WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Integrating statistical pair potentials into protein complex prediction

Julian Mintseris,[1] Brian Pierce,[1] Kevin Wiehe,[1] Robert Anderson,[1] Rong Chen,[1] and Zhiping Weng[1,2]*

[1] Bioinformatics Program Boston University, Massachusetts 02215

[2] Biomedical Engineering Department, Boston University, Massachusetts 02215

ABSTRACT

The biophysical study of protein–protein interactions and docking has important implications in our understanding of most complex cellular signaling processes. Most computational approaches to protein docking involve a tradeoff between the level of detail incorporated into the model and computational power required to properly handle that level of detail. In this work, we seek to optimize that balance by showing that we can reduce the complexity of model representation and thus make the computation tractable with minimal loss of predictive performance. We also introduce a pair-wise statistical potential suitable for docking that builds on previous work and show that this potential can be incorporated into our fast fourier transform-based docking algorithm ZDOCK. We use the Protein Docking Benchmark to illustrate the improved performance of this potential compared with less detailed other scoring functions. Furthermore, we show that the new potential performs well on antibody-antigen complexes, with most predictions clustering around the Complementarity Determining Regions of antibodies without any manual intervention.

Proteins 2007; 69:511–520.

Key words: protein interactions; protein recognition; protein interfaces; protein complexes.

INTRODUCTION

Protein function is usually described in terms of interactions with other proteins and thus ability to describe, model, predict, and design protein interactions is important for both theoretical and applied proteomics. The current state of the art in prediction of complex structures was recently described in comprehensive reviews by Smith and Sternberg[1] and Halperin *et al.*[2] Both identified three key ingredients: (1) representation of the system, (2) global conformational space search, and (3) reranking of top solutions based on a scoring function. Halperin *et al.* notes that these steps are similar to the traditional approach to protein folding. Indeed they are typical of many approaches to problems in computational biology. In this work, we have chosen the popular fast-fourier transform (FFT) approach to docking and focused on optimizing the protein representation to attain an efficient way of incorporating accurate scoring functions into the search procedure.

Use of FFT correlation for protein docking, first proposed by Katchalski-Katzir *et al.*,[3] requires mapping the atoms of both molecules onto a 3D grid and assigning values to the grid cells, which, in the process of computation, evaluate to various components of the docking scoring function. It has been successfully shown that this approach can rather naturally evaluate shape complementarity and electrostatics energies.[4–6] Essentially, this is possible because these energy calculations can be broken down into two components, each placed as a value on one of the two docked molecules, allowing them to evaluate to the final score in the context of an FFT correlation computation. When it comes to implementing a pair-wise calculation, such as a statistical potential score, the identities of the grid cells (or protein atoms) become important. This means that in order to use the FFT correlation method to evaluate the statistical potential score using a 20 × 20 amino acid interaction preference table, we would need to construct individual FFT grids for each amino acid (or atom type). In the past, our ZDOCK program used an averaged version of statistical potentials, compressing the 18 × 18 atomic contact energies (ACE) table down to an 18-long vector and using that as an approximation of the pair potential ACE score.[7] Here, we show that despite the high reported correlation between pair-wise and averaged potential scores, there is a significant amount of information lost through this approximation.

Taking advantage of the fact that the FFT evaluates both real and imaginary parts of a complex function, it is possible to compute the pair-wise potential score of two amino acids or atom types with each FFT correlation. Nevertheless, the computational cost of 10 FFT correlations for the 20 amino acid alphabet or 9 FFT correla-

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

*Correspondence to: Zhiping Weng, Department of Biomedical Engineering and Bioinformatics Program, Boston University, 44 Cummington St., Boston, MA 02215. E-mail: zhiping@bu.edu

Received 22 September 2006; Revised 2 March 2007; Accepted 8 March 2007

Published online 10 July 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21502

tions for 18 ACE atom types remains prohibitive. Using the results of our recent work on derivation of optimized atom types from protein interfaces for any alphabet size, we found that we can decrease the number of atom types needed to represent protein structures without any loss of accuracy in the pair-wise score evaluation.

