranks. The success rate of 53.94% of ResDock in top 5 structures is much better than 15.64% of InterEvDock2 with InterEvScore, 16.74% of InterEvDock2 with SOAP-PP, 14.65% of InterEvDock2 with Frodock 2.1, and 11.82% of ZDOCK server. When applying ResDock on all the 1417 targets in the PPI4DOCK dataset, a success rate of 93.79% is accomplished, while ZDOCK ends up with an 85.25% success rate (see Supplementary Fig.S2).

In this paper, best pose of a target refers to the predicted structure that is closest to the crystal structure of a target protein. An investigation on the goodness of best structures of 1417 targets in the dataset reveals that ResDock works pretty well for very_easy targets, as illustrated in Fig. 13. The percentage of high-quality structures in very_easy targets is 88.07% and of medium-quality structures is 11.53%. In the case of easy targets, medium-quality structures dominate. As hardness increases, the quality of the structures diminishes. ResDock was successful in generating a feasible solution for only 30% of super_hard targets.

According to [29], very_easy, easy, and hard targets in the dataset are enough to investigate the efficiency of rigid-body docking techniques. Table 2 shows the list of super_hard targets in the dataset for which acceptable or better solutions are obtained using either ZDOCK or the proposed method. InterEvDock2 excluded this class of targets from the analysis.

### Table 2

| Target   | Proposed Method Rank | ZDOCK Rank | Proposed Method CAPRI Quality | ZDOCK CAPRI Quality |
|----------|----------------------|------------|--------------------------------|--------------------|
| 1yrt_AB  | 10                   | Acceptable | 103                            | Acceptable         |
| 2fo1_CE  | 50                   | Acceptable | 253                            | Acceptable         |
| 3c4f BD  | 41                   | Medium     | 10                             | Medium             |
| 3hls_AB  | 10                   | Medium     | 251                            | Acceptable         |
| 3ohm_AB  | 471                  | Medium     | 103                            | Acceptable         |
| 3gyj_AC  | 250                  | Acceptable | 42                             | Acceptable         |
| 3pvt_BD  | 10                   | Acceptable | 52                             | Acceptable         |
| 3tgx_AB  | 42                   | Acceptable | 471                            | Acceptable         |
| 3uc0_AB  | 52                   | Acceptable | 227                            | Acceptable         |
| 3w8l_AB  | 90                   | Acceptable | 103                            | Acceptable         |
| 4tw1_ED  | 250                  | Acceptable | 52                             | Acceptable         |

3.4 Discussion

The paper describes a new protein-protein docking technique, which utilizes the shape complementarity property of the protein surface. The robustness of the method lies in the accuracy of the second-order derivative test in finding the seed points, the effectiveness of surface variance in describing the shape of the region, and their interplay with the scoring function in finding the appropriate poses. Though no explicit measures for checking geometric clashes are employed in the method, filtering the poses using a scoring function suffice to achieve the same. Van der Waals, being distance-dependent potential, can be used to eliminate unwanted poses. When the distance between atoms is
minimal, the repulsive force gets dominated, and the value of the potential becomes very high. At a decent distance, the attractive and repulsive terms almost cancel each other. Thus one can infer about the geometry of the complex formed using Lennard-Jones potential. In short, evaluating poses generated by shape complementarity based methods using shape-based scoring functions can identify the plausible complex structures.

An attempt to replace the coarse-grained model with an all-atom model revealed the sheer power of ResDock in checking the size of conformation space. The results obtained with all-atom model on a sample set of targets is given in Supplementary Table S2. When all surface atoms are considered, the surface's unevenness is exposed, and the algorithm perhaps identifies more seed points. Consequently, the size of the conformation space will increase and hence the execution time. Coarse-graining will smoothen the protein surface and thus saves time and space. In other words, it helps to avoid trivial intricacies on protein surface which may mislead docking.

The proposed method is compared with Haddock-CG, HawkRank, ATTRACT, S-Dock, SP-Dock, PatchDock, LzerD, ZDOCK, Cluspro, and InterEvDock2 using targets in Protein-Protein Docking Benchmark set and PPI4DOCK dataset. We got comparable or better results on these datasets having unbound structures of individual proteins. The method considerably reduces the size of conformation space as expected. Furtherance on managing conformation space can be achieved by aligning only those regions that agree on their shape and size rather than shape only.

