Docking Unbound Proteins Using Shape Complementarity, Desolvation, and Electrostatics

Rong Chen¹ and Zhiping Weng²*³

¹Bioinformatics Program, Boston University, Boston, Massachusetts
²Department of Biomedical Engineering and Bioinformatics Program, Boston University, Boston, Massachusetts

Abstract A comprehensive docking study was performed on 27 distinct protein-protein complexes. For 13 test systems, docking was performed with the unbound X-ray structures of both the receptor and the ligand. For the remaining systems, the unbound X-ray structure of only molecule was available; therefore the bound structure for the other molecule was used. Our method optimizes desolvation, shape complementarity, and electrostatics using a Fast Fourier Transform algorithm. A global search in the rotational and translational space without any knowledge of the binding sites was performed for all proteins except nine antibodies recognizing antigens. For these antibodies, we docked their well-characterized binding site—the complementarity-determining region defined without information of the antigen—to the entire surface of the antigen. For 24 systems, we were able to find near-native ligand orientations (interface Cα root mean square deviation less than 2.5 Å from the crystal complex) among the top 2,000 choices. For three systems, our algorithm could identify the correct complex structure unambiguously. For 13 other complexes, we either ranked a near-native structure in the top 20 or obtained 20 or more near-native structures in the top 2,000 or both. The key feature of our algorithm is the use of target functions that are highly tolerant to conformational changes upon binding. If combined with a post-processing method, our algorithm may provide a general solution to the unbound docking problem. Our program, called ZDOCK, is freely available to academic users (http://zlab.bu.edu/~rong/dock/). Proteins 2002;47:281–294. © 2002 Wiley-Liss, Inc.

Introduction The biological activities of many proteins depend on the specific recognition of one or more partner proteins. Enzyme/inhibitor, antibody/antigen, hormone/hormone receptor, and protein kinase/substrate are examples of protein-protein complexes. Recent developments in proteomics technologies such as mass spectroscopy, genome-scale yeast 2-hybrid, and display cloning are uncovering many novel protein-protein interactions.¹⁻⁴ In the meantime, structural genomics⁵ has considerably accelerated the experimental determination of protein structures. As of June 19, 2001, there are 15,435 entries in the Protein Data Bank (PDB⁶). However, the determination of protein-protein complex structures remains difficult with a few hundred deposited coordinate sets, most of which are from a highly limited variety of proteins. Thus, computational determination of complex structures by docking studies will continue to be an invaluable tool, not only for the insight they yield on the thermodynamics of molecular recognition, but also for their potential utility in understanding protein interaction networks in a cell.

There has been a wealth of research on protein-protein docking, described in several reviews.⁷⁻¹¹ In the simplest situation, the input receptor and ligand structures are taken directly from the complex and reassembled. This is frequently referred to as the bound docking problem, and excellent results have been obtained for a large number of test cases.¹²⁻¹⁵ In reality, proteins undergo conformational changes upon complex formation. Even though structural flexibility is mostly restricted to surface side chains,¹⁶ it makes docking unbound molecules extremely difficult. This is referred to as the unbound docking problem. The algorithm by Abagyan and colleagues¹⁷,¹⁸ allows surface side chain flexibility; however, such algorithms typically require hundreds of hours for each complex,¹⁷ which is not feasible for genome-scale applications. The alternative is to adopt the rigid-body approach, which only searches the six rotational/translational degrees of freedom, but uses target functions that are tolerant to some overlap of the two proteins being matched. This approach is called soft docking.

The large number of novel protein-protein interactions uncovered by proteomics techniques demands unbound docking algorithms that do not require binding site information. A few algorithms are capable of searching the entire 6-dimensional (6-D) rotational and translational space of the ligand. Notably, Fast Fourier Transform (FFT) based methods¹⁹⁻²² performed well in two blind trials.¹⁸,²³²⁴ Others include an algorithm that matches surface cubes,²⁵ a computer-vision-based algorithm,²⁶ a

Abbreviations: ACE, atomic contact energy (a desolvation free energy measure); CDR, complementarity determining region; FFT, Fast Fourier Transform; RMSD, root mean square deviation.

Grant sponsor: NSF; Grant numbers: DBI-9806002 and DBI-0078194.
*Correspondence to: Zhiping Weng, Department of Biomedical Engineering and Bioinformatics Program, Boston University, 44 Cummington Street, Boston, MA 02215. E-mail: zhiping@bu.edu

Received 6 July 2001; Accepted 6 December 2001
method based on Boolean operations, and a method based on spherical polar Fourier correlations. A recent paper uses a Genetic Algorithm to match protein surface dots.

In this paper, we focus on the initial stage of unbound docking with minimal assumption of the binding site. Since it is computationally infeasible to perform a full 6-D search with explicit surface side chain flexibility, we have chosen a soft-docking approach. Our goal is to rank near-native structures within the top 1,000 orientations for a wide variety of complexes without any manual intervention. This subset of structures can then be subjected to further filtering using more complete binding free energy functions and side chain conformational searches.