Furthermore, we built upon previous work by Moont *et al.*[6] to derive a "surface fraction" pair potential, which evaluates the interaction preferences of protein surface atoms based on a curated set of transient protein interfaces.[8]

The final results show a significant improvement in performance of ZDOCK on a nonredundant benchmark of 63 protein docking cases, while taking advantage of a reduced atom type representation to alleviate computational costs.

AVERAGED AND PAIR-WISE STATISTICAL POTENTIALS

Knowledge-based statistical potentials have been used extensively and successfully for a variety of problems in computational biology – protein structure prediction as well as prediction of interactions of proteins with other proteins, small molecules, and DNA. They were first introduced three decades ago by Levitt and Warshel[9] and Tanaka and Scheraga[10] and further developed by Miyazawa and Jernigan,[11] Sippl,[12] and many others. The interpretation of statistical potentials most often relies either on statistical Bayesian approaches or on the more energetically appropriate Inverse Boltzmann Law, but regardless of the details, the general formulation is as follows:

$$E_{\text{structure}} = \sum_{i,j} \text{Observed Structure}_{i,j} \log\left(\frac{\text{Observed Database}_{i,j}}{\text{Reference}_{i,j}}\right), \quad (1)$$

where i and j are amino acid or atom types. Given a residue or atom-level protein representation with k types, we define n_i as the number of atoms of types i (1...k; 0 for solvent) in one interacting protein, n_{ij} as the number of contacts between two atoms of different types. Numerous variations on this theme exist in the literature but most energy expressions take the more specific form

$$e_{ij} = \ln\frac{n_{ij}}{c_{ij}}, \quad (2)$$

where we define ensemble values for native atomic contacts

$$n_{ir} = \sum_{j=1}^{k} n_{ij}; \;\; n_{rr} = \sum_{i=1}^{k} n_{ir}; \;\; N = \sum_{i=1}^{k} n_i \quad (3)$$

where the subscript r refers to any protein atom.

In the context of the FFT, where each grid cell can represent only one atom or residue type, a full pair-wise calculation would require a separate correlation between an atom/residue type on one protein and all atom/residue types on the interacting partner. An average contact energy for each atom type can be defined as

$$e_i = \frac{1}{n_{ir}} \sum_{j=1}^{k} e_{ij} n_{ij} \quad (4)$$

Calculation of the interaction energy using this kind of averaged potential can be achieved in a single FFT correlation since the FFT grid can contain all the atom/residue type labels simultaneously. However, as we will see, the loss of information upon averaging the potentials also results in substantial performance cost. In this study, we propose some approaches that strive to achieve a more nuanced balance between computational cost and algorithm performance. To do this, we turn our attention to parts of the above equations, which are often neglected—the identities of i and j.

OPTIMIZED ATOM TYPE ALPHABETS

Despite the extensive work on statistical potentials, one aspect that so far has defied rigorous treatment is the definition of i and j in Eq. (1). Essentially, this is an issue of finding an appropriate way to represent protein structures. Many studies use 20 amino acids—the intuitive choice. Earliest efforts in this area by Warme and Morgan[13] started shortly after statistical potentials for protein folding were first introduced. Considerable theoretical work exists on the simplest of alphabets involving just two groups- the so-called HP (Hydrophobic-Polar) model.[14,15] Most studies have used a variety of chemical, physical, and biological properties of amino acids and their functional groups to derive atom type schemes, while incorporating different investigators' understanding of protein energetics.[7,16–18] Recently, we addressed this problem in a rigorous, data-driven manner, deriving atom type schemes over a range of alphabet sizes from nonredundant sets of proteins and protein complexes (http://www.jsbi.org/journal/IBSB04/IBSB04F021.html).[19] Briefly, we used mutual information (MI) as an optimization criterion to find sets of atom types that have the most distinct (informative) protein environments.

$$\text{MI} = \sum_{i,j} P(i,j) \;\; \log\frac{P(i,j)}{P(i)P(j)}, \quad (5)$$

where $P(i,j)$ is the joint probability that an atom of type i forms a contact with an atom of type j, and $P(i)$ and $P(j)$ are the marginal probabilities. Note the similarity

DOI 10.1002/prot

between the definition of MI and the statistical potential in Eqs. (1) and (2).

