Using Correlated Parameters for Improved Ranking of Protein–Protein Docking Decoys

PRALAY MITRA,[1,2] DEBNATH PAL[1]
[1]*Bioinformatics Centre, Indian Institute of Science, Bangalore, Karnataka 560012, India*
[2]*Supercomputer Education Research Centre, Indian Institute of Science, Bangalore, Karnataka 560012, India*

Received 22 April 2010; Revised 6 July 2010; Accepted 6 August 2010
DOI 10.1002/jcc.21657
Published online 12 October 2010 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: A successful protein–protein docking study culminates in identification of decoys at top ranks with near-native quaternary structures. However, this task remains enigmatic because no generalized scoring functions exist that effectively infer decoys according to the similarity to near-native quaternary structures. Difficulties arise because of the highly irregular nature of the protein surface and the significant variation of the nonbonding and solvation energies based on the chemical composition of the protein–protein interface. In this work, we describe a novel method combining an interface-size filter, a regression model for geometric compatibility (based on two correlated surface and packing parameters), and normalized interaction energy (calculated from correlated nonbonded and solvation energies), to effectively rank decoys from a set of 10,000 decoys. Tests on 30 unbound binary protein–protein complexes show that in 16 cases we can identify at least one decoy in top three ranks having $\leq$10 Å backbone root mean square deviation from true binding geometry. Comparisons with other state-of-art methods confirm the improved ranking power of our method without the use of any experiment-guided restraints, evolutionary information, statistical propensities, or modified interaction energy equations. Tests on 118 less-difficult bound binary protein–protein complexes with $\leq$35% sequence redundancy at the interface showed that in 77% cases, at least 1 in 10,000 decoys were identified with $\leq$5Å backbone root mean square deviation from true geometry at first rank. The work will promote the use of new concepts where correlations among parameters provide more robust scoring models. It will facilitate studies involving molecular interactions, including modeling of large macromolecular assemblies and protein structure prediction.

© 2010 Wiley Periodicals, Inc. J Comput Chem 32: 787–796, 2011

Key words: interaction; quaternary structure; prediction; method; scoring; benchmark; ranking; docking

Introduction

Protein–protein docking involves two major steps: (i) sampling the potential quaternary structures and (ii) scoring to rank them according to the structural proximity to the true quaternary structure geometry. Models generated from quaternary structure sampling are called decoys, some of which may be native-like. Several efficient rigid body methods are currently being used[1–3] for sampling and decoy generation. Similarly, many scoring functions have been developed,[4] including those that use experimental information.[5] A majority of the functions uses physicochemical parameters for scoring, such as surface complementarity (SC), solvation, and nonbonding energy, that are experimentally known to modulate the free energy of binding.

The task of assigning top ranks to decoys having near-native quaternary structure continues to be an enigma. This is because no generalized functions exist that consistently output score values in an order that reflects the actual stability of decoys in physiological condition. Several problems interfere with the consistent scoring of the quaternary structures. (i) The surface of protein is highly irregular[6] because of which there exists a possibility of significant compactness variation at the interfaces. Usually, solvent and ligand molecules fill the interfacial voids, but in most cases, we ignore them in the docking studies. (ii) The nonbonding interaction and solvation energy (SE) at the interface may also vary significantly based on the residue composition at the interface. However, because unnormalized energy values computed from empirical/semi-empirical methods are used for scoring, a small difference in estimation parameters

Additional Supporting Information may be found in the online version of this article.

Correspondence to: D. Pal; e-mail: dpal@serc.iisc.ernet.in

Contract/grant sponsor: Department of Biotechnology, New Delhi, India

© 2010 Wiley Periodicals, Inc.

may permit wide discrepancy of inference between different research groups. This may result in properties like SE suggested to be important by one research group[7] to be declared ineffective by the other.[8] Sophisticated scoring functions are likely to under-perform[9] compared with methods at low resolution.[10]

In this study, we attempted to circumvent some of the issues faced by the current scoring methods by developing a regression model for geometric compatibility and using normalized interaction energies to score and rank decoys. This is achieved by using a set of correlated parameters useful in assessing protein–protein interfaces in quaternary structures. This conceptually simple method performs better than state-of-art methods, such as ZDOCK,[11] FIREDOCK,[12] and GRAMMX[13] without using experiment-guided restraints, evolutionary and statistical information, or sophisticated interaction energy equations. Our study offers a new way for scoring and ranking that may allow for improved and consistent results.

Materials and Methods

Parameter Definition and Calculation

Interface Area

Interface area (IA) is defined as the accessible surface area (ASA) buried on a complex formation. ASA is defined as the area accessible to the solvent (usually water) and estimated by rolling a probe atom over the surface of the protein.[14] To compute IA, the ASA of all the atoms in the individual subunits and that of the complex are calculated. An atom is defined to be an interface atom if it loses its ASA by >0.1 Å^2 on complex formation. The summation of the loss of ASA by all the interface atoms divided by 2 is the IA for the dimer.

Surface Complementarity

SC is an area-based measure to estimate the compactness of the protein interface. At first, a suitable origin-transformation is given to the pair of molecules whose SC is to be computed. A two-dimensional Delaunay tessellation is thereafter applied on the protein subunit surface to describe it in terms of triangular tiles. The distance and angle between the tiles across the two subunit's interface are evaluated (with some corrections to the interface rim regions) to ascertain which of them packed properly. The SC is expressed as the ratio of the minimum of the two "packing" tile areas available from the two subunits and the total tile area of the interface. A value of 0 means no complementarity, whereas a value of 1 indicates perfect complementarity. Full details of the calculations can be found from Mitra and Pal.[15]

Interface Packing

Interface packing (IP) is a volume-based measure for estimating compactness of the protein interface. An envelope covering a 4-Å slice across the interface is first calculated covering all the atoms and interatomic voids. The ratio between the sum of the van der Waals volumes of the atoms enclosed in the envelope and the total volume enclosed in the envelope considering it a sphere gives IP. A value of 0 means no packing at the interface, whereas a value of 1 indicates full packing at the interface. Full details of the calculations can be found from Mitra and Pal.[15]

Nonbonded Energy

Nonbonded energy (NE) is an estimate of the strength of noncovalent interaction between the interface atoms across the protein subunits. For a collection of atoms, NE is defined as the summation over the atom pairs, where i and j belong to two interacting subunits,

$$\mathrm{NE} = \sum_{i<j} \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{4\Pi\varepsilon R_{ij}}\right) \quad (1)$$

where, q_i and q_j, are the partial charges, R_{ij}, the interatomic distance, and A_{ij} and B_{ij} are the constants with respect to the atoms i and j, and ε is the dielectric constant, considered as 1.0.

