Docking Unbound Proteins Using Shape Complementarity, Desolvation, and Electrostatics

Rong Chen[1] **and Zhiping Weng**[2]*
[1]*Bioinformatics Program, Boston University, Boston, Massachusetts*
[2]*Department 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 C_α 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 genomics[5] has considerably accelerated the experimental determination of protein structures. As of June 19, 2001, there are 15,435 entries in the Protein Data Bank (PDB[6]). 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 colleagues[17,18] 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 methods[19–22] performed well in two blind trials.[18,23,24] 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,[27] and a method based on spherical polar Fourier correlations.[28] A recent paper uses a Genetic Algorithm to match protein surface dots.[29]

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.[30–24]

A number of groups have also investigated the initial stage of unbound docking. Jiang and Kim[25] 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.[21] Norel et al. applied a computer vision-based algorithm to unbound docking.[26] 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.[27] 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.[28] 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.[22]

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.[29] 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.[19,21] 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 ($l,m,n = 1,2,\ldots N$) assigned the following value:

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

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

where $i = \sqrt{-1}$ and ρ is a positive number. In this paper ρ = 9. We also tested ρ = 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 1IAI. 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[35] with the radius of the water probe set to 1.40 Å. If a protein atom has more than 1 Å^2 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 ρi 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 ρi 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 ρi 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 ρi 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}:

$$S_{SC}(o, p, q) = \mathrm{Re}\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R_{SC}(l, m, n) \cdot L_{SC}(l + o, m + p, n + q)\right] - \mathrm{Im}\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R_{SC}(l, m, n) \cdot L_{SC}(l + o, m + p, n + q)\right] \quad (2)$$

where Re[ ] and 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 $(\rho i)^2 = -81$ to the correlation, to penalize core-core clashes that are physically forbidden. A surface-core contact results in a mild penalty of $\mathrm{Im}[\rho i] = -9$.

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

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

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 $N^3 \cdot \log_2(N^3)$ 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 $360{*}360{*}360/15^3$ orientations. Among these, only 6,389 are non-redundant.[36] 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 of15°).

Desolvation Free Energy (DS)

We have implemented the Atomic Contact Energy (ACE) developed by Zhang et al.[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.[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.[37-40]

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.[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.[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:

$$S_{DS}(o,p,q) = \frac{1}{2} \times \mathrm{Im}\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R_{DS}(l,m,n) \cdot L_{DS}(l+o, m+p, n+q)\right] \quad (4)$$

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

$$S_{DS} = \frac{1}{2} \times \mathrm{Im}\left[\frac{1}{N^3}\,\mathrm{IFT}(\mathrm{IFT}(R_{DS}) \cdot \mathrm{DFT}(L_{DS}))\right] \quad (5)$$

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.,[21] except that we used the partial charges in the CHARMM19 potential. This was motivated by the previous success of the ACE desolvation when combined with an electrostatics energy term based on the CHARMM19 partial charges.[37–40] In addition, grid points in the core of the receptor are assigned a value of 0 for the electric potential, to avoid the contributions from non-physical receptor-core/ligand contacts.

Combining SC, DS, and ELEC

In previous applications, ACE was combined with a Coulombic electrostatics term without any scaling.[37–40] However, in an unbound docking study, surface side chains tend to be at non-optimal 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.[37–40] 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:

$$S = \alpha S_{SC} + S_{DS} + \beta S_{\mathrm{ELEC}} \quad (6)$$

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 C^{α} 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.[21]

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.[37] The electrostatics energy was calculated using CHARMM,[41] 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 C_{β} atoms. A residue is on the surface only if *all* of its side chain atoms (C_{β} 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