Interpretation of the reduced representation problem in information-theoretic terms is straightforward. MI between two variables I and J (representing a grouping of the protein atom types) is a measure of how much information one variable reveals about the other. If i and j are instances of I and J, where the number of such instances is governed by the size k of the atom type alphabet, we want to define i and j such that the MI is maximized. Each instance i or j is a grouping of protein atoms of one type. It is easy to see from Eq. (5) that if i and j are chosen randomly, the probability of the joint distribution would be equal to the product of marginal distributions resulting in zero MI. On the other extreme, the maximum possible MI for a given alphabet size can be determined if we take $P(i,j) = P(i) = P(j) = 1/k$. Equation (5) then reduces to:

$$\mathrm{MI}_{\max} = \sum_{i,j} P(i) \log \frac{P(i)}{P(i)P(j)} = \log(\text{alphabet size}) \quad (6)$$

Another way to think about this is to realize that grouping atoms with similar biochemical properties—atoms that are commonly found in protein structures in similar environments, tends to increase MI by increasing the certainty that a specific atom type will occur in a given protein environment. Thus MI is a rigorous and intuitive measure suitable for optimization.

Notice that MI is also a measure of independence. If the variables I and J are randomly distributed, they reveal no information about each other, as shown earlier. Assuming under a null hypothesis (H0) that I and J are independent and an alternative hypothesis (H1) that they are not, it can be shown that a log likelihood ratio test is exactly equivalent to the definition of MI.[20]

In the statistical context of the test of independence, the objective of finding the representation with maximum MI is equivalent to maximizing the significance of the test of independence between the atom types. The problem of finding such an optimal reduced protein representation for a given target alphabet size is essentially equivalent to maximum likelihood estimation. We have a model as described earlier and a comprehensive nonredundant dataset of proteins and protein complexes. The distributions of heavy atoms into a given number of atom types, or the probabilities of membership of each heavy atom in an atom type group are the parameters to estimate. An exhaustive solution to this problem is impossible: the number of ways of distributing k objects into m bins grows very quickly. We use Monte Carlo methods to estimate the best reduced representations by randomly perturbing the bin memberships and accepting/rejecting based on the Metropolis criterion. This is in many ways similar to K-means clustering but here we optimize a metric global to the whole dataset as opposed to a distance from every point to some cluster center.

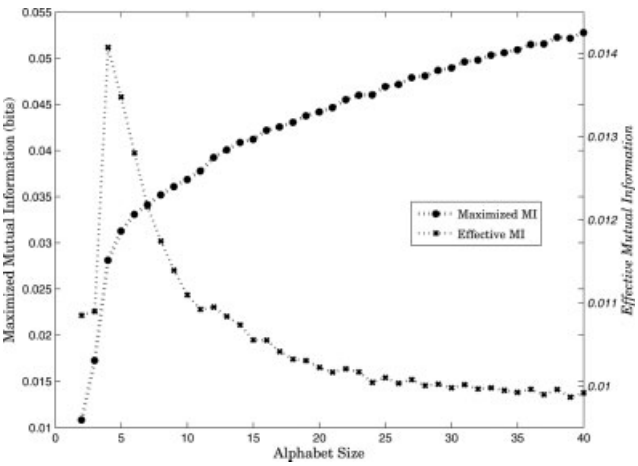

Figure 1

Maximized mutual information increases monotonically with alphabet size. Normalizing for maximum information available for a given alphabet size reveals a peak for atom types. See text for details.

Figure 1 shows that optimized MI increases monotonically with increasing alphabet size. This is to be expected given the definition of MI in Eq. (5). To get a better sense of how the optimized MI relates to maximum possible MI for a given alphabet size [calculated from Eq. (6)], we also plot the ratio of optimized MI to this maximum—the effective MI—in Figure 1. The strongest peak of Effective MI occurs at alphabet of size four, corresponding to hydrophobic, polar, positive, and negative types.[19] This suggests that the major energetic driving forces in protein interactions are made up of the hydrophobic effect and charge interactions. Further subdividing the atom space into larger alphabets leads to quickly diminishing returns, with effective MI declining below the level of the first two atom types past the alphabet of size 12. The breakdown of these 12 atom types is presented in supporting Table I and we revisit them later.