The first two terms in eq. (1) represent the Lennard-Jones (LJ) potential function estimating the van der Waals interaction and the third term, the Coulomb interaction.[16] The parameters for the LJ potential and the Coulomb potential function are assigned from the Amber united atom force field version 9.[17] We used VMD 1.8.6[18] to incorporate hydrogen atoms, which are needed to be attached to the heavy atom as described in the Amber united atom force field. NE was computed by summing up contribution from each nonbonded interaction between atoms within a sphere radius of 7.0 Å at the interface. The 7.0 Å distance threshold was decided through a data mining study. The NE calculated from the interfaces of both the subunits is averaged to give the final NE. The value of NE linearly scales with the IA; therefore, we divide the NE by the IA to get normalized NE.

Solvation Energy

SE is a measure of entropic contribution to binding free energy estimated from the burial of surface area on complex formation. The calculation of SE using the loss of ASA on complex formation is based on the method of Eisenberg and McLachlan[19]:

$$\mathrm{SE} = \sum \Delta\sigma(\text{Atom type}) \times \Delta\mathrm{ASA} \quad (2)$$

where ΔASA is the loss of ASA by interface atom on complex formation, and the value of $\Delta\sigma$ are atomic solvation parameters for specific atom types (Supporting Information Table S1). $\Delta\sigma$ values reflect the polar and nonpolar character of the atoms. The SE calculated from the interfaces of both the subunits is averaged to give the final SE at the interface. This value of SE linearly scales with the IA; therefore, we divided it by the IA to get normalized SE.

Pseudoscaling

Pseudoscaling is introduced to account for residual steric clashes in the decoys after geometric optimization. Essentially, the clashing atoms are made clash-free before evaluation of NE and SE.

1096987x, 2011, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.21657 by Indian Institute Of Technology, Kharagpur Central Library, Wiley Online Library on [11/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Nonbonded Energy. To a given docking decoy, we introduce hydrogen atoms in accordance to the Amber united atom Version 9 force field. Thereafter, to account for the steric clashes, a translation to all the atoms within 7.0 Å sphere radius across the subunit interface was iteratively applied so that there were no steric clashes with respect to the atom on which the calculations for LJ and Coulomb energy were being made. The direction and magnitude of translation are determined by minimizing the Euclidian norm of the clashing atoms X_i and X_j across the subunits. This is done by solving the equation:

$$\sum_{i,j} || \overrightarrow{X_i} - \overrightarrow{X_j} ||_2 \tag{3.1}$$

subject to the constraints,

$$|| \overrightarrow{X_i} - \overrightarrow{X_j}||_2 \geq r_i + r_j \tag{3.2}$$

where, r_i and r_j are the van der Waals radii of the clashing atoms. The possibilities of disulfide bond (bond length 2.0 Å) between two sulfurs of Cys and salt bridge (bond length 2.8 Å) between Asp/Glu and Arg/Lys were considered separately. The transformed coordinates were used to calculate the pseudoscaled NE using eq. (1).

Solvation Energy. Some corrections are also introduced to the calculation of ASA of the docking decoys to account for the steric clashes (Supporting Information Fig. S1). If the ASA of an atom A is computed when there is a steric clash with atom B, the ratio of the computed ASA with and without the steric clash gives a correction term. We computed all possible steric clashes at interatom increments of 0.1 Å with respect to all possible atoms and compiled a static list of correction terms. For any steric clash at a given distance, the correction term was multiplied to the calculated ASA to derive the actual ASA at the interface without a steric clash. These new ASA values were used to calculate pseudoscaled SE using eq. (2).

Ranking Algorithm

The ranking algorithm is applied on a decoy set available from the programs like FTDOCK[20] or ZDOCK.[11] For our evaluation, ranking is first done on a selected smaller subset based on interface size of the decoys and on all 10,000 decoys available from the mentioned programs. For the smaller select subset, the interface size is chosen after identifying the right-skew region of the interface size distribution of the decoys. We have used restricted side-chain optimization option of the FIREDOCK[12] program to optimize the contact geometry of the side-chain packing in case of unbound docking. Whenever the programs did not return any decoy with near-native structures, the cases were dropped for rank evaluation.

The ranking procedure consists of several consecutive steps (Fig. 1). The four physicochemical parameters, IP, SC, pseudoscaled NE (or NE per unit IA), and SE (or SE per unit IA) are used as input for final ranking. At first, the SC and IP value of the docking decoy is used to calculate a distance SP from an expected SC and IP value, derived from the regression curve

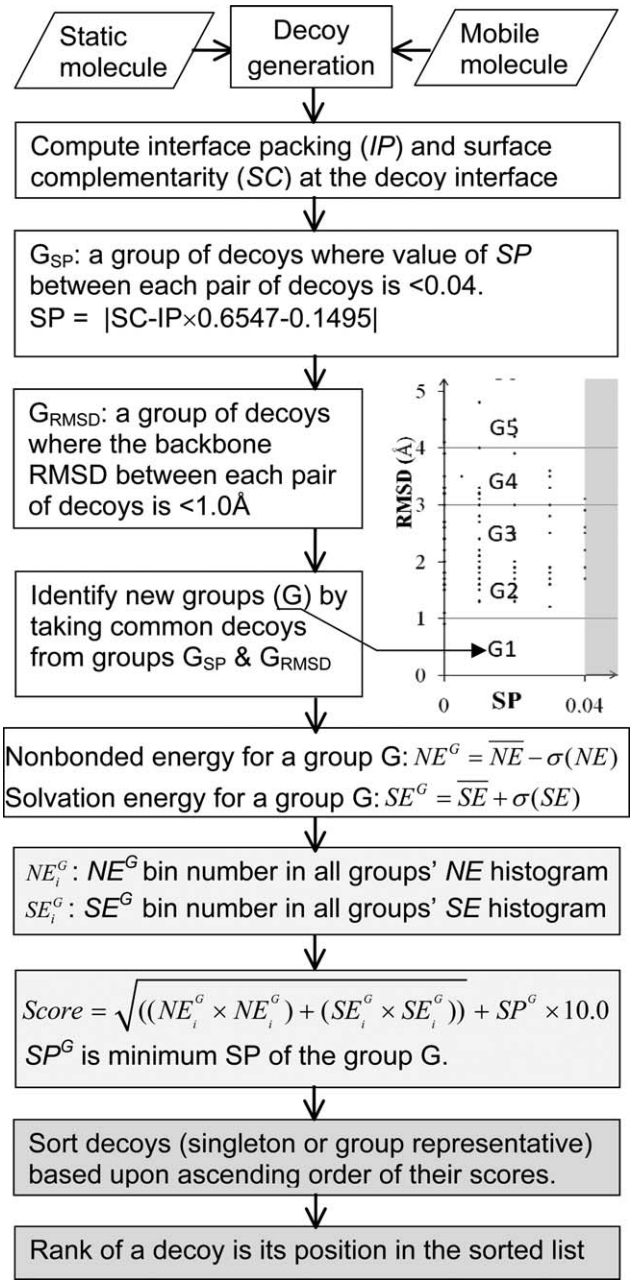

Figure 1. Flowchart summarizing the ranking algorithm. Each step is shown in a square box. A graph of RMSD *versus* SP is included to elaborate the grouping method.

described in Figure 2A. Mathematically, SP (*SC*/*IP*) is: |SC − IP × 0.6547 − 0.1495|. The constants in the equation come from regression fit between SC and IP. The line represents an expected model for geometric compatibility of protein–protein interfaces.