A number of groups have also investigated the initial stage of unbound docking. Jiang and Kim developed an algorithm based on matching surface cubes and counting the number of favorable and unfavorable pairwise interactions. They tested two unbound complexes and were able to retain near-native structures for both complexes in the top few thousand possibilities. Gabb et al. applied an FFT algorithm with a steric-electrostatic-combined target function on ten unbound complexes, and were able to rank tens of near-native structures in the top 4,000 for nine of them without knowing the binding site. Norel et al. applied a computer vision-based algorithm to unbound docking. Their test set encompassed six distinct unbound complexes, and for all of them near-native orientations were obtained in the top 1,000 using a target function that only accounted for shape complementarity. Palma et al. proposed a steric docking algorithm based on Boolean operations followed by filtering using electrostatic, solvation, and residue contact energies. They tested the algorithm on 19 unbound complexes; for 16 of them the algorithm was able to rank near-native structures in the top 1,000 without any knowledge of the binding site. Ritchie and Kemp introduced an algorithm based on spherical polar Fourier correlations and used a target function that was a linear combination of electrostatic, steric, and hydrophobic energies. They applied the algorithm to 18 unbound complexes and successfully found near-native structures in the top thousands for 13 complexes, when the search was constrained to the receptor binding site. Unfortunately, they did not report any results on a full 6-D search although the algorithm was clearly fast enough to do so. Mandell et al. studied three distinct unbound complexes using an FFT-based method with a target function that was the sum of a Poisson-Boltzmann electrostatics energy and a van der Waals energy. The composite target function identified 1 to 15 near-native structures in the top 500.

The above studies have collectively shown that some algorithms can retain near-native structures in the top few thousands for some protein-protein complexes. However, it is not clear if one algorithm would work for all protein complexes. The other goal of the initial stage of unbound docking is to retain as many near-native structures as possible, since this greatly affects the success rate of further filtering using more time-consuming algorithms. In the above studies, the numbers of retained near-native structures have been small.

In this paper, we have compiled all distinct unbound complexes tested by the above studies to form a large dataset for a comprehensive analysis. The dataset, in the order of increasing difficulty, includes: 5 homodimers, 11 enzyme/inhibitor, 9 antibody/antigen, and 2 others. We have chosen to use an FFT-based method, since it is mathematically elegant, computationally efficient, and capable of optimizing any target function that can be expressed as a linear combination of correlations. We implemented target functions based on shape complementarity, desolvation, and electrostatics. We noticed the work by Gardiner et al. only after the completion of this paper. They tested 34 unbound complexes and performed a full 6-D search for some of them. Nine of these complexes are also included in our dataset, and we will compare the results of both studies in Discussion and Conclusions.

We could unambiguously identify the correct complex structures for three systems (one protease/inhibitor and two dimers), even though our algorithm is designed for the initial stage of unbound docking. For an additional 13 complexes, we either ranked a near-native structure in the top 20 or obtained 20 or more near-native structures in the top 2,000 or both. We anticipate that a post-processing algorithm could identify the correct binding mode for these complexes relatively easily. Using combined target functions, we were able to obtain near-native structures in the top 2,000 for 24 complexes. We found that shape complementarity and desolvation performed well for different types of complexes. Electrostatics provided significant improvement for five complexes without worsening the performance for others. Our results were obtained without any postprocessing such as filtering, clustering, or rescoring using more sophisticated target functions. The program is completely automatic and does not require any manual intervention. The combination of our algorithm and a postprocessing algorithm could provide a general solution to the unbound protein-protein docking problem.

**METHODS**

Shape Complementarity (SC)

The basic search procedure used in this study is analogous to previous work on FFT docking. The algorithm searches exhaustively the entire rotational and translational space of the ligand with respect to the receptor, which remains fixed at the origin. The rotational search is performed by explicitly rotating the ligand around each of its three Cartesian angles by a certain increment, 15° in this study. For every rotation, the algorithm rapidly scans the translational space using FFT.

Two discrete functions \( R_{SC} \) and \( L_{SC} \); SC stands for shape complementarity) are used to describe the geometric characteristics of the receptor \( R \) (the larger protein) and the ligand \( L \) (the smaller protein), respectively. They are obtained by discretizing \( R \) and \( L \) using a \( N \times N \times N \) grid with each grid point \((i,m,n = 1,2,...N)\) assigned the following value:
DOCKING UNBOUND PROTEINS

\[ R_{SC}(l, m, n) = \begin{cases} 
1 & \text{surface of } R \\
\rho & \text{core} \\
0 & \text{open space}
\end{cases} \]

\[ L_{SC}(l, m, n) = \begin{cases} 
1 & \text{surface of } L \\
\rho & \text{core} \\
0 & \text{open space}
\end{cases} \]

where \( i = \sqrt{-1} \) and \( \rho \) is a positive number. In this paper \( \rho = 9 \). We also tested \( \rho = 8 \) and 10 and these values did not change the results significantly. Throughout our study, we used a grid spacing of 1.2 Å. \( N \) must be sufficiently large so that the grid can cover the sum of the maximal spans of \( R \) and \( L \) plus 1.2 Å. We used \( N = 128 \) for the largest complex 11AI. For others we used \( N = 100 \) since it significantly saved time although Fast Fourier Transform typically performs best when \( N \) is a power of 2.

We identify the surface and core of a protein by computing the solvent accessible area with the radius of the water probe set to 1.40 Å. If a protein atom has more than 1 Å² solvent accessible area, it is considered to be a surface atom. Otherwise, it is a core atom.