INTERFACE ATOMIC CONTACT ENERGIES

Potentials are based on the idea that the native conformations exist at the global thermodynamic minimum. The parameters describing the native state are easy to compute by surveying the current structure databases. The somewhat controversial part of statistical potentials is in defining the so-called reference state, which in protein structure prediction implies the "unfolded" state presumably at a very high energy.[21] Defining such a reference state is difficult since proteins do not really exist in completely "unfolded" states under physiological condi-

tions and it is very difficult to determine the properties of whatever high energy state they do occupy. Interestingly, the definition of a reference state for protein–protein docking, at least for transient interactions of independently folded proteins, is much simpler and less controversial. This goes along with the fact that the definition of docking decoys is also much more straightforward. In docking, we are only making one assumption—that the two (or more) molecules in question, do, in fact, interact. Given this assumption, the reference state can be thought of as the average of a distribution of all possible docking conformations, with all possible surface patches touching each other. For any two proteins, a large number of such conformations can be generated using many of the existing docking algorithms relying just on shape complementarity. However, doing this for a sufficiently large number of protein complexes is still rather computationally expensive. Instead, we can define the reference state mathematically. This can be done several ways and we choose to follow the formulation developed by Miyazawa and Jernigan[11] for residue level protein folding potentials and subsequently adapted for atom-level potentials.[7]

Moont *et al.* identify two simple ways to compute the reference state - expected residue pairs at the interface of two interacting proteins[6]:

$$c_{(\text{mole}-\text{fraction})i,j} = n_{rr}\frac{n_i}{N}\frac{n_j}{N};\;\; c_{(\text{contact}-\text{fraction})i,j} = n_{rr}\frac{n_{ir}}{N}\frac{n_{jr}}{N} \tag{7}$$

Using the mole-fraction method, the authors assume the number of contacts in the reference state to be proportional to the product of the atom types (or residues) in the pair of proteins. For the contact-fraction method, the expected reference contacts are proportional to the propensities of the native contacting atom types (residues). The most obvious shortcoming of these approaches with respect to protein docking is that the contact fraction method ignores any portion of the protein not in contact in a given docking conformation, while the mole-fraction method incorporates all residues, even those in the protein core. The authors conclude that the mole-fraction method performs better in re-ranking docking predictions.[6] This is not surprising since that approach takes into account a greater fraction of accessible surface in the protein. To take the next step toward a more accurate and realistic modeling of protein interaction in terms of statistical potentials, we need to include only the atoms that are exposed to the solvent and thus have the potential to become part of the protein interface. We can find these atoms by calculating the average number of contacts a given atom type is capable of forming when completely buried (coordination number) and then comparing to the observed number of contacts in the protein of interest.[7,11]

As in the previous works, we use atom type-specific coordination numbers q_i to define a relationship between the number of atoms and the number of atom-atom contacts within in a 6 Å sphere:

$$q_i n_i = \sum_{j=1}^{k} n_{ij} + n_{i0} \tag{8}$$

We further define n_{i0} to represent contacts with water and then use the definition of coordination number to derive the solvent-accessible atoms on the surface:

$$n_{i0} = q_i n_i - \sum_{j=1}^{k} n_{ij};\;\; n_{r0} = \sum_{i=1}^{k} n_{i0} \tag{9}$$

We compute the contacts for our reference state representing all possible contacts between the surface atoms of the interacting proteins. By extrapolation from Moont *et al.*[6] this could be called "surface fraction" potential:

$$c_{\text{surface}-\text{fraction};ij} = \frac{\sum n_{i0;p1}\, n_{j0;p2}}{\sum_1^k n_{i0;p1} \sum_1^k n_{i0;p2}} \cdot n_{rr;c} \tag{10}$$

The summations in the above equation run over all the complexes c and their interacting components p_1 and p_2. Combining Eqs. (2) and (10) allows us to define the Interface Atomic Contact Energies

$$e_{\text{IFACE};ij} = \ln \frac{n_{ij}}{c_{\text{surface}-\text{fraction};ij}}, \tag{11}$$

The non-redundant set of protein complexes used to derive the statistical potentials includes 150 complexes. These complexes were obtained by filtering the list obtained in Mintseris and Weng[8] to remove rigid-body complexes in the Docking Benchmark.[22]