Grouping

The SP value is used to group the docking decoys (G_{SP}) using a threshold value of <0.04. The threshold was determined from a

1096987x, 2011, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.21657 by Indian Institute Of Technology, Kharagpur Central Library, Wiley Online Library on [11/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

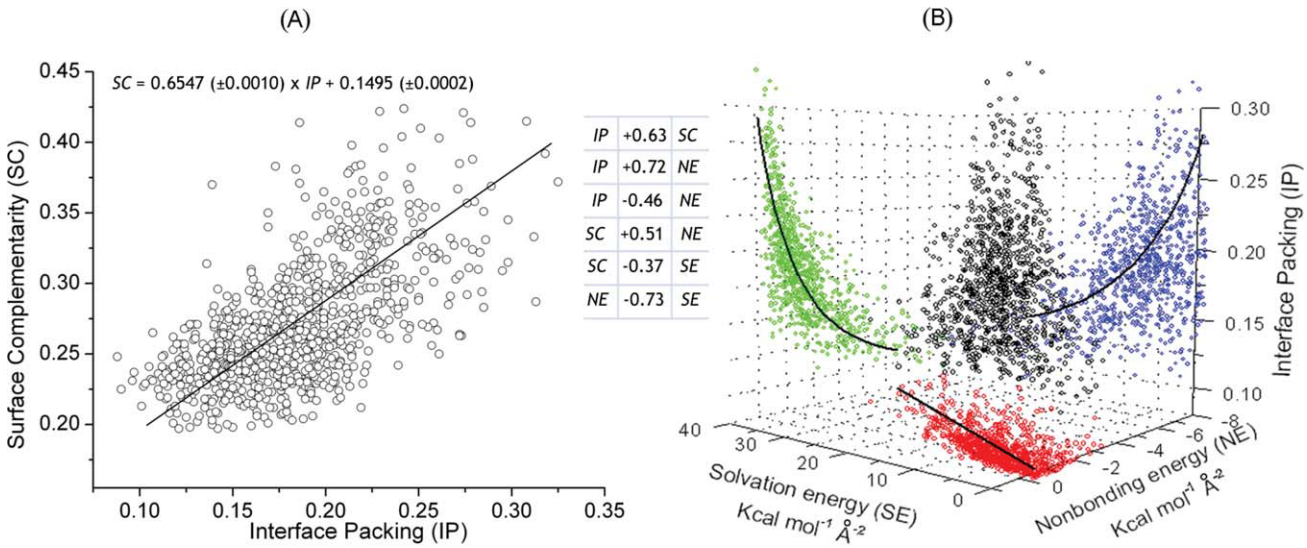

Figure 2. Correlation among the four physicochemical properties at the protein interfaces. (A) A scatter plot showing linear correlation between IP and SC at the protein interface in the training set. The standard deviations of the coefficients of the linear regression curve shown in the figure were calculated using the leave one out test. (B) A three-dimensional scatter plot showing correlation among IP, SE, and NE. The black circles represent the actual data points, and the gray ones represent points projected to the respective two-dimensional planes. Inset: The correlation coefficients obtained between the various parameters plotted in (A) and (B) after removing a few outlier cases.

histogram of SP distances calculated between docking decoy pairs at 1.0 Å root mean square deviation (RMSD), where the said threshold gave 85% coverage with minimum false positives (Supporting Information Fig. S2). A threshold of 1.0 Å RMSD was used because we used it as another criterion for structurally grouping (G_{RMSD}) docking decoys whose quaternary structures were nearly similar. RMSD between a decoy pair is computed between the smaller subunits using the backbone trace after superimposing the backbone of the larger subunits. Docking decoys common to SP-based and RMSD-based groups are selected to construct new groups (G) (see SP *versus* RMSD graph in Fig. 1). Each group is then assigned a representative SP, pseudoscaled NE, and SE value denoted as SP^G, NE^G, and SE^G, respectively. For a group *G*, the representative decoy is one with the lowest SP value among all the group members; its SP value is assigned to the SP^G. NE^G is computed as one standard deviation less than the mean of the entire group members' NE. SE^G is computed as one standard deviation more than the mean of entire group members' SE. Whenever a decoy does not satisfy any grouping criteria, it is taken alone for further scoring.

Scoring

The score of a docking decoy is based on the contribution from SP^G, NE^G, and SE^G. A 5 × 5 matrix is drawn using NE^G and SE^G, which is basically a histogram of NE^G and SE^G with five numbers of bins (Supporting Information Fig. S3). (If per unit area of NE^G or SE^G is used, then the grid is changed to 10 × 10 size.) Because a more stable interaction is denoted by a lower value of NE, and a high SE denotes biologically relevant stable complexes with large interface burial on a complex formation, the matrix is drawn in a way such that from origin along a row NE^G increase, and for column, SE^G decrease. The score is calculated as:

$$\text{Score} = \sqrt{((\text{NE}_i^G \times \text{NE}_i^G) + (\text{SE}_i^G \times \text{SE}_i^G))} + 10.0 \times \text{SP}^G \qquad (4)$$

where NE_i^G is the bin number of NE^G in all groups' NE histogram, and SE_i^G is the bin number of SE^G in all groups' SE histogram. The combined output from the square-root term gives an estimate of normalized interaction energy. We multiply SP^G with 10.0 to give it comparative weight in the final score.

Ranking

The scores are sorted in increasing order of their value. Rank of a decoy or a group of decoys (represented by a group representative) is defined as the position of the complex in this sorted list.