TABLE I. Test Systems

Complex name	Receptor[a]	Ligand[a]	Complex[a]	RMSD[b]
α-chymotrypsinogen/human pancreatic secretory trypsin inhibitor	1CHG	1HPT	1CGI	1.53
α-chymotripsin/ovomucoid 3rd domain	5CHA	2OVO	1CHO	0.62
Kallikrein A/bovine pancreatic trypsin inhibitor	2PKA	6PTI	2KAI	0.91
β-trypsinogen/bovine pancreatic trypsin inhibitor	2PTN	6PTI	2PTC	0.31
Subtilisin Novo/chymotrypsin inhibitor 2	1SUP	2CI2	2SNI	0.37
Subtilisin BPN/subtilisin inhibitor	1SUP	3SSI	2SIC	0.39
Barnase/Barstar	1A2P	1A19	1BRS	0.47
TEM-1 β-lactamase/BLIP	1BTL	N/A	N/A	0.67
Mouse Acetylcholinesterase/Fasciculin II	1MAA	1FSC	1MAH	0.60
Torpedo Acetylcholinesterase/Fasciculin II	2ACE	1FSC	1FSS	0.76
Uracil-DNA Glycosylase/inhibitor	1AKZ	1UGH	1UGH	0.28
1gG1 D44.1 Fab fragment/lysozyme	1MLB	1LZA	1MLC	0.85
1gG1 HyHel-10 Fab fragment/lysozyme	3HFM	1LZA	3HFM	0.44
Jel42 Fab fragment/histidine phosphocarrier protein (HPr)	2JEL	1POH	2JEL	0.19
IgG1 HyHel-5 Fab fragment/lysozyme	1BQL	1DKJ	1BQL	0.52
IgG1 Fv fragment/lysozyme	1JHL	1GHL	1JHL	0.28
IgG1 Fab fragment/lysozyme	1FBI	1HHL	1FBI	0.49
IgG1 Idiotypic Fab/Igg2A anti-idiotypic Fab	1AIF	1IAI	1IAI	1.06
IgG1 Fv fragment/lysozyme	1VFA	1LZA	1VFB	0.59
V_H single-domain antibody/lysozyme	1MEL	1LZA	1MEL	0.65
Hemoglobin dimer	3SDH	3SDH	3SDH	0.20
Desulforeredoxin dimer	1DXG	1DXG	1DXG	0.19
Erabutoxin b dimer	6EBX	6EBX	6EBX	0.23
HIV-2 protease with peptide inhibitor (dimer)	2MIP	2MIP	2MIP	0.30
Troponin dimer	1CTA	1CTA	1CTA	0.91
Yeast cytochrome C peroxidase/yeast iso-1-cytochrome C	1CCP	1YCC	2PCC	0.39
IgG1 Fab fragment/protein G	1IGC	1IGD	1IGC	0.38

[a] 4-letter PDB code for the crystal structures used in this study. The structures were downloaded from PDB[6] except for BLIP and TEM-1 β-lactamase/BLIP, which were kindly provided by Dr. Natalie Strynadka.
[b] The RMSD (in Å) of the interface C_α atoms for input receptor and ligand after superposition onto the co-crystallized complex structure. See Methods for details on RMSD calculations.

SGI origin 2000 with 32 R10000 CPUs. Complexes with grid size N equal to 100 require less than 10 hours of CPU time on a single processor. The run time of complex 1IAI, which requires $N = 128$, is 19 hours. The program is freely available to academic users (http://zlab.bu.edu/~rong/dock/).

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_α 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 S_{SC}, another list with the best shape complementarity and desolvation score $\alpha S_{SC} + S_{DS}$, and a third list with the best shape complementarity, desolvation, and electrostatics score $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$. The RMSDs of the kept structures can be computed and the numbers of hits obtained.

Shape Complementarity Alone (S_{SC})

The second column in Table II shows that among the top 1,000 structures ranked according to S_{SC}, we were able to retain 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.

TABLE II. Docking Results[†]

PDB ID of complex	Number of hits in top 1,000			S_{SC} & $\alpha S_{SC} + S_{DS} + \beta S_{\mathrm{ELEC}}$ (number of hits in top 2,000)			Energies for crystal complexes[c]	
	S_{SC}	$\alpha S_{SC} + S_{DS}$	$\alpha S_{SC} + S_{DS} + \beta S_{\mathrm{ELEC}}$	Number of hits	Rank[a]	RMSD (Å)[b]	S_{DS}	S_{ELEC}
1CGI	1	55	50	51	11	1.94	−23	−17
1CHO	1	22	27	28	72	1.53	−14	−18
2KAI	0	20	16	16	11	1.91	−9	−16
2PTC	0	21	45	45	6	1.62	−5	−24
2SNI	0	43	49	49	2	1.20	−16	−15
2SIC	0	112	84	84	1	1.93	−18	−17
1BRS	13	0	0	13	1,024	1.62	10	−40
TEM-BLIP	11	9	10	20	57	0.97	10	−41
1MAH	0	16	33	33	5	1.86	−7	−32
1FSS	0	0	3	3	210	1.34	−2	−33
1UGH[d]	10	3	20	28	6	1.20	2	−37
1MLC	5	2	3	8	18	2.00	3	−22
3HFM[d]	0	0	0	0	—	—	14	−6
2JEL[d]	4	0	0	4	1,085	2.42	8	−5
1BQL[d]	1	9	37	38	2	1.39	7	−18
1JHL[d]	1	0	0	1	1,640	0.62	10	−14
1FBI[d]	1	0	0	1	1,573	1.63	11	−29
1IAI[d]	0	16	13	13	94	2.08	−3	−3
1VFB	0	0	0	0	—	—	10	−9
1MEL[d]	8	16	13	21	38	1.33	−6	−13
3SDH[d]	23	0	0	23	1,008	0.81	9	−20
1DXG[d]	90	223	236	324	1	0.65	−27	−18
6EBX[d]	0	10	9	9	430	2.40	−3	−4
2MIP[d]	17	183	185	199	1	0.67	−45	−12
1CTA[d]	16	99	106	122	5	0.90	−25	11
2PCC	0	0	0	0	—	—	7	−42
1IGC[d]	0	10	3	3	727	2.45	−4	−3