While existing methods match a convex patch of the receptor with a concave patch of ligand and vice versa, the proposed method aligns the concave regions of a receptor with convex regions of a ligand only. This approach potentially reduces the number of poses generated. The approach was checked for correctness by interchanging the receptor and ligand, i.e., poses are generated by aligning the concave region of ligand and convex region of the receptor. In a separate trial, results of both combinations, convex of the receptor with concave of ligand and vice versa, were also checked. The quality and I-rmsd value of targets checked is tabulated in Table 3. In the table, ‘-1’ indicates that no acceptable solutions are generated for the target. The column named RL and LR in this table gives I-rmsd and quality of the best pose generated when aligning concave patch of receptor and convex patch of ligand and vice versa respectively. The last column, ‘both’, shows the result when both LR and RL cases are considered. It is clear from the table that results are at par with the assumptions of the proposed method.

So, as the number of predictions increases, the chance of near-native structure getting selected increases. A different scoring function may improve the performance of proposed method.

The tool is found to be effective for bound docking also. Predicted structure of a sample antigen-antibody complex is given in Supplementary Fig. S5.

3.5 Computational aspects

The program was run on an Intel Xeon CPU E3-2640. The size of the interacting proteins vary, so does the execution time. The running time of the algorithm can be segregated as follows:

1) Time for identifying the seed points and extracting the surface patches,
2) Time for calculating the surface descriptor,
3) Time for pose generation,
4) Time for scoring the poses

To demonstrate the effectiveness of the proposed method in saving execution time and reducing the size of the conformation space, the following scenarios are considered:
A new shape descriptor based on the eigenvector of the covariance matrix is adopted in this method. Transformation on convex regions of ligand to orient with concave regions of the receptor, guided by shape descriptors, yield different poses. Additionally, one-patch matching implemented in the proposed method takes advantage of the residues’ propensity to interact. The method could generate feasible structures for 76.88% of targets in the Protein-Protein Docking Benchmark set and 93.79% of targets in the PPI4DOCK dataset. It could significantly reduce the size of the conformation space. Analysis of the performance proved that docking tools could work with a reduced representation of the protein surface without compromising the accuracy. As an improvement to the proposed method, the size of the patches can be considered in addition to the shape to reduce the size of conformation space. The introduction of parallel algorithms can improve the speed of execution.

Acknowledgment

The authors gratefully recognize the Department of Computer Science, NIT Calicut for providing infrastructure facilities. Sincere gratitude to Central Computing Center, NIT Calicut, for providing remote access to the system even during lockdown period.

References

[1] T. B. Thompson, P. Chaggar, E. Kuhl, A. Goriely, A. D. N. Initiative, et al., “Protein-protein interactions in neurodegenerative diseases: A conspiracy theory,” PLoS computational biology, vol. 16, no. 10, p. e1008267, 2020.
[2] D. Chakravarty, G. McElfresh, P. J. Kundrotas, and I. A. Vakser, “How to choose templates for modeling of protein complexes: Insights from benchmarking template-based docking,” Proteins: Structure, Function, and Bioinformatics, 2020.
[3] S. Vangaveti, T. Vreven, Y. Zhang, and Z. Weng, “Integrating ab initio and template-based algorithms for protein–protein complex structure prediction,” Bioinformatics, vol. 36, no. 3, pp. 751–757, 2020.
[4] Y. Yan, H. Tao, J. He, and S.-Y. Huang, “The hdock server for integrated protein–protein docking,” Nature protocols, vol. 15, no. 5, pp. 1829–1852, 2020.
[5] J. Fernández-Recio, M. Trottov, and R. Abagyan, “Soft protein–protein docking in internal coordinates,” Protein Science, vol. 11, no. 2, pp. 280–291, 2002.
[6] A. Tovchigrechko and I. A. Vakser, “Gramm-x public web server for protein–protein docking,” Nucleic acids research, vol. 34, no. suppl_2, pp. W310–W314, 2006.
[7] V. Venkatraman, L. Sael, and D. Kihara, “Potential for protein surface shape analysis using spherical harmonics and 3d zernike descriptors,” Cell biochemistry and biophysics, vol. 54, no. 1–3, pp. 23–32, 2009.
[8] I. Anishchenko, P. J. Kundrotas, and I. A. Vakser, “contact potential for structure prediction of proteins and protein complexes from potts model,” Biophysical journal, vol. 115, no. 5, pp. 809–821, 2018.
[9] Í. T. Desta, K. A. Porter, B. Xia, D. Kozakov, and S. Vajda, “Performance and its limits in rigid body protein-protein docking,” Structure, Function, and Bioinformatics, 2020.
[10] E. Katchalski-Katzir, I. Sharir, M. Eisenstein, A. A. Friesem, C. Affalo, and I. A. Vakser, “Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques,” Proceedings of the National Academy of Sciences, vol. 89, no. 6, pp. 2195–2199, 1992.
[11] D. Kozakov, R. Brenke, S. R. Comeau, and S. Vajda, “Piper: an fft-based protein docking program with pairwise potentials,” Proteins: Structure, Function, and Bioinformatics, vol. 65, no. 2, pp. 392–406, 2006.
[12] R. Chen, L. Li, and Z. Weng, “Zdock: an initial-stage protein-docking algorithm,” Proteins: Structure, Function, and Bioinformatics, vol. 52, no. 1, pp. 80–87, 2003.
[13] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho, “Cluspro: an automated docking and discrimination method for the prediction of protein complexes,” Bioinformatics, vol. 20, no. 1, pp. 45–50, 2004.
[14] Y. Yan and S.-Y. Huang, “Pushing the accuracy limit of shape complementarity for protein-protein docking,” BMC bioinformatics, vol. 20, no. 25, p. 696, 2019.
[15] C. Zhang and L. Lai, “Sdock: A global protein-protein docking program using stepwise force-field potentials,” *Journal of computational chemistry*, vol. 32, no. 12, pp. 2598–2612, 2011.