The Data Set

Our data set was constructed from binary interfaces in the asymmetric unit of X-ray crystallography-derived structures from the protein data bank (PDB) solved within resolution and R-factor of 2.5 Å and 0.2, respectively. We restricted ourselves by selecting binary protein complexes with both the individual subunits having at least 25 residues and with no major errors, like chain breaks on the protein surface. The data set was screened for ligand-mediated interaction, and proteins with ligand(s) containing at least five heavy atoms at the dimer interface were dis-

1096987x, 2011, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.21657 by Indian Institute Of Technology, Kharagpur Central Library, Wiley Online Library on [11/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

carded. If any protein had disordered atoms, we kept the one with the highest occupancy (atom labeled A in PDB file). Complexes sharing >90% sequence identity at interfaces in both protein subunits was excluded, and only one representative chosen arbitrarily; this gave us a set of 947 protein dimers of which 828 were homodimers and the rest 119 were heterodimers (Supporting Information PDF 1). To exclude PDB dimers with >90% identical interface, we first identified the atoms at the interface. All residues associated with these atoms were noted and marked on the primary structure. These marked residues were used to build a virtual sequence representing the primary structure at the interface. These sequences were aligned for each subunit pair to identify whether there exists more than 90% sequence identity at the interface.

To screen the binary complexes whether they were biological or not, we used the >800 Å^2 criteria[21] in addition to review of literature on 198 protein complexes with <1000 Å^2. We used PiQSi database[22] (version 2009-9-5) wherever possible; however, the amount of information in the PiQSi database is not as large as that in the Protein Quaternary Structure (PQS) database[23] and was, therefore, unsuited for use as a single source for creation of a large training set. For dimer complexes with IA >1000 Å^2, we used the PQS database to check whether the interfaces are biological. In case a given interface was not found to be biological, we checked the literature and accepted or discarded it accordingly.

The Training Data

Of the 947 dimers in our data set, we retained 829 (773 and 56) dimers as training data (values in the parentheses show number of homodimers and heterodimers). Only 204 complexes in the training set had PiQSi entries. We also classified the data set into fold class based on Structural Classification of Proteins (SCOP) definition.[24] The subunits in the homodimers and heterodimers were distributed between 357 and 30 numbers of SCOP classes (release version 1.75). Training data were used to compute the regression line for SP estimation described in Materials and Methods. It was also used to find the correlation between NE and SE parameters.

Bound Docking Test Set

Two unique dimer data sets at 30% and 35% sequence identity at the interface were created independent of the training set. A total of 25 (9 and 16) and 118 (55 and 63) dimers were present in these two data sets spanning interface ranges 522–6923 and 404–6923 Å^2, respectively (values in parenthesis indicate the number of homodimers and heterodimers). Other screening criteria, such as R-factor, resolution, length of protein, and literature/PQS/PiQSi evidence were same as above. Only 22 complexes could be found in PiQSi database. The subunits in the homodimers and heterodimers were distributed between 54 and 66 numbers of SCOP classes. Before docking calculations, individual subunits were randomized for initial orientation by giving arbitrary rotations and translations. The first 10,000 decoys returned by the FTDOCK program were taken for scoring and ranking.

Unbound Docking Test Set

We culled unbound dimers and their targets from Benchmark 2.0,[25] Benchmark 3.0,[26] Gottschalk et al.,[27] and Bernauer et al..[28] To confirm the oligomeric state, we used information from both PQS[23] and PISA[29] server, aside from literature when additional information was necessary. We did not apply resolution and R-value criteria on the data set, and after excluding structures derived from NMR, or with chain breaks, missing residues or not of minimum chain length 25, we had a total of 32 numbers of test cases of which 26 were unbound–unbound cases and six were unbound–bound cases. The unbound docking partner has been taken from literature wherever available. If the unbound partner information was not present in the literature, then we selected the structure using BLASTP[30] against PDB database using a threshold e-value of 2E-10 and modeled them. This covered 86% of the dimers in Benchmark 3.0,[26] covering rigid body docking category. Two cases PDB:1ACB and PDB:2HQS were discarded from the final list of data set reported because no solution could be obtained from ZDOCK search with acceptable accuracy within 10,000 decoys. This gave us a total of 30 test cases to evaluate our scoring method.

Results

Correlation Among Parameters

A positive correlation of +0.63 is observed between IP and SC (Fig. 2A) allowing us to build a regression model for geometric compatibility. This correlation is further improved to 0.98 if both the parameters are normalized by IA.[15] We use un-normalized SC and IP in our study because normalization offers no specific advantage, IA being a common denominator for SC and IP during decoy evaluation. A negative correlation of −0.73 is observed between nonbonding energy (NE) and SE (Fig. 2B). This allows us to build a 5 × 5 (or 10 × 10) matrix for evaluation of normalized interaction energy. What is also observed from the plot is that geometric compatibility measures IP and SC also have moderate correlation with NE and SE. At low IP and SC values, NE and SE tend to be high. This means that loose IP is compensated by a higher contribution from NE and SE for stable association.

Evaluation Definitions

In our evaluation, we use the canonical accuracy criteria proposed by Mendez et al.[31] Three parameters L_{RMSD}, I_{RMSD}, and F_{NAT} are used for this purpose. The RMSD value between the smaller subunit after superposing a pair of dimers on the larger subunit is called L_{RMSD}; if the RMSD calculation is restricted to interface residues only, then it is called I_{RMSD}. F_{NAT} is the fraction of native contacts at the interface defined as the number of native (correct) residue–residue contacts in the decoy divided by the number of contacts in the true complex. For a true dimer in the test set if there exists at least one decoy in the acceptable accuracy category ($F_{NAT} \geq 0.1$ and (5.0 Å $< L_{RMSD} \leq$ 10.0 Å or 2.0 Å $< I_{RMSD} \leq$ 4.0 Å)), we define it as true prediction (TP); else we define it as no prediction (NP). To evaluate the perform-

1096987x, 2011, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.21657 by Indian Institute Of Technology, Kharagpur Central Library, Wiley Online Library on [11/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