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

[a] Ranked according to $\alpha S_{SC} + S_{DS} + \beta S_{\mathrm{ELEC}}$. – indicates that no hit was found, likewise for other columns and tables.

[b] RMSD for the best ranked hit.

[c] The desolvation and electrostatics energies (in kcal/mol) for co-crystallized protein complexes (see Methods for the details of energy calculations).

[d] 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

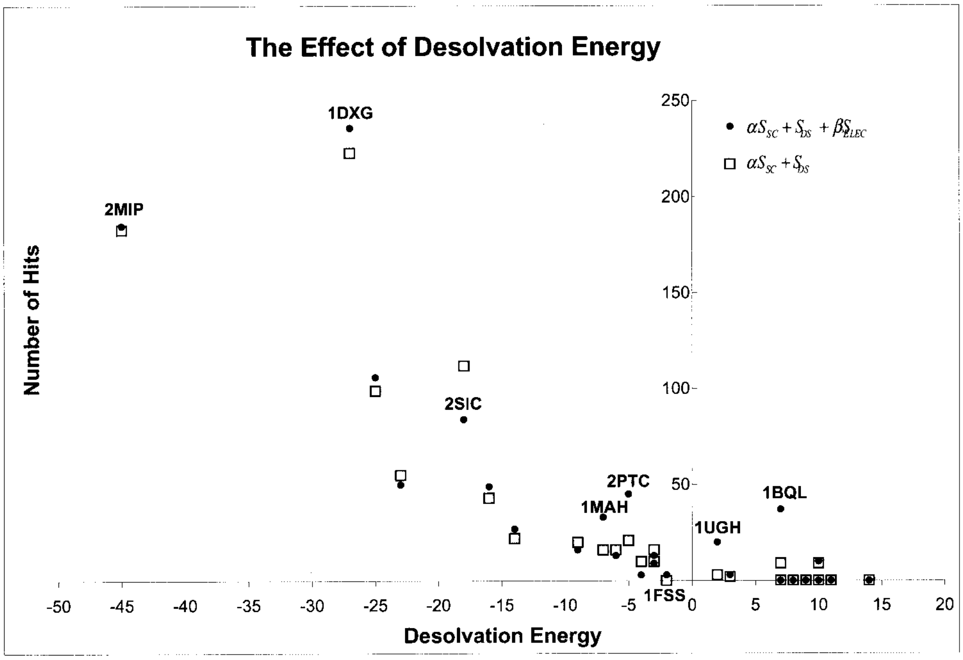

Fig. 1. The effect of desolvation energy on the number of near-native structures (hits) in the top 1,000 structures identified using target functions αS_{SC} + S_{DS} + βS_{ELEC} (solid circles) and αS_{SC} + S_{DS} (open squares). Desolvation energies (in kcal/mol) were calculated for co-crystallized complexes using pairwise ACE scores.[37] They are listed in the second-last column of Table II (labeled S_{DS}). The numbers of hits are listed in the third and forth columns of Table II (labeled αS_{SC} + S_{DS}, αS_{SC} + S_{DS} + βS_{ELEC}, respectively).

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 against the desolvation energy of the crystal complex (open squares). There is a clear correlation between the above two quantities. For all complexes that are desolvation driven (desolvation energy below −3 kcal/mol), there are at least 10 hits. In addition, as the desolvation energy becomes more negative, αS_{SC} + S_{DS} is able to identify more hits. For complexes whose desolvation energies are −3 kcal/mol or above, there are at most 10 hits. The desolvation energies for complexes without any hits are all ≤−3 kcal/mol. The results for some of these complexes are significantly improved by adding electrostatics to the target function (see Shape Complementarity, Desolvation, and Electrostatics).

Shape Complementarity, Desolvation, and Electrostatics (αS_{SC} + S_{DS} + βS_{ELEC})

The αS_{SC} + S_{DS} + βS_{ELEC}column in Table II shows the improvement that resulted from adding the simple Coulombic electrostatics energy to the target function. Compared to αS_{SC} + S_{DS}, αS_{SC} + S_{DS} + βS_{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 αS_{SC} + S_{DS} failed to find any hits for 1FSS while αS_{SC} + S_{DS} + βS_{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 αS_{SC} + S_{DS} + βS_{ELEC}). They do however all have highly favorable electrostatics energies (Table II).