To evaluate the effectiveness of various pair potentials to discriminate near-native protein complexes from decoys, we used 63 complexes taken from the Protein Docking Benchmark. These cases were classified as "rigid-body" because of small amount of conformational change occurring upon binding. We used an early version of the ZDOCK algorithm,[4] which used only shape complementarity to produce 54,000 conformations for each benchmark case. Forty one of these 63 cases produced at least one hit and were then re-scored using both the mole-fraction potential from Moont *et al.*[6] and the surface-fraction potential described above. Both were derived using our new 150 complex dataset over a range of atom type alphabets. Results are presented in Figure 2 in terms of Area under the ROC (Received Operator Characteristic) curve. Clearly, the surface-fraction potential is better at ranking near-native hits (as defined in Methods) than the mole fraction potential. The pair-wise potential using the full matrix is significantly better than the averaged

 DOI 10.1002/prot

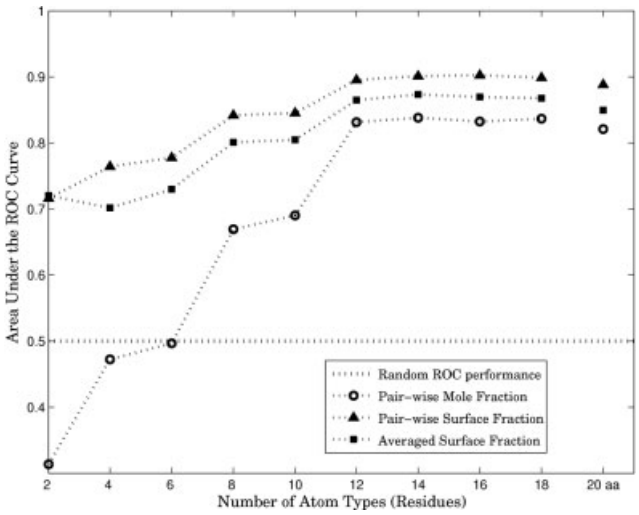

Figure 2

Starting with ZDOCK predictions based solely on shape complementarity, the complexes are reranked based on the mole-fraction potential of Moont et al. *as well as the surface-fraction potential presented in this work. Note that the performance levels off after 12 atom types and reaches levels slightly higher than that of the original 20 amino acids*

one using just a vector to represent atom type-specific energy preferences (*P*-value 0.0078 according to the paired-sample Wilcoxon signed rank test). Finally, we see in Figure 2 that the effectiveness of all potentials flattens out at alphabet of size 12. A sample plot of ROC curves is presented in Supporting Figure 3.

INTEGRATING POTENTIALS INTO FFT PROTEIN–PROTEIN DOCKING

In the previous version of the ZDOCK algorithm, shape complementarity (SC), electrostatics (EL), and desolvation were incorporated in the FFT framework.[4] The desolvation component of the scoring function was a simplified version of the ACE energy parameters,[7] where the parameters used were averaged over each atom type [Eq. (4)], thus allowing the calculation to be completed in a single FFT correlation. For details, the reader may refer to the works of Chen and Weng,[4,5] Moont *et al.*[6] as well as the seminal paper by Katchalski-Katzir *et al.*[3]

Briefly, each interacting protein is mapped to a three-dimensional grid. The cells of the grid are assigned appropriate values representing qualities of the protein such as desolvation parameters, partial or full charges, or values representing surface exposure. A separate 3D function is needed to represent each of these physical parameters but some may be combined by taking advantage of the real and complex parts and their resulting complex product. For any given starting orientation of the two molecules, FFT is then used to speed up the translational 3D search, thus calculating the correlation and the value of a parameter in question such as shape complementarity, electrostatics, or desolvation.

Here, to take advantage of the information in our newly derived pair-wise potential energies, we use both the real and imaginary parts of k/2 FFT correlation functions to compute the sum of all pair-wise energies over k atom types. To make a fair comparison with previous ZDOCK versions, we first used the 18 ACE atom types.