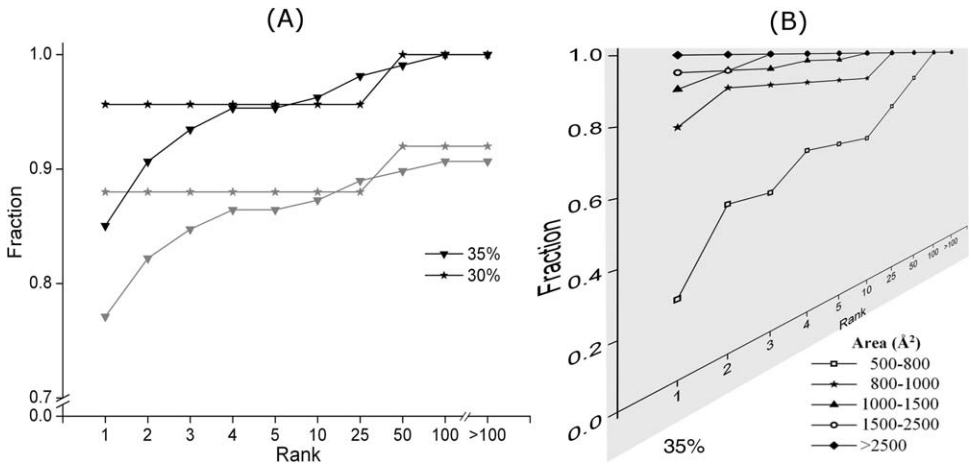

Figure 3. Accuracy of our algorithm at various sequence identity thresholds. (A) Plot showing the variation of accuracy of the algorithm with rank in two different sequence identity-based test sets. The darker curve shows the accuracy where the dimers were prefiltered based on interface size. The light gray curve shows the accuracy over the whole data set using 10,000 decoys. (B) Plot showing the variation of accuracy with rank when the targets are divided into various interface size ranges (prefiltering using interface size performed for these cases). The test set of 30% identity is excluded from the plot owing to few cases.

ance, we used three metrics: rank, coverage, and accuracy. For a true dimer, all docking decoys (singleton or group representatives) are arranged in the ascending order of the score value. Rank is defined as the position of a docking decoy in the sorted list. Coverage is defined to show what fraction of the test set has a true prediction (TP/(TP + NP)). Accuracy is defined as the fraction of true dimers in the test set matched by a docking decoy within a certain rank at acceptable accuracy category.

Bound Docking

The algorithm can assign a large fraction of the top-ranking docking decoys to the true docking geometry. All the results had best ranks with $\leq$5 Å L_{RMSD} after screening, and satisfied the medium accuracy criteria of Mendez.[31] The bound test set of 118 dimers with 35% sequence redundancy at the interface show an accuracy of 77% for identifying decoys nearest to the true dimer at rank 1 (Fig. 3A). The accuracy at rank 1 improves to 85% when only the filtered subset based on interface size is taken; the coverage achieved is around 91%. If we divide the results based on IA ranges (Fig. 3B), for cases with true IA $>$ 1500 Å^2, the accuracy for identifying correct decoy as first rank is >95%. Taking only cases with IA $\geq$ 1000 Å^2, the accuracy at first rank is 93%; if we take first five ranked decoys, the accuracy improves to 99%. The corresponding coverage achieved is >92% and >98%, respectively. For IA $<$ 1000 Å^2, the minimum accuracy is 79% within first five ranks and 100% for within 100 rank; the minimum coverage is 52% and 66%, respectively. The accuracy is relatively poor for <800 Å^2 interfaces (Fig. 3B, □). However, we show in the Example Predictions section that some of these predictions with small interface size may indeed be correct. Raw data on all the predictions may be found in Supporting Information pdf 2. The ranks output by the FTDOCK program alone using the grid-correlation function[1] was able to screen only 37% cases correctly for the true dimer geometry at rank 1.

Example Predictions

We identified interfaces from six protein dimers, which were different from the 539–1305 Å^2 range interfaces predicted by PQS (Table 1). These new interfaces were generated by symmetry operations on the PDB asymmetric unit coordinates. The top-ranked decoys we identified and ranked were within 5 Å L_{RMSD} of these symmetry-generated interfaces. It can be seen that the area of the protein interfaces predicted by our method are larger and has more negative nonbonding energy, suggesting a more stable interface; also the interfaces are more hydrophobic, indicating a better biological association. Literature survey confirmed that the first four proteins listed in Table 1 exist as dimers in solution, whereas there is no publication for the last two (PDB:2IWO and 2O16). PQS proposed 2O16 as monomer, but PISA,[29] an improved scheme for biological dimer detection corroborated all the entries as dimers, as well as their individual quaternary structure predictions.

As a test case we sought experimental evidence on PDB:2O16 as it formed a large compact interface of 1404 Å^2 area. Analytical gel filtration using Agilent HPLC pump to identify the oligomeric state showed the 18-kDa protein to elute at 5.0 min in comparison with two standards: carbonic anhydrase (29 kDa, elution time: 5.5 min) and bovine serum albumin (66 kDa, elution time 4.5 min). (J. M. Sauder, New York SGX Research Center for Structural Genomics (NYSGXRC), personal communication). The standard logarithmic curve of molecular weight to elution volume adjusted for a larger hydrodynamic radius for the protein because of its disk-like shape confirmed its dimeric state in the solution.

1096987x, 2011, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.21657 by Indian Institute Of Technology, Kharagpur Central Library, Wiley Online Library on [11/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1. Example Cases Predicted in This Study with Better Protein Interface Features Compared with the Conventional Interface Available from the PDB.

PDB	Symmetry	Rank (PDB)[a]	Rank (Symm)[b]	Interface (PDB) (Å^2)	Interface (Symm) (Å^2)	L_{RMSD} (Symm) (Å)	ΔSC[c]	ΔIP[c]	ΔNE (kcal/mol)[c]	ΔSE (kcal/mol)[c]
1EX2	−1/2−X,−Y,1/2+Z	–[d]	1	539	1151	1.5	−0.01	0.03	0.53	−1.9
1F9M[e]	1−X,−1/2+Y,−Z	–	3	573	869	4.3	0.00	0.04	0.08	−1.7
1L6R	1/2−X,1/2−Y,−1/2+Z	–	1	718	1292	4.5	0.07	−0.01	0.58	−5.0
2AE2	1−Y,1−X,5/6−Z	3	1	1305	1829	2.8	0.03	0.00	0.5	−3.8
2IWO	1/2−X,−1/2+Y,1−Z	56	1	790	1168	2.1	0.02	0.03	0.52	−4.3
2O16	−X,−1/2+Y,−Z	–	1	830	1403	3.0	0.03	0.01	0.23	−7.9

[a]Ranking done by assuming coordinates in PDB file to represent true quaternary structure.
[b]Ranking done by assuming symmetry-generated complex to represent true quaternary structure.
[c]Δ is the difference between (PDB) and (Symm) associated values.
[d]"–" indicates no prediction.
[e]Engineered thioredoxin F. Wild-type thioredoxin F is a monomer.