\( R_{SC} \) is computed as follows. All grid points are initialized with 0. We first assign \( \rho / 10 \) to grid points that are within \( \sqrt{1.5} \) van der Waals (vdW) radii of a core atom or within \( \sqrt{0.8} \) vdw radii of a surface atom. \( \sqrt{1.5} \) is used to avoid assigning 0 to core grid points and \( \sqrt{0.8} \) is used to allow for some softness. Next, we assign 1 to all grid points that have not been assigned \( \rho / 10 \) and are within 3.4 Å plus the radius of any surface atom. This is equivalent to using a 3.4 Å surface layer for the receptor; 3.4 Å corresponds to the diameter of a water molecule.

\( L_{SC} \) is computed as follows, which is similar to \( R_{SC} \) except for the extra surface layer. All grids are initialized with 0. First, we assign \( \rho / 10 \) to grid points that are within \( \sqrt{1.5} \) times of the vdW radii of a core atom. Next, we assign 1 to all grid points that are within the vdW radius of any surface atom. Finally, if a grid point is assigned \( \rho / 10 \) and two of its nearest neighboring grid points have value 0, it is changed to 1.

Shape complementarity can be described using the correlation function between \( R_{SC} \) and \( L_{SC} \):

\[
S_{SC}(o, p, q) = \text{Re} \left[ \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} R_{SC}(l, m, n) \cdot L_{SC}(l + o, m + p, n + q) \right] \\
- \text{Im} \left[ \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} R_{SC}(l, m, n) \cdot L_{SC}(l + o, m + p, n + q) \right]
\]

where \text{Re} \{ \} and \text{Im} \{ \} denote the real and imaginary parts of a complex function, and \( o, p, \) and \( q \) are the number of grid points by which ligand \( L \) is shifted with respect to receptor \( R \) in each dimension. When there is no contact between \( R \) and \( L \), \( S_{SC} \) equals zero. A surface grid point–surface grid point contact contributes 1 to the correlation. A core-core contact contributes \((\rho/10)^2 = -81\) to the correlation, to penalize core-core clashes that are physically forbidden. A surface-core contact results in a mild penalty of \( \text{Im} \{ \} = -9 \).

Discrete Fourier Transform (DFT) and Inverse Fourier Transform (IFT) can be used to calculate the correlation between two functions:

\[
S_{SC} = \text{Re} \left[ \frac{1}{N^3} \text{IFT}(\text{IFT}(R_{SC}) \cdot \text{DFT}(L_{SC})) \right] \\
- \text{Im} \left[ \frac{1}{N^3} \text{IFT}(\text{IFT}(R_{SC}) \cdot \text{DFT}(L_{SC})) \right]
\]

Both DFT and IFT can be rapidly calculated using the FFT algorithm, which requires in the order of \( N^3 \) \( \log_2 N^3 \) steps of multiplication and summation for transforming a 3-D function of \( N^3 \) grid points. As a result, computing \( S_{SC} \) for all possible translational orientations using equation [3] requires in the order of \( N^3 \cdot 3 \log_2(N^3) \) calculations, which is significantly less than \( N^6 \), the order of calculations required by equation [2].

In order to complete a 6-D exhaustive search, the correlation function \( S_{SC} \) needs to be evaluated for all possible rotational orientations of \( L \) given fixed \( R \). Here, we sample the three angles around each Cartesian axis at a fixed interval of 15°; therefore, \( S_{SC} \) must be calculated for \( 360 \times 360 \times 360 / 15^3 \) orientations. Among these, only 6,389 are non-redundant. Since \( N^3 \) translations are considered for each rotation, approximately 7 billion orientations are examined in total. The enhancement of speed by FFT is more pronounced if we take into account the rotational search, since the computational complexity needs to be multiplied by the number of rotations (6,389 for an interval of 15°).

Desolvation Free Energy (DS)

We have implemented the Atomic Contact Energy (ACE) developed by Zhang et al.\textsuperscript{37} ACE is a desolvation free energy measure, defined as the free energy of replacing a protein-atom/water contact, by a protein-atom/protein-atom contact. ACE was derived from 90 high-resolution protein crystal structures, based on the method of Miyazawa and Jernigan, with a number of improvements.\textsuperscript{37} Combined with an electrostatics term, ACE has been shown to accurately predict the free energy changes associated with transferring amino acids from protein interior to water, site-specific mutations of hydrophobic amino acids, and protease-inhibitor binding.\textsuperscript{37,39,40}

The ACE scores were obtained for all pairs of 18 atom types. The total desolvation free energy of complex formation (\( S_{DS} \); DS stands for desolvation) is calculated by summing the ACE scores of all atom pairs between the receptor and the ligand that are within 6 Å.\textsuperscript{37} Expressed in the form of correlations, the computation of \( S_{DS} \) requires 18 DFT and 36 IFT. In order to speed up the calculation, we decided to use 18 non-pairwise ACE scores, representing the score between one protein atom of a specific type and another protein atom of any type.\textsuperscript{37} The desolvation
energies calculated using pairwise and non-pairwise ACE scores are highly correlated. The correlation coefficient was 0.94 for all crystal complexes used in this study.