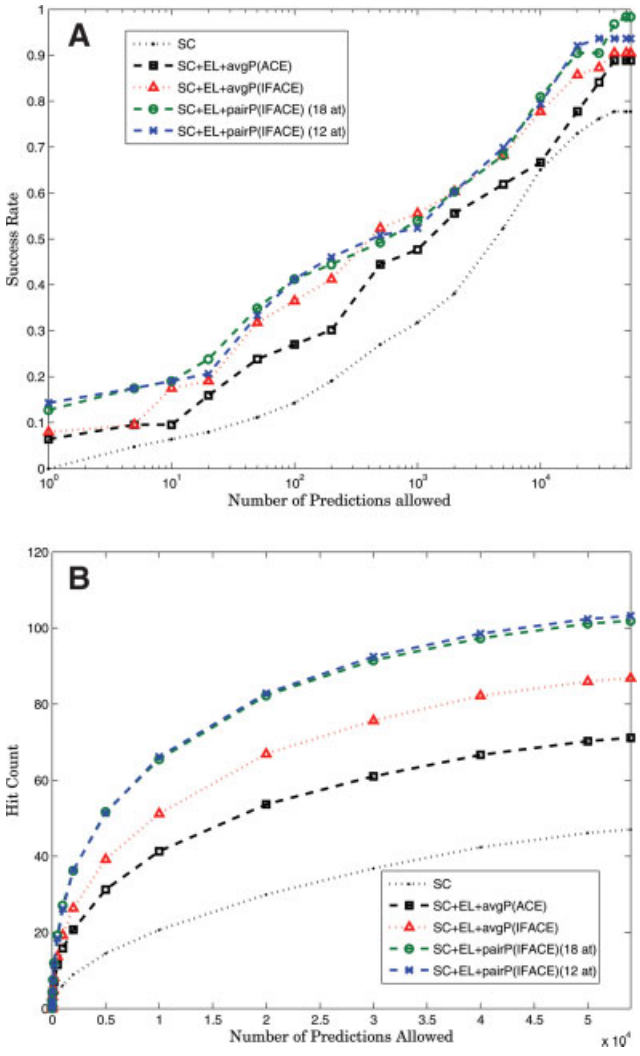

Figure 3

*Comparison of (**A**) success rates and (**B**) hit countsobtained from different ZDOCK scoring functions, using 6° sampling. Success rate is defined as the fraction of tested cases with at least one near-native conformation given some number of predictions. Hit count is defined as the total number of near-native conformations among cases tested given some number of predictions. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

Thus, if we define 9 discrete functions for each atom of type $i\{1, 3, 5, 7, \ldots, 17\}$ in a protein ligand L:

$$\mathrm{Re}[L_i] = \begin{cases} 1 \text{ if grid cell is occupied by a ligand atom of type i} \\ 0 \text{ otherwise} \end{cases}$$

$$\mathrm{Im}[L_i] = \begin{cases} 1 \text{ if grid cell is occupied by a ligand atom of type } (i+1) \\ 0 \text{ otherwise} \end{cases} \quad (12)$$

and 9 discrete functions for each possible atom type $i\{1, 3, 5, 7, \ldots, 17\}$ in contact with a protein receptor R atom of type j:

$$\mathrm{Im}[R_i] = \begin{cases} \sum e_{\mathrm{IFACE};i,j} & \text{Neighbors within 6 Å,} \\ 0 & \text{non-neighbor atoms,} \end{cases}$$

$$\mathrm{Re}[R_i] = \begin{cases} \sum e_{\mathrm{IFACE};(i+1),j} & \text{Neigbbors within 6 Å,} \\ 0 & \text{non-neighbor atoms} \end{cases} \quad (13)$$

The sum of resulting FFT correlations of a cubic grid will give the total desolvation energy summed over all atom types

$$E_{\mathrm{IFACE}} = \sum_{i=1}^{k}\sum_{j=1}^{k} e_{ij} n_{ij} \cong \sum_{i=1,3,5,\ldots}^{k/2} \left[\sum_x \sum_y \sum_z L_i \times R_i \right] = \sum_{i=1,3,5,\ldots}^{k/2} \left[\sum_x \sum_y \sum_z \mathrm{Re}[R_i] \times \mathrm{Im}[L_i] + \mathrm{Im}[R_i] \times \mathrm{Re}[L_i] \right] \quad (14)$$