In Figure 2, the difference in the numbers of hits by αS_{SC} + S_{DS} + βS_{ELEC} and αS_{SC} + S_{DS} is plotted against the electrostatics energy of the crystal complex calculated using CHARMM.[41] For the 13 electrostatics driven complexes (electrostatics energies below −17 kcal/mol), either αS_{SC} + S_{DS} + βS_{ELEC} found more hits than αS_{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 αS_{SC} + S_{DS} + βS_{ELEC} found many fewer hits than αS_{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 αS_{SC} + S_{DS} + βS_{ELEC}

Although S_{SC} generally performed much worse than αS_{SC} + S_{DS} + βS_{ELEC}, it did identify hits for complexes on which αS_{SC} + S_{DS} + βS_{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

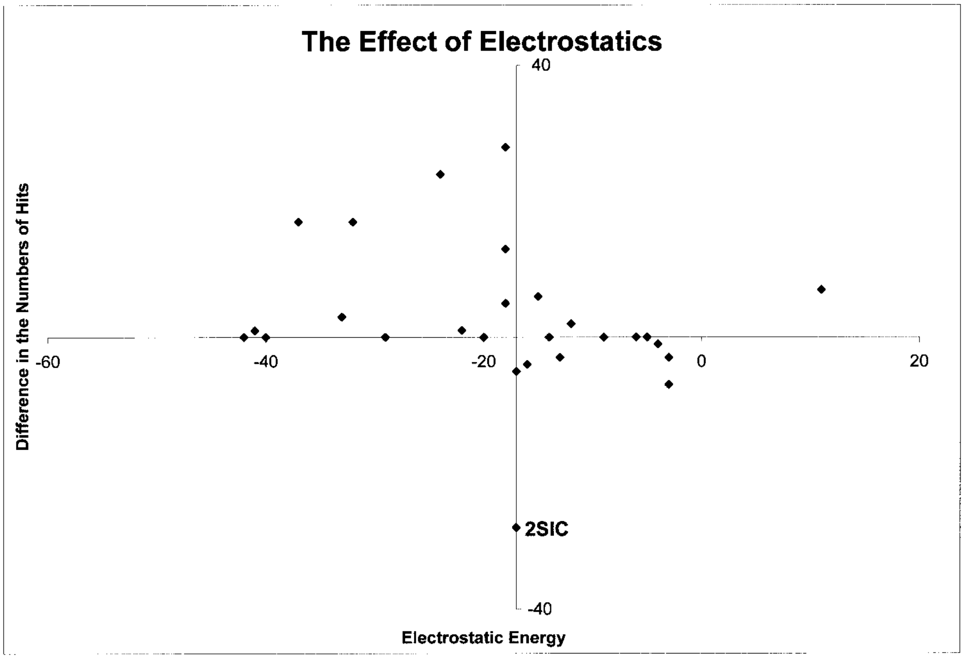

Fig. 2. The difference of the number of hits in the top 1,000 structures identified using target functions $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ and $\alpha S_{SC} + S_{DS}$ plotted against the electrostatic energy. The electrostatics energies (in kcal/mol) were computed for co-crystallized complexes using CHARMM.[41] They are listed in the last column of Table II. Electrostatics significantly increased the number of hits for complexes with energies below −17 kcal/mol. Even though the addition of electrostatics decreased the number of hits for 2SIC from 112 to 84, its complex structure was still identified unambiguously [see Fig. 3(a)].

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$\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$. 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 −4 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.