We use two discrete functions, $R_{DS}$ and $L_{DS}$, to describe the desolvation (DS) properties of the receptor and the ligand. They are similar to $R_{SC}$ and $L_{SC}$ in Equation [1] but with different values. First, both $R_{DS}$ and $L_{DS}$ are initialized to 0. We then loop over every grid point and assign the real portion of $R_{DS}$ to the sum of the ACE scores of all receptor atoms that are within 6 Å. Next, we loop over all receptor atoms and assign 1 to the imaginary portion of $R_{DS}$ for the nearest grid point (regardless of the real portion). $L_{DS}$ is computed in exactly the same way for the ligand. $S_{DS}$ can be expressed in the following correlation:

$$S_{DS}(o, p, q) = \frac{1}{2} \times \text{Im} \left[ \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} R_{DS}(l, m, n) \right]$$

This formulation of $S_{DS}$ requires only one DFT and two IFT to compute:

$$S_{DS} = \frac{1}{2} \times \text{Im} \left[ \frac{1}{N^2} \text{IFT} \left( \text{IFT} \left( R_{DS} \cdot \text{DFT} \left( L_{DS} \right) \right) \right) \right]$$

Electrostatics (ELEC)

The electrostatics energy can be expressed as a correlation between the electric potential generated by the receptor and the charges of the ligand atoms, as described by the Coulombic formula. We adopted the approach by Gabb et al.,

$$\text{Calculation of electrostatics energy}$$

Combining SC, DS, and ELEC

In previous applications, ACE was combined with a Coulombic electrostatics term without any scaling. However, in an unbound docking study, surface side chains tend to be at non-optimial positions and thus the resulting electrostatics energies can be inaccurate and even unrealistic. Artifacts can also result from the grid representation. In addition, here we use the non-pairwise ACE scores, instead of the pairwise ones in previous applications. With the above considerations, we need to scale the electrostatics energy $S_{ELEC}$. We also included the shape complementarity score with another scaling factor. The final target function is:

$$S = \alpha S_{SC} + S_{DS} + \beta S_{ELEC}$$

The default values for scaling factors are $\alpha = 0.01$ and $\beta = 0.06$ in this study. We varied both factors from 50 to 200% of the default values but observed no significant changes in the results.

Label Non-CDR Residues of Antibodies

Since the binding sites of immunoglobulins are invariant and well characterized (the Complementarity Determining Regions or CDRs), we can limit docking to the CDR region. This was easily achieved by assigning the surface layer only to the CDR atoms of antibodies ($R_{SC}$ in Equation [1]). The CDR1, CDR2, and CDR3 residues were based on the Kabat definition with three extra residues in both directions (http://www.biochem.ucl.ac.uk/~martin/abs/GeneralInfo.html). We also included CDR4, extended for two residues in both directions. The detailed list of CDR residues can be found at http://zlab.bu.edu/~rong/dock/cdrlist.html.

Calculate Root Mean Square Deviation (RMSD)

The quality of a docked structure was assessed by the RMSD of its interface Cα atoms after superposition onto the cocrystallized protein complex. The interface was defined as all residues with at least one atom located within 10 Å of the other protein. We computed the RMSD of superposed complex, shown in the last column of Table I. Throughout the paper, all predictions with interface RMSD less than or equal to 2.5 Å are defined as near-native structures or hits. Our RMSD definition and the 2.5 Å cutoff defining near-native structures are the same as in Gabb et al.

Desolvation and Electrostatics for Co-crystallized Protein Complexes

To investigate the performance of various energy terms, we calculated the detailed desolvation and electrostatics energies for all co-crystallized protein complexes. The desolvation energy was calculated using the pairwise ACE scores in Zhang et al. The electrostatics energy was calculated using CHARMM, after 100 steps of energy minimization with all non-hydrogen atoms constrained to the starting positions in the crystal structure. The results are shown in the last two columns of Table II.

Collapsed Side Chain

Surface side chains at the binding site frequently change conformations upon complex formation. As a result, they can overlap core atoms in near-native structures and cause large penalties in $S_{SC}$. After comparing the bound and free structures, we found Arg and Lys side chains to have the greatest conformational changes. Therefore, we analyzed the effect of “collapsed side chains,” where the side chain atoms of all surface Arg and Lys residues in the ligand were “collapsed” to the positions of the corresponding Cα atoms. A residue is on the surface only if all of its side chain atoms (Cα and beyond) are defined as surface atoms.

Computational Implementation

We have implemented our algorithm in a C program that consists of approximately 1,400 lines of code. The program is called ZDOCK. All test cases were run on an
RESULTS

Table I lists the 27 distinct protein-protein complexes used in this study. For 13 of them, the unbound structures of both proteins were used in docking. We call these unbound/unbound complexes. For the other 14, the unbound structure of only one molecule was available, and the bound structure of the other molecule was used. We call these unbound/bound complexes. Monomers of a dimer do not exist freely in solution. For the five dimers, the structure of one monomer was used to dock to itself. Therefore, dimers fall into the unbound/bound category. For the nine antibody/antigen complexes, we restricted the search to the antibody’s CDR regions. For antigens and both molecules in the remaining complexes, we assumed no binding site information and performed a full search.

Our program ZDOCK inputs two PDB files and outputs a list of ligand orientations with the best scores. Since we focus on the initial stage of unbound docking, we do not perform any post-processing of the resulting structures. Our goal is to retain as many near-native structures in the top 2,000 as possible, for as many complexes as possible. A near-native structure (or hit) is defined as a docked structure with interface C atoms RMSD ≤ 2.5 Å.

For every rotational orientation, ZDOCK keeps the 10 best translational orientations. Three lists are kept at all times, each containing 1,000 ligand orientations: one list with the best shape complementarity score SSC, another list with the best shape complementarity and desolvation score $SSC + SDS$, and a third list with the best shape complementarity, desolvation, and electrostatics score $SSC + SDS + SELEC$. The RMSDs of the kept structures can be computed and the numbers of hits obtained.