where the imaginary part of the complex product evaluated as a result of the correlation accomplishes the summation of the energy components over atoms in contact. The new version of the ZDOCK algorithm was implemented in C++ to perform in parallel using MPI. We tested the implementation on 63 rigid-body cases from the Docking Benchmark. The results, obtained with 6° sampling, are presented in Figure 3 in terms of success rate and hit count. Success rate is defined as the fraction of all cases producing at least one near-native hit given a certain number of predictions. Hit count is the number of near-native hits obtained given a certain number of predictions. Near-native hits were assigned as described in Methods. The first observation to be made from Figure 3, is that the newly derived potentials perform much better than ACE across the entire range of allowed prediction numbers (P-value 1.5×10^{-5} when compared with pair-wise 18 atom type potential). Secondly, when comparing within the new potential group, pair-wise versions of the potential outperform the averaged ones (P-value 0.004 for 18 atom-type IFACE averaged vs. pair-wise potentials), especially when we allow a small number of predictions. The probability of a 1st-ranked near-native hit almost doubles with the pair-wise scoring scheme [Fig. 3(A)]. Although the performance in success rate tends to even out with greater number of predictions, the hit rates are clearly higher for the pair-wise potentials across the range of predictions [Fig. 3(B)]. Thirdly, we note that the difference between pair-wise potentials with 18 and 12 atom types is not statistically significant, as was predicted by preliminary analysis in Figure 2. Detailed breakdown of the 12 optimized atom types as well as the derived potential is described in Supporting Tables I and II.

PERFORMANCE OF PAIR POTENTIAL FFT DOCKING ON ANTIBODY–ANTIGEN CASES

In all the results presented so far, we have not used any additional biological information, which is often known to significantly improve the results. In particular, for antibody-antigen complexes, the common assumption that the binding of antibody involves the Complementarity Determining Regions (CDRs) usually holds and points to that relatively small portion of the antibody surface as the likely binding region. In the CAPRI (Critical Prediction of Protein Interactions) experiment, most groups use this assumption successfully. However, as most groups found out in one of the early CAPRI rounds, the CDR assumption does not hold absolutely - a few of the antibody-antigen complexes turned out to involve the interaction of significant portions of the framework regions. These special kinds of antibodies are found in camelids and are made up of a single chain. The framework residues (non-CDR) of camelid antibodies have been shown to constitute 25–40% of the interface.[23] Since most groups restricted the search to CDRs, predictions for those complexes were not successful.[24,25]

In order to see the effect of blocking non-CDRs and to compare this effect across the variations of the algorithm, we repeated the analysis above for the rigid-body antibody-antigen cases in the benchmark with the non-CDR residues blocked. To help focus on the differences, Figure 4 shows the difference in success rate and difference in hit count, which we obtained by counting the hits with blocking and subtracting the results in Figure 3. Figure 4(B) clearly illustrates that the effect of blocking non-CDR residues decreases for more sophisticated versions of the algorithm. In other words, the better the original algorithm, the less useful the additional biological information. Another striking result evident from Figures 4(A,B) is the difference of the effect of additional biological information on the performance of ZDOCK