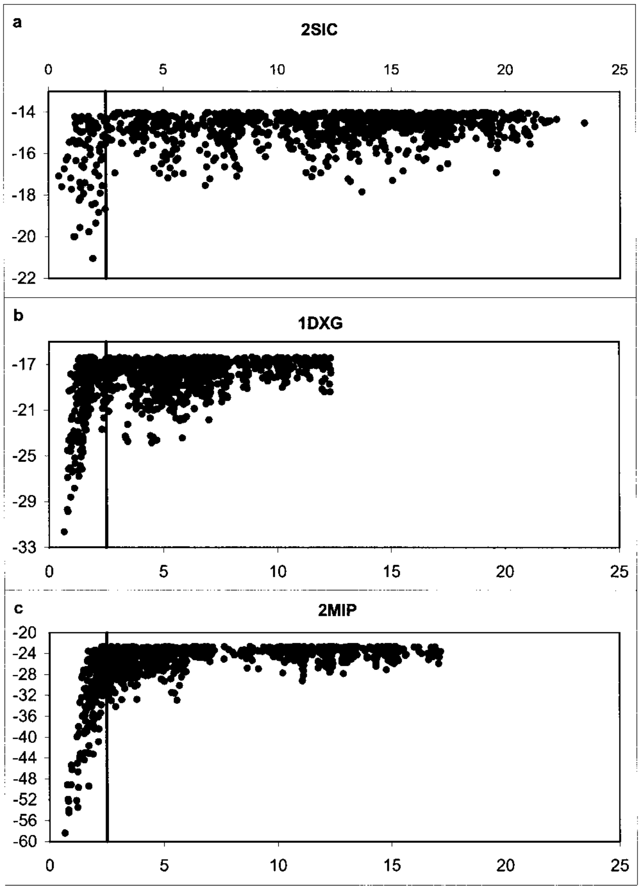

Fig. 3. $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ (ordinate) vs. RMSD (abscissa) for the top 1,000 structures of 2SIC, 1DXG, and 2MIP.

Our method failed to find any near-native structures in the top 2,000 for three complexes: 3HFM, 1VFB, and 2PCC. The first two are antibody/antigen and the last one is cytochrome C/peroxidase. All of them have highly unfavorable desolvation energies (S_{DS} = 14, 10, and 7 kcal/mol, respectively; Table II). For the cytochrome C/peroxidase complex, a heme is at the binding site. Since our desolvation free energy was derived from proteins, we did not have parameters for heme atoms. We simply used the corresponding values from protein atoms of the same type as defined by the CHARMM19 potential. This could have led to inaccurate desolvation calculations for 2PCC. This complex has highly favorable electrostatics energy (S_{ELEC} = −42 kcal/mol; Table II). We also tried to combine electrostatics directly with shape complementarity without desolvation, and were able to find hits for 2PCC, although the performance for other complexes was poor. For the two antibody/antigen complexes, even the electrostatics energies are only marginally favorable (Table II). In fact, they are not sufficient to compensate for the unfavorable desolvation energies and, as a result, S_{DS} + S_{ELEC} values are positive. The only other complex with positive $S_{DS} + \beta S_{ELEC}$ is 2JEL, another antibody/antigen. For 2JEL, shape complementarity alone was able to discover 4 near-native structures. Unfortunately, for 3HFM 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 C_β 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 S_{SC} 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.[19] and further developed by Sternberg and colleagues.[11,21,32] The key component of our target function is a desolvation term based on the Atomic Contact Energy (ACE) developed by Zhang et al.[37] 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,[21,27,28] 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.[27]