Shape Complementarity Alone ($SSC$)

The second column in Table II shows that among the top 1,000 structures ranked according to $SSC$, we were able to identify 1 to 90 hits for 15 complexes. For seven of these, we had 10 or more hits. These include three protease/inhibitor and four dimers. Among the 12 complexes for which we did not identify any hits, eight were unbound/unbound and only four were unbound/bound complexes.
The above results indicate that shape complementarity alone can identify near-native structures for unbound docking. However, the success rate is only 15/27, and the number of hits is small, fewer than 10 for half of the successful cases. In addition, it is highly sensitive to structural variations, since most of the unsuccessful cases involve unbound/unbound complexes.

**Shape Complementarity and Desolvation \((\alpha S_{SC} + S_{DS})\)**

The third column in Table II shows the number of near-native structures among the top 1,000 structures ranked by \(\alpha S_{SC} + S_{DS}\). We were able to identify between 2 and 223 hits for 18 complexes. For 14 complexes, we generated 10 or more hits. For some complexes, we generated many hits, namely 112 for 2SIC, 223 for 1DXG, 183 for 2MIP, and 99 for 1CTA. The unsuccessful cases include four unbound/unbound and five unbound/bound complexes.

Although the success rates between the desolvation target function \(\alpha S_{SC} + S_{DS}\) and the shape complementarity target function \(S_{SC}\) do not differ drastically (18/27 vs. 15/27), desolvation identified many more hits than shape complementarity. Specifically, desolvation generated 10 or more hits for 14 out of the 18 successful cases. The performance was particularly good for dimers and protease/inhibitor. For antibody/antigen, shape complementarity alone did better. Desolvation succeeded for eight complexes on which shape complementarity failed, but failed on five complexes for which shape complementarity alone succeeded. This indicates that the binding mechanism differs among protein-protein complexes.

The performance of the desolvation target function did not seem to differ between unbound/unbound or unbound/bound complexes. In fact, more than half of the unsuccessful cases involved unbound/bound complexes. The only unbound/bound protease/inhibitor (1UGH) performed much worse than most of the unbound/unbound protease/inhibitor complexes. Intrigued by this, we computed the desolvation energies for co-crystallized complexes and for superposed complexes using unbound structures; the correlation coefficient was 0.94 for all complexes. This indicates

### Table II. Docking Results

<table>
<thead>
<tr>
<th>PDB ID of complex</th>
<th>Number of hits in top 1,000</th>
<th>(S_{SC} &amp; \alpha S_{SC} + S_{DS} + \beta S_{ELEC}) (number of hits in top 2,000)</th>
<th>Energies for crystal complexes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(S_{SC})</td>
<td>(\alpha S_{SC} + S_{DS})</td>
<td>(\beta S_{ELEC})</td>
</tr>
<tr>
<td>SSB</td>
<td>1</td>
<td>55</td>
<td>50</td>
</tr>
<tr>
<td>SSB</td>
<td>1</td>
<td>22</td>
<td>27</td>
</tr>
<tr>
<td>SSB</td>
<td>0</td>
<td>20</td>
<td>16</td>
</tr>
<tr>
<td>SSB</td>
<td>0</td>
<td>21</td>
<td>45</td>
</tr>
<tr>
<td>SSB</td>
<td>0</td>
<td>43</td>
<td>49</td>
</tr>
<tr>
<td>SSB</td>
<td>0</td>
<td>112</td>
<td>84</td>
</tr>
<tr>
<td>SSB</td>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SSB</td>
<td>11</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>SSB</td>
<td>0.16</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>SSB</td>
<td>0.16</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>SSB</td>
<td>0.16</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>SSB</td>
<td>0.16</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>SSB</td>
<td>0.16</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>

\(^1\)Hits (or near-native structures) are defined as docked structures with interface C<sub>r</sub> RMSD ≤2.5 Å from the crystal complex. See Methods for details on RMSD calculations.

\(^2\)Ranked according to \(\alpha S_{SC} + S_{DS} + \beta S_{ELEC}\). — indicates that no hit was found, likewise for other columns and tables.

\(^3\)Unbound/bound complexes.
that our desolvation function is highly tolerant to conformational changes upon binding.

To further investigate why desolvation failed on some complexes, we computed the desolvation and electrostatics energies for all co-crystallized structures (the last two columns of Table II). Figure 1 plots the number of hits in the top 1,000 structures identified using target functions $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ (solid circles) and $\alpha S_{SC} + S_{DS}$ (open squares). Desolvation energies (in kcal/mol) were calculated for co-crystallized complexes using pairwise ACE scores. They are listed in the second-last column of Table II (labeled $S_{des}$). The numbers of hits are listed in the third and forth columns of Table II (labeled $\alpha S_{SC} + S_{DS}, \alpha S_{SC} + S_{DS} + \beta S_{ELEC}$, respectively).

**Shape Complementarity, Desolvation, and Electrostatics ($\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$)**

The $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ column in Table II shows the improvement that resulted from adding the simple Coulombic electrostatics energy to the target function. Compared to $\alpha S_{SC} + S_{DS}, \alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ identified more hits for 12 complexes, fewer hits for seven complexes, and the same number of hits for the remaining eight complexes. For five complexes, electrostatics significantly increased the number of hits: 1MAH, 2PTC, 1FSS, 1UGH, and 1BQL, labeled in Figure 1. In particular, both $S_{SC}$ and $\alpha S_{SC} + S_{DS}$ failed to find any hits for 1FSS while $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ successfully identified 3. Interestingly, none of above five complexes have highly favorable desolvation energies (filled circles in Fig. 1 indicate the number of hits by $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$). They do however all have highly favorable electrostatics energies (Table II).