Detailed investigation of PDB:1EX2,[32] a conserved *Bacillus subtilis* protein Maf (Fig. 4), provides an evolution and chemical compatibility-based evidence for the correctness of our result. In our prediction, residues Ile 152, Val 160, Ile 163, and Gly 165 are evolutionarily conserved and are at the core area of the interface; of these, Ile 163 and Gly 165 are in contact with each other across the interface. In contrast, the interface of the original PDB, corroborated by PQS databases does not have any conserved residue at the interface. The nature of the interface is largely hydrophilic as highlighted by the red and gray colors, predominantly present at the interface (Fig. 4A). It should also be noted that these colors are not present at complementary position across the interface, indicating a limited physicochemical compatibility. In contrast, our prediction has a large hydrophobic patch consisting of residues Leu 20, Leu 21, Ala 157, Leu 158, Ile 163, and Ile 175, highlighted by yellow in Figure 4B. The IA is also substantially large and includes two aromatic residues Tyr 167 and Tyr 168, interacting across the interface forming an aromatic cluster (shown in green color). The colors are evenly arranged across the interface, indicating good chemical compatibility. These features were found to hold for other dimer entries in Table 1 as well.

Unbound Docking

The unbound test proteins present a more difficult case of predictive docking. Our test set consisted of two parts, unbound–unbound and unbound–bound cases, numbering 24 and six cases, respectively (Table 2; see Supporting Information Table S2 for more details). Of the total 30 cases, we have been able to predict at <100 ranks for 25 cases (i.e., ~83%) using an interface size filter in a prescreening step. Of these, in seven cases we predict correctly at rank 1, one at rank 2, and six at rank 3, totaling 14 (47%) predictions within top 3. If we look at the statistics for top 10 predictions, then in 17 of 30 cases give 57% coverage. In 5 of 30 cases, where we could not achieve a <100 rank after using interface size based filter, if we screen on 10,000 decoys, we find in two cases, we get a rank in top 20,

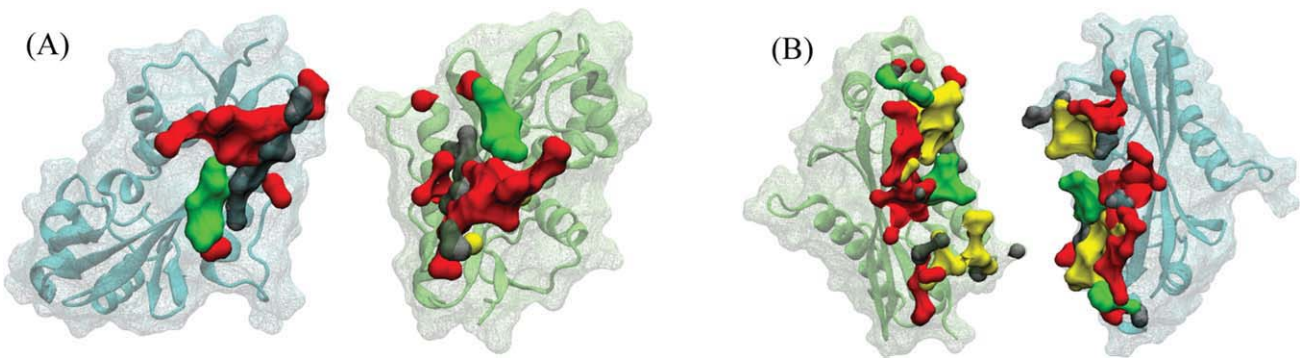

Figure 4. Residue property at the interface of the protein PDB:1EX2. The interacting subunits have been rotated out of plane for better visualization. The red color represents the charged residues (Asp, Glu, Lys, Arg, and His); the green color represents the aromatic residues (Phe, Trp, and Tyr); the yellow color represents the hydrophobic residues (Ala, Ile, Leu, Met, and Val), and the rest are gray (Pro, Gly, Ser, Cys, Thr, Asn, and Gln). (A) The conventional interface from the original PDB file, corroborated by PQS database. (B) The interface from our prediction. Please note the enhanced compatibility of the color distribution at the larger interface in our prediction.

1096987x, 2011, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.21657 by Indian Institute Of Technology, Kharagpur Central Library, Wiley Online Library on [11/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2. Ranking of Best Near-Native Quaternary Structure Output by Various Predictive Docking Methods.

	Rank[b]					
	Our method					
PDB[a] (target IA) (Å^2)	With interface size filter	No filter	ZDOCK	GRAMMX	PATCHDOCK	FIREDOCK
Unbound–unbound cases (24)						
1AVXAB (792)	>100	17	1	>100	*>100*	*1*
1BRCEI (657)	60	3	>100	>100	*17*	*1*
1BUHAB (662)	>100	>100	7	>100	>100	>100
1BVNPT (1111)	3	>100	1	22	*74*	*6*
1CGIEI (1026)	3	43	33	15	*>100*	*10*
1CSEEI (744)	1	>100	>100	7	*>100*	*2*
1DFJEI (1291)	3	>100	1	>100	>100	>100
1J2JAB (605)	20	1	1	56	>100	26
1JLTAB (1426)	3	>100	2	1	18	5
1K9OIE (914)	>100	43	>100	>100	>100	>100
1KXPAD (1670)	9	100	1	7	15	1
1LW6EI (853)	1	10	1	1	–	–
1MA9AB (1786)	14	>100	4	>100	>100	>100
1PPEEI (844)	1	27	1	1	*3*	*1*
1R0REI (704)	51	>100	>100	5	>100	9
1UDIEI (1011)	1	>100	69	19	*8*	*5*
1UGHEI (1096)	10	>100	14	6	*42*	*1*
1XD3AB (1140)	3	>100	7	43	>100	>100
1YVBAI (871)	70	>100	5	4	>100	2
1Z0KAB (893)	>100	12	3	58	>100	1
2BTFAP (1031)	46	>100	>100	>100	>100	>100
2HLEAB (1058)	2	>100	39	81	>100	31
2SNIEI (814)	49	64	7	3	*>100*	*13*
2UUYAB (640)	>100	>100	>100	>100	>100	>100
Unbound–Bound cases (6)						
1CT0EI (523)	67	>100	>100	>100	39	>100
1E44BA (1268)	1	71	1	1	1	1
1JTDBA (1090)	3	74	>100	>100	6	21
1JTGCB (1302)	1	>100	1	1	20	3
1STFEI (895)	1	6	5	1	*5*	*1*
2TECEI (780)	6	>100	>100	1	*26*	*1*

[a]The four-letter PDB code followed by the chain identifier (underlined).
[b]Ranking is done based on acceptable accuracy according to Mendez definition.[31] Cases italicized are directly taken from literature.[12]

and one in top 50, the remaining two cannot be improved below <100 rank. If we take a union of the best results with or without using the interface size filter, we are able to identify 16 correct decoys at the top 3 rank in 53% cases, and top 10 rank in 60% cases. These are indeed significantly improved ranking compared with other methods (see later).