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

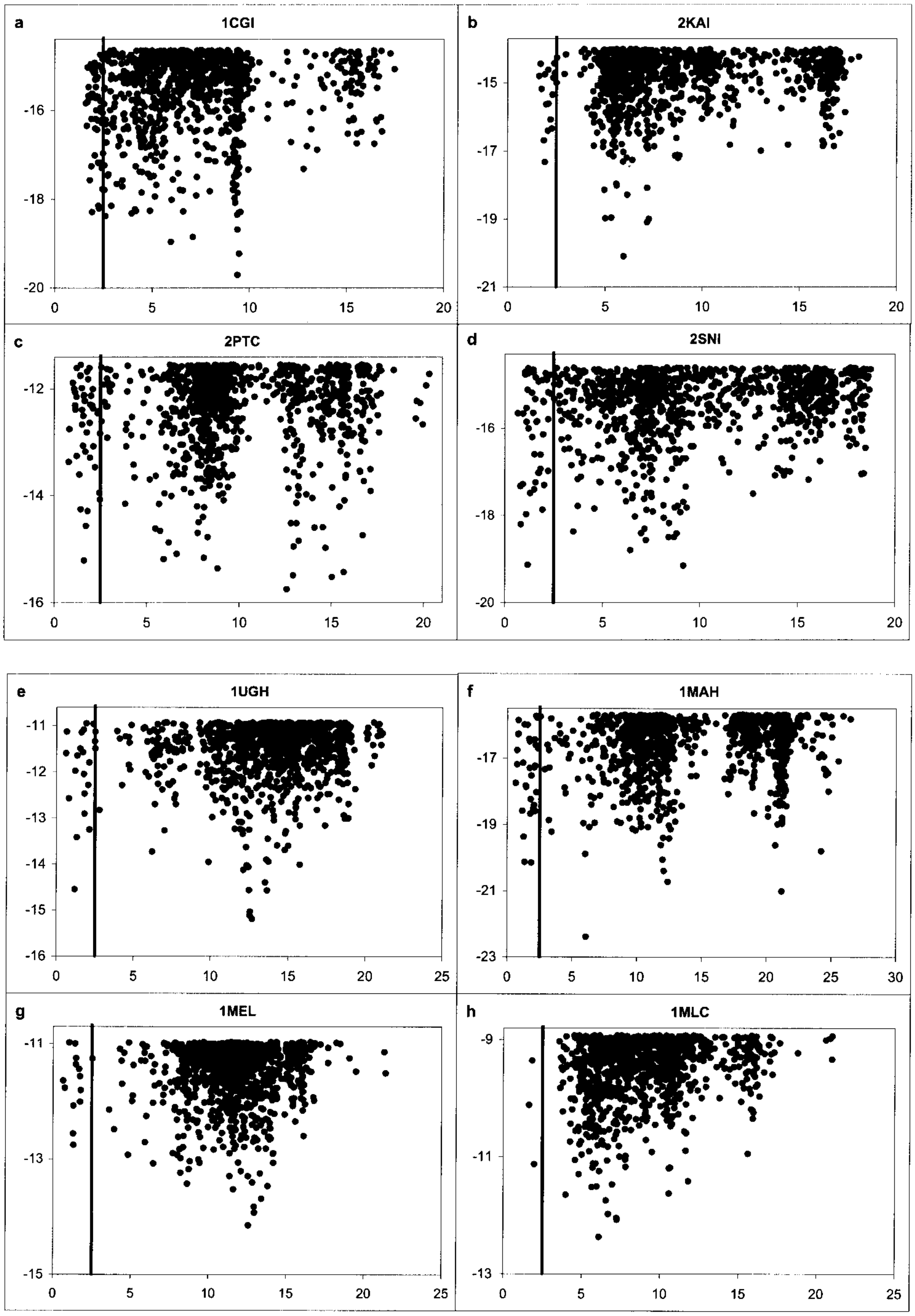

Fig. 4. $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ (ordinate) vs. RMSD (abscissa) for the top 1,000 structures of the 13 complexes that either have one or more near-native structure in top 20 or have 20 or more near-native structures in top the 2,000.

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

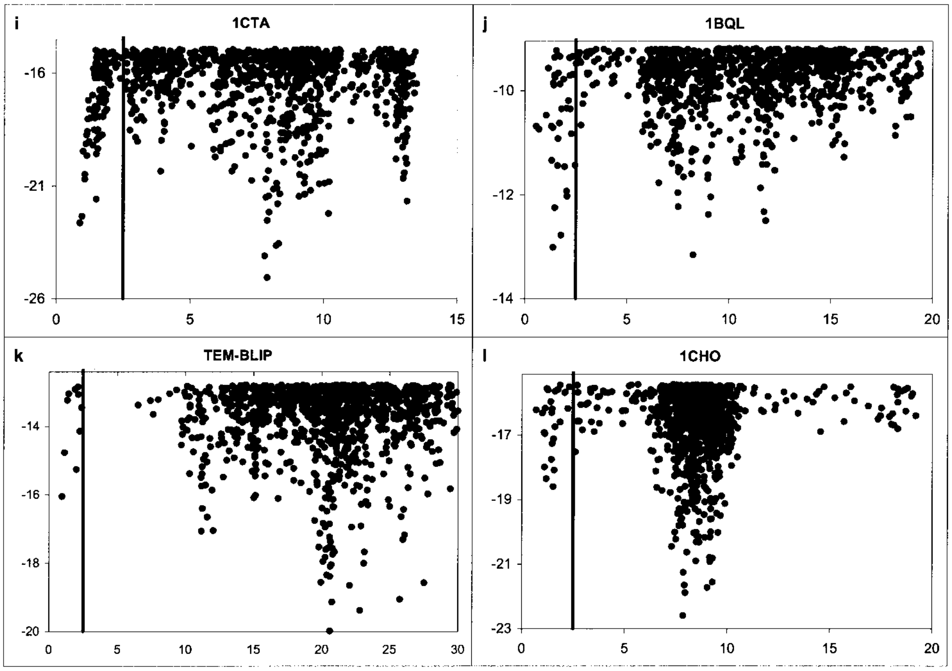

Figure 4. (Continued.)

definition requires knowledge of the complex structure, whereas our CDR definition does not. Our definition 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 defined 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, 2SNI, 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 (S_{SC}), desolvation (S_{DS}), and electrostatics S_{ELEC}. The key is the desolvation term S_{DS}. Since S_{DS} alone does not prevent steric overlap between the receptor and the ligand, in this study we combined S_{DS} with S_{SC}. Table II shows that $\alpha S_{SC} + S_{DS}$ 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 $\alpha S_{SC} + S_{DS} + \beta S_{ELEC}$ (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 -3 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_{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-