[16] D. Schniedman-Duhovny, Y. Inbar, R. Nussinov, and H. J. Wolfon, “Patchdock and symmdock: servers for rigid and symmetric docking,” *Nucleic acids research*, vol. 33, no. suppl_2, pp. W363–W367, 2005.

[17] A. Axenopoulos, P. Daras, G. E. Papadopoulos, and E. N. Houstis, “Sp-dock: protein-protein docking using shape and physicochemical complementarity,” *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 10, no. 1, pp. 135–150, 2012.

[18] S.-Y. Huang, “Exploring the potential of global protein–protein docking using region-based 3d zernike descriptors,” *BMC bioinformatics*, vol. 10, no. 1, p. 407, 2009.

[19] P. Koukos and A. Bonvin, “Integrative modelling of biomolecular complexes,” *Journal of Molecular Biology*, 2019.

[20] V. Venkatraman, Y. D. Yang, L. Sael, and D. Kihara. “Protein-protein docking using region-based 3d zernike descriptors,” *BMC bioinformatics*, vol. 10, no. 1, p. 407, 2009.

[21] S. J. de Vries, C. E. Schindler, I. C. de Beauchêne, and M. Zacharias, “A web interface for easy flexible-protein-protein docking with attract,” *Biophysical journal*, vol. 108, no. 3, pp. 462–465, 2015.

[22] G. Van Zundert, J. Rodrigues, M. Trellet, C. Schmitz, P. Kastritis, E. Karaca, A. Melquiond, M. van Dijk, S. De Vries, and A. Bonvin, “The haddock2. 2 web server: user-friendly integrative modeling of biomolecular complexes,” *Journal of molecular biology*, vol. 428, no. 4, pp. 720–725, 2016.

[23] J. Roel-Touris, C. G. Don, R. V. Honorato, J. P. Rodrigues, and A. M. Bonvin, “Less is more: Coarse-grained integrative modeling of large biomolecular assemblies with haddock,” *Journal of chemical theory and computation*, vol. 15, no. 11, pp. 6358–6367, 2019.

[24] J. Roel-Touris, B. Jiménez-García, and A. M. Bonvin, “Integrative modeling of membrane-associated protein assemblies,” *Nature communications*, vol. 11, no. 1, pp. 1–11, 2020.

[25] S.-Y. Huang, “Exploring the potential of global protein–protein docking: an overview and critical assessment of current programs for automatic ab initio docking,” *Drug Discovery Today*, vol. 20, no. 8, pp. 969–977, 2015.

[26] H. Hwang, T. Vreven, J. Janin, and Z. Weng, “Protein–protein docking benchmark version 4.0,” *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 15, pp. 3111–3114, 2010.

[27] T. Vreven, I. H. Moal, A. Vangone, B. G. Pierce, P. L. Kasprzyk, H. Hwang, T. Vreven, J. Janin, and Z. Weng, “Zdock server: interactive docking prediction of protein–protein complexes and symmetric multimers,” *Bioinformatics*, vol. 30, no. 12, pp. 1771–1773, 2014.