The ZDOCK, GRAMMX, PATHDOCK, and FIREDOCK rankings show that of the 30 cases, ZDOCK has 11 top 3 ranks, GRAMMX 8 top 3 ranks, PATCHDOCK 2 top 3, and FIREDOCK 11 top 3 ranks. These are lower than 14/16 top 3 ranks achieved by our ranking method. Also the number of cases with >100 ranks are 8 (26%) for ZDOCK, 10 (33%) for GRAMMX, 16 (53%) for PATCHDOCK, 8 (26%) for FIREDOCK compared with 5 (16%) for our method when an interface filter is used in ranking, and only 2 (7%) when results from with or without filter are combined. Interestingly, in most cases, when one of the other programs gave a good answer, our ranking has also worked effectively. In addition, our method has worked well for cases where none of the other methods gave superior ranking. For example, in PDB:2HLE, our rank is 2 whereas ZDOCK rank is 39, GRAMMX: 81, PATCHDOCK: >100, and FIREDOCK: 31. If we look closely at the docking geometry output by all the programs, they dock on the same site with the same orientation (Supporting Information Fig. S4). The difference is how we score the geometric compatibility and the interaction energy. The loop and the β-strand regions on the surface ensure the surface compatibility; but the binding surface is not equally strong in electrostatic potential. Although one of the subunit interfaces has moderately positive charge, the charge on the interface from other subunit is highly negative.

It is known that statistical potentials are particularly powerful, when they are custom built based on sufficient prior infor-

1096987x, 2011, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.21657 by Indian Institute Of Technology, Kharagpur Central Library, Wiley Online Library on [11/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

mation on the interacting proteins. Therefore, we additionally checked how RPScore[33] and Pair Score[34] performed on our data set (Supporting Information S3). When RPScore was applied to our 30 unbound targets with interface size filtered decoy sets, we got 17% at the top 3 rank, 33% at top 10, and 67% at top 100; evaluating the same with full 10,000 decoys, we got 27% at top 3, 37% at top 10, and 50% at top 100. It is evident that RPScore ranks equally well on interface size filtered set and unfiltered decoy set. Similar evaluation with Pair Score on the unbound target set with the interface-filtered decoys showed only 0, 20, and 60% cases at top 3, top 10, and top 100 cases, respectively. Evaluation with 10,000 decoys returned a weak 13, 13, and 23% cases at top 3, top 10, and top 100 cases, respectively. But these were not significant against the 53% top 3 ranking achieved by our method. We tried to see whether attaching the RPScore and Pair Score to our scoring method would improve the ranking efficiency of our results. Testing our RPScore-attached scoring function on 21 cases where interface size filtered top 50 ranked decoys contained at least one acceptable solution, we obtained improved ranking in 33% cases, 19% had no change, and 48% returned worse results; Pair Score returned improved ranks in 14% cases, another 14% had no change, and in 71% cases ranks worsened (Supporting Information Table S4). Interestingly, among all the cases that showed improvement in rank, except one, all had native structure interface size >1000 Å^2. It is known that discriminating large interfaces as biological is easier than the smaller ones. This suggests that general statistical potentials are of limited use when incorporated in our scoring scheme or directly applied to our data set.

Random Evaluation of Bound Targets

Predictive docking of unbound complexes requires structural refinement of the decoys because side-chain and main-chain conformations change after the complex is formed. Although a structural refinement is performed, it rarely reconstructs the exact conformation (both side- and main-chain) seen in the bound target complex. We, therefore, compared the docking and ranking on the bound targets to assess how our method would perform if we had the opportunity to obtain unrefined decoys directly from the subunits of the bound targets. Table 3 shows ranking for 10 arbitrarily chosen targets from 30 entries in Table 2. We find that, in all cases, the rank is 1 except one, where it is 3. What is most interesting is that 1K9O, which could be predicted with 43 rank in unbound targets, can now be predicted at rank 1. This indicates a very critical role of docking sampling and structural refinement in effective ranking of the decoys. What is also notable is that the F_{NAT} values recovered are close to 1.0, which is rarely in unbound docking.

Table 3. Evaluation of Randomly Selected Bound Complexes from Table 2.

PDB	Interface area (Å^2) of decoy	L_{RMSD} (Å)	I_{RMSD} (Å)	F_{NAT}	Rank[a]
1AVX	992	1.7	1.6	0.9	3
1BVN	1246	1.3	1.3	0.9	1
1CGI	1098	1.7	1.4	1.0	1
1DFJ	1531	1.5	1.4	0.9	1
1K9O	*996*	*2.1*	*1.6*	*1.0*	*1*
1LW6	913	1.7	1.6	0.9	1
1MA9	1974	1.9	2.1	0.8	1
1PPE	901	1.7	1.6	0.9	1
1STF	946	2.1	1.7	1.0	1
2SNI	924	1.6	1.6	0.9	1

Cases in italics are results obtained without interface size filter.
[a]Ranking is done based on acceptable accuracy according to Mendez definition.[31]

Discussion

Our work focuses specifically on ranking docking decoys available from docking search methodologies (namely FTDOCK and ZDOCK used in this study). Our method is simple in which we combine estimates from a regression model for geometric compatibility and use normalized interaction energy to score the decoys to rank them effectively. More specifically, the regression line shown in Fig. 2A gives the expected values for geometric compatibility, whereas the 5 × 5 matrix of NE and SE (10 × 10 in case on normalized NE and SE) gives a reference to calculate normalized NE and SE values. We are not aware of any other ranking algorithm, which works on a combination of regression model and normalized interaction energy to score and rank decoys.

Some parts of our parameter estimation have commonality with other ranking algorithms, but they differ significantly in how they are implemented. For example, in SOFTdock[8] the SC is determined using Voronoi tessellation, compared with Delaunay tessellation in our work.[15] In Delaunay tessellation, the atoms represent the vertex of the surface triangulation, whereas, in Voronoi the midpoint between the atoms represents a vertex of the triangles. As a result, in the former, the subtle variation in the surface is better captured and probably aids a better estimate of SC. It has already been shown that our method can better discriminate for biological dimers.[15]

The concept of softdocking has been tried elsewhere,[2] but we adapt this for our work in a unique way. We use SP and RMSD measures individually to band docking solutions together into groups and used their intersection to obtain group-based parameters (SP^G, NE^G, SE^G) so that we do not work with original values but representatives and standard deviations (Fig. 1). Conventionally, softdocking has been described in cases where researchers have assumed, for example, the receptor surface as "soft,"[2] or allowed for penetration of the receptor surface during docking search.[35] Technically speaking, having a rigid impenetrable receptor surface and a soft scoring function, or a soft penetrable receptor surface and a hard scoring function allow the same "softdocking" in terms of the final result.