TABLE III. Comparison Between the Normal and Collapsed Side Chain Schemes[†]

	S_{SC} (number of hits in top 1,000)		$\alpha S_{SC} + S_{DS} + \beta S_{\mathrm{ELEC}}$ (number of hits in top 1,000)		S_{SC} & $\alpha S_{SC} + S_{DS} + \beta S_{\mathrm{ELEC}}$ (number of hits in top 2,000)	
Complex	Normal	Collapsed	Normal	Collapsed	Normal	Collapsed
1CGI	1	2	50	59	51	60
1CHO	1	0	27	33	28	33
2KAI	0	0	16	18	16	18
2PTC	0	0	45	38	45	38
2SNI	0	0	49	42	49	42
2SIC	0	0	84	80	84	80
1BRS	13	22	0	0	13	22
TEM-BLIP	11	11	10	10	20	20
1MAH	0	0	33	30	33	30
1FSS	0	0	3	1	3	1
1UGH	10	7	20	16	28	21
1MLC	5	8	3	3	8	11
3HFM	0	0	0	0	0	0
2JEL	4	2	0	0	4	2
1BQL	1	1	37	20	38	21
1JHL	1	3	0	0	1	3
1FBI	1	5	0	0	1	5
1IAI	0	0	13	10	13	10
1VFB	0	0	0	0	0	0
1MEL	8	8	13	12	21	20
3SDH	23	26	0	2	23	26
1DXG	90	94	236	232	324	325
6EBX	0	0	9	4	9	4
2MIP	17	14	185	187	199	199
1CTA	16	24	106	103	122	127
2PCC	0	0	0	0	0	0
1IGC	0	0	3	0	3	0

[†] Hits (or near-native structures) are defined as docked structures with interface C_α RMSD $\leq$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)[29†]

		Results in this paper			Results by Gardiner et al. (2001)[29]		
Complex	RMSD cutoff for hits (Å)	Number of hits in top 100	Best rank[a]	RMSD of the best hit	Number of hits in top 100	Best rank	RMSD of the best hit
1CGI	5.48	45	6	2.61	8	34	5.04
2KAI	5.00	15	7	4.97	8	4	4.39
2PTC	4.74	10	6	1.62	6	1	4.38
2SIC	4.38	38	1	1.93	5	25	3.91
2SNI	4.85	26	2	1.20	2	51	2.05
1CHO	4.33	2	72	1.53	9	1	1.79
1UDI	4.37	7	6	1.20	3	13	2.23
1BRS	4.47	0	—	—	2	1	3.14
TEM-BLIP	4.88	2	57	0.97	0	—	—

[†] Hits are docked structures with interface RMSD smaller than the corresponding cutoff as defined in Gardiner et al. (2001).[29] Results by Gardiner et al. are adopted from Table IV of Gardiner et al.[29] with normal angle ±30° and $J = 3.0$, the parameter set that gave them the best overall performance.

[a] Ranked according to $\alpha S_{SC} + S_{DS} + \beta S_{\mathrm{ELEC}}$.

mentarity than other complexes.[42,43] However, a recent study using Voronoi volumes concluded that these two types of complexes have comparable degrees of shape complementarity.[44] 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.[45] 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. 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[21] and Mandell et al. used 1 Å.[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). 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

1. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae* [see comments]. Nature 2000;403:623–627.
2. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci USA 2000;97:1143–1147.
3. Pandey A, Mann M. Proteomics to study genes and genomes. Nature 2000;405:837–846.
4. Weng Z, DeLisi C. Protein therapeutics: promises and challenges for the 21st century. Trends Biotechnol 2002;20:29–35.
5. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S. Structural genomics: beyond the human genome project. Nat Genet 1999;23:151–157.
6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
7. Cherfils J, Janin J. Protein docking algorithms: simulating molecular recognition. Curr Opin Struct Biol 1993;3:265–269.
8. Janin J. Protein-protein recognition. Prog Biophys Mol Biol 1995; 64:145–166.
9. Shoichet BK, Kuntz ID. Predicting the structure of protein complexes: a step in the right direction. Chem Biol 1996;3:151–156.
10. Sternberg MJ, Gabb HA, Jackson RM. Predictive docking of protein-protein and protein-DNA complexes. Curr Opin Struct Biol 1998;8:250–256.
11. Sternberg MJ, Gabb HA, Jackson RM, Moont G. Protein-protein docking. Generation and filtering of complexes. Methods Mol Biol 2000;143:399–415.
12. Norel R, Lin SL, Wolfson HJ, Nussinov R. Molecular surface complementarity at protein-protein interfaces: the critical role played by surface normals at well placed, sparse, points in docking. J Mol Biol 1995;252:263–273.
13. Fischer D, Lin SL, Wolfson HL, Nussinov R. A geometry-based suite of molecular docking processes. J Mol Biol 1995;248:459–477.
14. Meyer M, Wilson P, Schomburg D. Hydrogen bonding and molecular surface shape complementarity as a basis for protein docking. J Mol Biol 1996;264:199–210.
15. Ackermann F, Herrmann G, Posch S, Sagerer G. Estimation and filtering of potential protein-protein docking positions. Bioinformatics 1998;14:196–205.
16. Betts MJ, Sternberg MJ. An analysis of conformational changes on protein-protein association: implications for predictive docking. Protein Eng 1999;12:271–283.
17. Totrov M, Abagyan R. Detailed ab initio prediction of lysozyme-antibody complex with 1.6 A accuracy. Nat Struct Biol 1994;1:259–263.
18. Strynadka NC, Eisenstein M, Katchalski-Katzir E, Shoichet BK, Kuntz ID, Abagyan R, Totrov M, Janin J, Cherfils J, Zimmerman F, Olson A, Duncan B, Rao M, Jackson R, Sternberg M, James MN. Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase [see comments]. Nature Struct Biol 1996;3:233–239.
19. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci USA 1992;89:2195–2199.
20. Vakser IA, Matar OG, Lam CF. A systematic study of low-resolution recognition in protein—protein complexes. Proc Natl Acad Sci USA 1999;96:8477–8482.
21. Gabb HA, Jackson RM, Sternberg MJ. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 1997;272:106–120.
22. Mandell JG, Roberts VA, Pique ME, Kotlovyi V, Mitchell JC, Nelson E, Tsigelny I, Ten Eyck LF. Protein docking using continuum electrostatics and geometric fit. Protein Eng 2001;14:105–113.
23. Vakser IA. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. Proteins 1997; Suppl 1:226–230.
24. Dixon JS. Evaluation of the CASP2 docking section. Proteins 1997;(Suppl 1):198–204.
25. Jiang F, Kim S. Soft docking: Matching of molecular surface cubes. Journal of Molecular Biology 1991;219:79–102.
26. Norel R, Petrey D, Wolfson HJ, Nussinov R. Examination of shape complementarity in docking of unbound proteins. Proteins 1999;36:307–317.
27. Palma PN, Krippahl L, Wampler JE, Moura JJ. BiGGER: A new (soft) docking algorithm for predicting protein interactions. Proteins 2000;39:372–384.
28. Ritchie DW, Kemp GJ. Protein docking using spherical polar Fourier correlations. Proteins 2000;39:178–194.
29. Gardiner EJ, Willett P, Artymiuk PJ. Protein docking using a genetic algorithm. Proteins 2001;44:44–56.
30. Jackson RM, Sternberg MJE. A continuum model for protein-protein interactions: Application to the docking problem. J Mol Biol 1995;250:258–275.
31. Weng Z, Vajda S, Delisi C, Prediction of protein complexes using empirical free energy functions. Prot Sci 1996;5:614–626.
32. Jackson RM, Gabb HA, Sternberg MJ, Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. J Mol Biol 1998;276:265–285.
33. Camacho CJ, Gatchell DW, Kimura SR, Vajda S, Scoring docked conformations generated by rigid-body protein-protein docking. Proteins 2000;40:525–537.
34. Camacho CJ, Vajda S, Protein docking along smooth association pathways. Proc Natl Acad Sci USA 2001;98:10636–10641.
35. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.

36. Lattman EE. Optimal sampling of the rotation function. Acta Crystallogr 1972;B28:1065–1068.
37. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 1997;267:707–726.
38. Zhang C, Cornette JL, Delisi C. Consistency in structural energetics of protein folding and peptide recognition. Prot Sci 1997;6:1057–1064.
39. Vasmatzis G, Zhang C, Cornette JL, DeLisi C. Computational determination of side chain specificity for pockets in class I MHC molecules. Mol Immunol 1996;33:1231–1239.
40. Zhang C, Chen J, DeLisi C. Protein-protein recognition: exploring the energy funnels near the binding sites. Proteins 1999;34:255–267.
41. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4:187–217.
42. Lawrence MC, Colman PM. Shape complementarity at protein/protein interfaces. J Mol Biol 1993;234:946–950.
43. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
44. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–2198.
45. Jackson RM. Comparison of protein-protein interactions in serine protease-inhibitor and antibody-antigen complexes: implications for the protein docking problem. Protein Sci 1999;8:603–613.