In Figure 2, the difference in the numbers of hits by $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ and $\alpha S_{SC} + S_{DS}$ is plotted against the electrostatics energy of the crystal complex calculated using CHARMM. For the 13 electrostatics driven complexes (electrostatics energies below $-17$ kcal/mol), either $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ found more hits than $\alpha S_{SC} + S_{DS}$, or both failed to find any hits. Notably, among the four complexes on which electrostatics also failed, the two most charged ones were both unbound/unbound complexes (1BRS and 2PCC). This indicates that the simple Coulombic electrostatics is highly sensitive to conformational changes upon binding. The only complex for which $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ found many fewer hits than $\alpha S_{SC} + S_{DS}$ was 2SIC (84 vs. 112; labeled in Figs. 1 and 2). Nonetheless, both target functions could identify the correct binding mode for this complex unambiguously (Fig. 3(a)).

**Combining $S_{SC}$ and $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$**

Although $S_{SC}$ generally performed much worse than $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$, it did identify hits for complexes on which $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ failed: 1BRS, 2JEL, 1JHL, 1FBI, and 3SDH. Since the focus of this study is to rank near-native structures in the top few thousands for as
many complexes as possible, combining the two lists generated by both target functions provides a much broader coverage. Table II shows the number of hits in the top 2,000, as well as the ranking and RMSD of the highest-ranking hit. The ranking was computed according to $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ and $\alpha S_{SC} + S_{DS}$. Note that we deleted redundant orientations when we combined the two lists. We were able to obtain near-native structures in the top 2,000 for 24 of 27 complexes. This represents a larger set of successfully docked complexes compared with previous unbound docking studies.21,22,25–28

For three complexes (2SIC, 1DXG, and 2MIP), a near-native structure was ranked number 1. Figure 3 plots $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ against the RMSD of the top 1,000 structures for each complex. Dozens of near-native structures are ranked above any false-positive structures. Therefore, the complex structures for these three systems can be identified unambiguously. Moreover, Figure 3(a) shows a large energy gap between the highest-ranking false-positive structure and many near-native structures, indicating the robustness of our algorithm. Among these three systems, 2SIC is a protein/inhibitor; 1DXG and 2MIP are both dimers. They are among the complexes with the most favorable desolvation energies (labeled in Fig. 1). This indicates that our desolvation energy calculation is a key component to the successful prediction of these complexes.

In addition to the above three systems, a near-native structure was ranked in the top 20 for nine complexes. Moreover, for complexes 1CHO, TEM-BLIP, 3SDH, and 1MEL, although the best near-native structure did not rank in the top 20, there were 20 or more near-native structures in the top 2,000. The above 13 complexes represent relatively easy cases for post-processing. Figure 4 plots $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ against the RMSD of the top 1,000 structures for each of these complexes. These plots are clearly far from random, and sometimes there are only a few false-positive structures scattered around. 3SDH is absent from Figure 4. Although there were 23 near-native structures in the top 2,000, none of them ranked higher than 1,000.

For eight other complexes, although we successfully identified near-native structures in the top 2,000, we did not retain many such structures and failed to rank them competitively. For four of these complexes (1BRS, 2JEL, 1JHL, and 1FBI), shape complementarity alone found near-native structures but $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ failed. For the other four (1FSS, 1IAI, 6EBX, and 1IGC), $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ succeeded but shape complementarity alone failed. These complexes represent difficult cases for post-processing. Interestingly, six of these are unbound/bound and the other two are unbound/unbound complexes. The RMSDs of the superposed structures (the last column in Table I) do not distinguish them from the more successfully predicted complexes. Meanwhile, the desolvation energies of these complexes are either unfavorable or only marginally favorable (with the lowest $S_{DS}$ value being $-17$ kcal/mol). For some complexes, this is compensated by favorable electrostatics energies (e.g., for 1BRS, 1FBI, and 1FSS), and for others even the electrostatics energies are not particularly favorable.
and 1VFB, shape complementarity as formulated in this paper failed. We are testing a few ideas and have been able to detect near-native structures for both 3HFM and 1VFB in the top 1,000 using a novel shape complementarity function (Chen and Weng, unpublished data).

**Comparing Normal and Collapsed Side Chains**

Conformational change of surface residues is a major contributor to the difficulty in unbound docking. One simple idea is to selectively “soften” the residues that most frequently change their conformations upon binding (i.e., Arg and Lys). The scheme we chose to test was to “collapse” the side chain atoms of all surface Arg and Lys residues in the ligand to the positions of the corresponding Cα atoms. With all parameters kept the same, the results of the “collapsed” side chain scheme are shown in Table III. Overall, the effect is small. The collapsed side chain scheme improved the performance of SSC alone, as indicated by the number of near-native structures. However, the performance of $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ was worse for the collapsed side chain scheme.