Our method has some primary difference from others because we use the original forms of the LJ, Coulomb, and SE functions to calculate the interaction energy. Our scoring does not incorporate any evolutionary information or other information (especially restraints) from experiments or knowledge bases. Many researchers, to improve the performance of their method, have

1096987x, 2011, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.21657 by Indian Institute Of Technology, Kharagpur Central Library, Wiley Online Library on [11/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

used experiment-guided restraints,[36] evolutionary information,[37] statistical propensities,[33] and also introduced modifications in the interaction energy equations to improve docking performance.[12,13,38] The novelty of our work is how we designed our ranking method exploiting the correlated parameters. In addition, the optimal values of parameters were chosen after extensive goodness-of-fit tests on our training set (Supporting Information PDF 3). The comparative results from Tables 2 and 3 also indicated that proper decoy sampling and structural refinement is of great importance. Inclusion of explicit solvent in refinement of docked structure can also improve binding geometry toward near-native contacts[39] and facilitate improved assessment of geometric compatibility and binding affinity.

Conclusion

We presented a conceptually new method for ranking decoys available from docking search algorithms. We exploit correlation among four physicochemical parameters, IP, SC, NE, and SE to build a regression model for geometric compatibility and obtain normalized interaction energy to compute a score to effectively rank the decoys according to their similarity to near-native quaternary structure geometry. The benchmarks show that the performance of our algorithm is superior to existing ranking methods despite its simplistic implementation. We can successfully rank targets where other methods are not so successful. Our approach has opened new possibilities in improvement of ranking decoys.

Acknowledgments

P.M. thanks All India Council for Technical Education, New Delhi, for the national doctoral fellowship. There is no conflict of interest in publication of this research study.

References

1. Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. Proc Natl Acad Sci U S A 1992, 89, 2195.
2. Palma, P. N.; Krippahl, L.; Wampler, J. E.; Moura, J. J. Proteins 2000, 39, 372.
3. Ritchie, D. W.; Kemp, G. J. Proteins 2000, 39, 178.
4. Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Proteins 2002, 47, 409.
5. Zhang, C.; Liu, S.; Zhou, Y. Proteins 2005, 60, 314.
6. Pettit, F. K.; Bowie, J. U. J Mol Biol 1999, 285, 1377.
7. Fernandez-Recio, J.; Abagyan, R.; Totrov, M. Proteins 2005, 60, 308.
8. Li, N.; Sun, Z.; Jiang, F. Proteins 2007, 69, 801.
9. Heifetz, A.; Katchalski-Katzir, E.; Eisenstein, M. Protein Sci 2002, 11, 571.
10. Vakser, I. A.; Matar, O. G.; Lam, C. F. Proc Natl Acad Sci U S A 1999, 96, 8477.
11. Wiehe, K.; Peterson, M. W.; Pierce, B.; Mintseris, J.; Weng, Z. Methods Mol Biol 2008, 413, 283.
12. Andrusier, N.; Nussinov, R.; Wolfson, H. J. Proteins 2007, 69, 139.
13. Tovchigrechko, A.; Vakser, I. A. Proteins 2005, 60, 296.
14. Lee, B.; Richards, F. M. J Mol Biol 1971, 55, 379.
15. Mitra, P.; Pal, D. FEBS Lett 2010, 584, 1163.
16. Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E. I.; De Bolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. A. Comp Phys Commun 1995, 91, 1.
17. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. J.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. J Am Chem Soc 1995, 117, 5179.
18. Humphrey, W.; Dalke, A.; Schulten, K. J Mol Graph 1996, 14, 33.
19. Eisenberg, D.; McLachlan, A. D. Nature 1986, 319, 199.
20. Gabb, H. A.; Jackson, R. M.; Sternberg, M. J. J Mol Biol 1997, 272, 106.
21. Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin, J. J Mol Biol 2004, 336, 943.
22. Levy, E. D. Structure 2007, 15, 1364.
23. Henrick, K.; Thornton, J. M. Trends Biochem Sci 1998, 23, 358.
24. Andreeva, A.; Howorth, D.; Chandonia, J. M.; Brenner, S. E.; Hubbard, T. J.; Chothia, C.; Murzin, A. G. Nucleic Acids Res 2008, 36(Database issue), D419.
25. Mintseris, J.; Wiehe, K.; Pierce, B.; Anderson, R.; Chen, R.; Janin, J.; Weng, Z. Proteins 2005, 60, 214.
26. Hwang, H.; Pierce, B.; Mintseris, J.; Janin, J.; Weng, Z. Proteins 2008, 73, 705.
27. Gottschalk, K. E.; Neuvirth, H.; Schreiber, G. Protein Eng Des Sel 2004, 17, 183.
28. Bernauer, J.; Aze, J.; Janin, J.; Poupon, A. Bioinformatics 2007, 23, 555.
29. Krissinel, E.; Henrick, K. J Mol Biol 2007, 372, 774.
30. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. J Mol Biol 1990, 215, 403.
31. Mendez, R.; Leplae, R.; De Maria, L.; Wodak, S. J. Proteins 2003, 52, 51.
32. Minasov, G.; Teplova, M.; Stewart, G. C.; Koonin, E. V.; Anderson, W. F.; Egli, M. Proc Natl Acad Sci U S A 2000, 97, 6328.
33. Moont, G.; Gabb, H. A.; Sternberg, M. J. Proteins 1999, 35, 364.
34. Ponstingl, H.; Henrick, K.; Thornton, J. M. Proteins 2000, 41, 47.
35. Del Carpio Munoz, C. A.; Peissker, T.; Yoshimori, A.; Ichiishi, E. Genome Inform 2003, 14, 238.
36. Dominguez, C.; Boelens, R.; Bonvin, A. M. J Am Chem Soc 2003, 125, 1731.
37. Tress, M.; de Juan, D.; Grana, O.; Gomez, M. J.; Gomez-Puertas, P.; Gonzalez, J. M.; Lopez, G.; Valencia, A. Proteins 2005, 60, 275.
38. Pierce, B.; Weng, Z. Proteins 2007, 67, 1078.
39. Krol, M.; Chaleil, R. A.; Tournier, A. L.; Bates, P. A. Proteins 2007, 69, 750.

1096987x, 2011, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.21657 by Indian Institute Of Technology, Kharagpur Central Library, Wiley Online Library on [11/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License