[28] C. Quignot, J. Rey, J. Yu, P. Tuffery, R. Guerois, and J. Andreani, “Intervedock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs,” *Nucleic acids research*, vol. 46, no. W1, pp. W408–W416, 2018.

[29] C. Quignot, J. Rey, J. Yu, P. Tuffery, R. Guerois, and J. Andreani, “812 cases from the ppi4dock database.” <https://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock2/table.html>, last accessed on May 3, 2020.

Sharon Sunny received her B.Tech from University of Kerala and M.Tech from Mahatma Gandhi University Kerala in 2010 and 2013 respectively. She is currently doing PhD in Department of Computer Science and Engineering, National Institute of Technology, Calicut. Her research mainly focus on protein-protein interactions, and structure prediction of protein complexes.

Jayaraj PB received his Ph.D in Computer Science from National Institute of Technology Calicut, India. Presently, he is working as an assistant professor in the Department of Computer Science and engineering, NIT Calicut, India. His research interests include Medical-informatics, Computational Drug Design and GPU Computing. He is author of a great deal of research studies published at national and international journals as well as conference proceedings.

[30] M. F. Sanner, A. J. Olson, and J.-C. Spehner, “Fast and robust protein interaction benchmarks: docking benchmark version 5 and multimeric inputs,” *Nucleic acids research*, vol. 46, no. W1, pp. W408–W416, 2018.

[31] J. H. Hou, “Hawkdock: a web server to predict and analyze the protein–protein complex based on computational docking and mm/gbsa,” *Nucleic acids research*, vol. 47, no. W1, pp. W322–W330, 2019.

[32] S. C. Lovell, J. W. Davis, W. B. Arendall III, P. I. De Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson, “Structure validation by co geometry: φ, ψ and cβ deviation,” *Proteins: Structure, Function, and Bioinformatics*, vol. 50, no. 3, pp. 437–450, 2003.

[33] J. Fernandez-Recio, et al., “Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling,” *Nucleic acids research*, vol. 32, no. suppl_2, pp. W363–W367, 2004.

[34] H. Liu, W. Xiong, X. Yang, J. Zheng, and J. Janin, “Protein–protein docking tested in blind predictions: the haddock2 web server,” *Journal of Molecular Biology*, vol. 450, 2003.

[35] G. Weng, E. Wang, Z. Wang, H. Liu, F. Zhu, D. Li, and T. Hou, “Hawkdock: a web server to predict and analyze the protein–protein complex based on computational docking and mm/gbsa,” *Nucleic acids research*, vol. 47, no. W1, pp. W322–W330, 2019.

[36] R. Méndez, R. Leplae, M. F. Lensink, and S. J. Wodak, “Assessment of capri predictions in rounds 3–5 shows progress in docking procedures,” *Proteins: Structure, Function, and Bioinformatics*, vol. 60, no. 2, pp. 150–169, 2005.

[37] S. Basu and B. Wallner, “Dockp: a quality measure for protein–protein docking models,” *PloS one*, vol. 11, no. 8, 2016.

[38] G. Weng, E. Wang, Z. Wang, H. Liu, F. Zhu, D. Li, and T. Hou, “Hawkdock: a web server to predict and analyze the protein–protein complex based on computational docking and mm/gbsa,” *Nucleic acids research*, vol. 47, no. W1, pp. W322–W330, 2019.

[39] S. C. Lovell, J. W. Davis, W. B. Arendall III, P. I. De Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson, “Structure validation by co geometry: φ, ψ and cβ deviation,” *Proteins: Structure, Function, and Bioinformatics*, vol. 50, no. 3, pp. 437–450, 2003.

[40] B. G. Pierce, K. Wiehe, H. Hwang, B.-H. Kim, T. Vreven, and Z. Weng, “Zdock server: interactive docking prediction of protein–protein complexes and symmetric multimers,” *Bioinformatics*, vol. 30, no. 12, pp. 1771–1773, 2014.

[41] C. Quignot, J. Rey, J. Yu, P. Tuffery, R. Guerois, and J. Andreani, “Intervedock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs,” *Nucleic acids research*, vol. 46, no. W1, pp. W408–W416, 2018.

[42] <https://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock2/table.html>, last accessed on May 3, 2020.