**DISCUSSION AND CONCLUSIONS**

We have presented a protein-protein docking algorithm ZDOCK focused on the initial stage of unbound docking. Our goal is to retain as many near-native structures as possible in the top 2,000 choices, for as many complexes as possible, with minimal assumption of the binding site. The search algorithm of ZDOCK is based on the FFT method proposed by Katchalski-Katzir et al. and further developed by Sternberg and colleagues. The key component of our target function is a desolvation term based on the Atomic Contact Energy (ACE) developed by Zhang et al. Two other components of our target function are shape complementarity and electrostatics. We tested ZDOCK on a large dataset (Table I). For antibodies in antibody/antigen, we restricted the search to the CDR region. For all other molecules, we did not assume any binding site information and performed a full 6-D search.

The first major finding of our study is that we can unambiguously identify the correct complex structures for three systems (2SIC, 1DXG, and 2MIP). In addition to a near-native structure being ranked first, there is a large number of near-native structures ranked above the first false positive [Fig. 3(a)]. 2SIC was also used in three previous unbound docking studies, none of which could identify the correct complex structure. The good performance on 1DXG and 2MIP should be somewhat discounted, since the RMSD between two monomers in a dimer tends to be smaller than that between the bound and unbound forms of a soluble protein. Palma et al. also obtained good results for these two complexes, with the best ranking of a hit being 2 for 1DXG and 82 for 2MIP.

The second major result of our study is that we can rank one or more near-native structures in the top 2,000 choices for the majority of complexes. The intricate balance of different energies in protein-protein interactions makes it important to test docking algorithms on many complexes. We successfully identified near-native structures for...
twenty-four complexes. The three complexes we failed on are 3HFM, 1VFB, and 2PCC. The electrostatics interaction energy for 2PCC is highly favorable (the last column in Table II); unfortunately, our simple Coulombic energy term did not improve the result. The more sophisticated continuum electrostatics enabled Mandell et al. to identify 1 hit in the top 500 choices for 2PCC.²²

We became aware of the recent work by Gardiner et al.²⁹ only after the completion of this work. They docked 34 unbound complexes using a Genetic Algorithm. Their target function depends primarily on shape complementarity, enhanced with hydrogen bonding information. For antibodies, they restricted searches to the binding site, defined as all residues within 8 Å of the antigen. This

Fig. 4. $\alpha S_{DC} + \beta S_{ELEC}$ (ordinate) vs. RMSD (abscissa) for the top 1,000 structures of the 13 complexes that either have one or more near-native structure in top 20 or have 20 or more near-native structures in top 2,000.
de

inition requires knowledge of the complex structure, whereas our CDR de

inition does not. Our de

inition includes many residues that are more than 8 Å from the antigen, and is therefore less restrictive than Gardiner et al.’s. This makes it impossible to compare the results of the two studies on antibody/antigen complexes. There are, however, nine other complexes for which both studies performed full 6-D searches; a comparison of these results is presented in Table IV. In Table IV, hits are de

ined according to Gardiner et al.’s criteria and performance is judged according to the number of hits and the ranking of the best hit in the top 100. For five complexes (1CGI, 2SIC, 2SN, 1UDI, and TEM-BLIP), ZDOCK performed better, i.e., we obtained more hits, as well as a higher ranking for the best hit. For two complexes (1CHO and 1BRS), Gardiner et al.’s method performed better than ours. For the remaining two complexes (2KAI and 2PTC), results were mixed; we obtained more hits while Gardiner achieved better rankings.

Three components are in our target function: shape complementarity (SSC), desolvation (SDS), and electrostatics (SELEC). The key is the desolvation term SDS. Since SDS alone does not prevent steric overlap between the receptor and the ligand, in this study we combined SDS with SSC. Table II shows that SSC + SDS can identify near-native structures for 18 complexes, which constitute the majority of the 24 successfully predicted complexes in this study. For the other six, one comes from adding electrostatics and the other five from shape complementarity alone. More importantly, for the three complexes that were solved completely by $\alpha S_{SC} + S_{DS}$, $\beta S_{\text{ELEC}}$ (2SIC, 1DXG and 2MIP), $\alpha S_{SC} + S_{DS}$ performed equally well or better. Moreover, $\alpha S_{SC} + S_{DS}$ succeeded for all complexes that were desolvation driven, namely with desolvation energy of the crystal complex being $-5$ kcal/mol and below (Table II, Fig. 1). This consistency indicates that our $S_{DS}$ calculation is robust. Furthermore, it is not very sensitive to conformational changes, indicated by the high correlation between $S_{DS}$ for the co-crystallized complexes and that for superposed complexes using unbound structures.

The simple Coulombic electrostatics significantly improved our results, in particular for the electrostatics-driven complexes (Fig. 2). Unlike the desolvation term, $S_{\text{ELEC}}$ is highly sensitive to conformational changes. This could explain the lack of improvement in two of the most electrostatics-driven complexes. Two previous unbound-docking studies considered the Coulombic electrostatics as part of the target function. Gabb et al. found it too noisy to be added to the scoring function and used it as a binary filter instead.21 Ritchie and Kemp added a Coulombic electrostatics term to the shape complementarity score using an adjustable parameter.28 Their Table V indicates that electrostatics did not result in significant improvement. The reason that we can take more advantage of the simple Coulombic electrostatics than the above two studies could be that our desolvation term is already quite accurate and furthermore is tolerant to conformational changes. This is supported by the following finding. We tested the target function of combining electrostatics directly with shape complementarity (without desolvation), and obtained generally poor performance compared to shape complementarity alone: the performance of five complexes improved, ten worsened, and twelve tied.

Among the complexes studied here, ZDOCK performed the best for protease/inhibitor and dimers. The overall performance for antibody/antigen complexes was somewhat worse. Previous studies analyzed different classes of protein-protein complexes. Two studies showed that antibody/antigen interfaces tend to have poorer shape comple-
mentarity than other complexes. However, a recent study using Voronoi volumes concluded that these two types of complexes have comparable degrees of shape complementarity. Our results support the latter study, since the performance our shape complementarity target function did not differ much between these two types of complexes. Jackson discovered that antibody/antigen complexes interact predominantly through charge-charge and charge/polar-tyrosine interactions. In agreement, we have noticed that most antibody/antigen complexes have unfavorable desolvation energies. As a result, our desolvation target function performed the worst for this type of complexes.

### TABLE III. Comparison Between the Normal and Collapsed Side Chain Schemes

<table>
<thead>
<tr>
<th>Complex</th>
<th>Normal</th>
<th>Collapsed</th>
<th>Normal</th>
<th>Collapsed</th>
<th>Normal</th>
<th>Collapsed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSC</td>
<td></td>
<td></td>
<td>αSSC + SDES + βSELLEC</td>
<td></td>
<td></td>
<td>αSSC + SDES + βSELLEC</td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

† Hits (or near-native structures) are defined as docked structures with interface Cα RMSD ≤2.5 Å from the crystal complex. See Methods for details on RMSD calculations.

### TABLE IV. Comparison to the Results by Gardiner et al. (2001)

<table>
<thead>
<tr>
<th>Complex</th>
<th>RMSD cutoff for hits (Å)</th>
<th>Number of hits in top 100</th>
<th>Best rank</th>
<th>RMSD of the best hit</th>
<th>Number of hits in top 100</th>
<th>Best rank</th>
<th>RMSD of the best hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1CGI</td>
<td>5.48</td>
<td>45</td>
<td>6</td>
<td>2.61</td>
<td>8</td>
<td>34</td>
<td>5.04</td>
</tr>
<tr>
<td>2KAI</td>
<td>5.00</td>
<td>15</td>
<td>7</td>
<td>4.97</td>
<td>8</td>
<td>4</td>
<td>4.39</td>
</tr>
<tr>
<td>2PTC</td>
<td>4.74</td>
<td>10</td>
<td>6</td>
<td>1.62</td>
<td>5</td>
<td>25</td>
<td>4.38</td>
</tr>
<tr>
<td>2SIC</td>
<td>4.38</td>
<td>38</td>
<td>1</td>
<td>1.93</td>
<td>6</td>
<td>1</td>
<td>3.91</td>
</tr>
<tr>
<td>2SNI</td>
<td>4.85</td>
<td>26</td>
<td>2</td>
<td>1.20</td>
<td>2</td>
<td>51</td>
<td>2.05</td>
</tr>
<tr>
<td>1CHO</td>
<td>4.33</td>
<td>2</td>
<td>72</td>
<td>1.53</td>
<td>9</td>
<td>1</td>
<td>1.79</td>
</tr>
<tr>
<td>1UDI</td>
<td>4.37</td>
<td>7</td>
<td>6</td>
<td>1.20</td>
<td>3</td>
<td>13</td>
<td>2.23</td>
</tr>
<tr>
<td>1BRS</td>
<td>4.47</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>2</td>
<td>1</td>
<td>3.14</td>
</tr>
<tr>
<td>TEM-BLIP</td>
<td>4.88</td>
<td>2</td>
<td>57</td>
<td>0.97</td>
<td>0</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

† Hits are docked structures with interface RMSD smaller than the corresponding cutoff as defined in Gardiner et al. (2001). Results by Gardiner et al. are adopted from Table IV of Gardiner et al. with normal angle ±30° and J = 3.0, the parameter set that gave them the best overall performance.

a Ranked according to αSSC + SDES + βSELLEC.
complexes. Since desolvation is the major component of ZDOCK, the overall performance for antibody/antigen is worse than other types of complexes.

We did not optimize the parameters in our shape complementarity calculations. For simplicity, we used the same grid spacing for all complexes. Due to the large sizes of some complexes, we could only use the grid resolution of 1.2 Å coarser than previous FFT-based unbound docking algorithms: Gabb et al. used grid sizes between 0.74 and 0.94 Å\(^2\) and Mandell et al. used 1 Å.\(^2\)\(^22\) We used a surface layer of 3.4 Å, similar to Eisentein and colleagues.\(^18\) However, Gabb et al. reported improved performance with smaller surface thickness (1.2 and 1.5 Å).\(^21\) We are currently testing different grid spacings and surface thicknesses, as well as other shape complementarity formulations. Also, our angle sampling is non-uniform and the spacing of 15° could be too coarse, and we are currently investigating these issues.

In summary, our unbound docking algorithm can sample all possible ligand orientations and successfully predict near-native structures in the top 2,000 for most of the complexes. For three complexes, ZDOCK can identify the correct complex structure unambiguously. For the other complexes, ZDOCK can be combined with a post-processing procedure to provide a general solution to the unbound protein-protein docking problem.

ACKNOWLEDGMENTS

We thank Joseph Szustakowski for critically reading the manuscript. We are grateful to the Boston University Scientific Computing Facilities for computing supports. We thank Dr. Natalie Strynadka for providing the coordinates of the TEM-1 β-lactamase/BLIP complex and the BLIP crystal structures.

REFERENCES