DOI 10.1002/prot

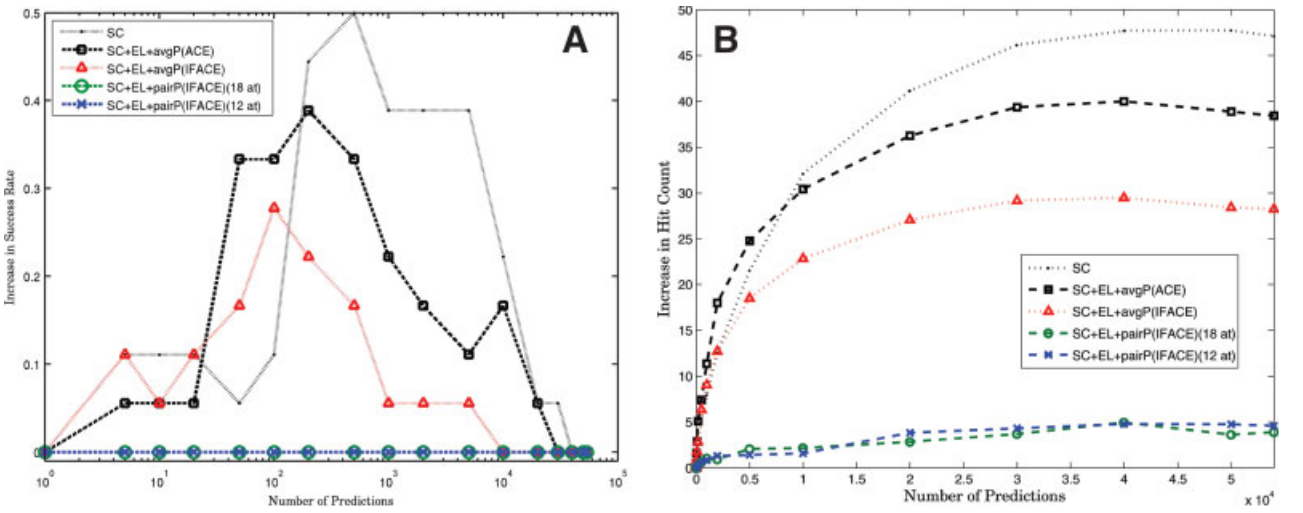

Figure 4

*(**A**) A subset of 20 antibody–antigen complexes was used to compare the effect of blocking non-CDR residues. Success rate (A) and hit count (**B**) without blocking is subtracted from that after blocking. Note that the difference is zero for all algorithms at high number of allowed predictions because all algorithms find at least one hit in both blocked and unblocked versions. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

with and without pair-wise statistical potentials. The success rate plot shows that for pair-wise potentials the effect is entirely absent across the whole range of allowed predictions [Fig. 4(A)]. From the hit count plot [Fig. 4(B)], it is clear that no more than 5 additional hits are obtained across the entire range of allowed predictions—a very small improvement, when compared with other versions of the algorithm. Since antibody-antigen complexes make up a substantial fraction of the benchmark rigid-body test cases, we present the results with these cases removed in Supporting Figure 1. Results indicate that other types of complexes also show substantial improvement.

To further understand the effect of new potentials on docking antibodies, we visualized the spatial distributions of hits and noticed that almost all predictions involved

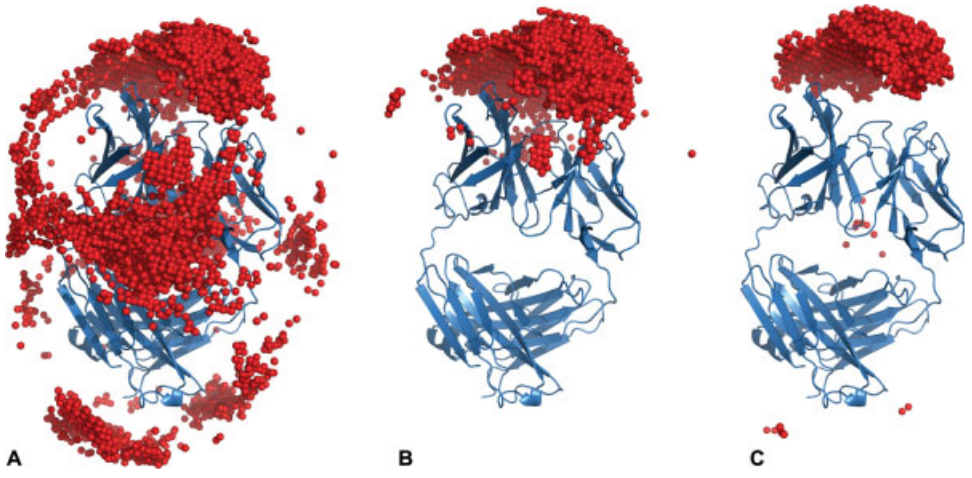

Figure 5

*Distribution of top 10000 orientations of lysozyme antigen (red) produced by docking to a Fab (blue) using (**A**) non-pair-wise (averaged) potential ZDOCK without blocking non-CDR residues, (**B**) non-pair-wise (averaged) potential ZDOCK with blocking non-CDR residues, (**C**) pair-wise potential ZDOCK without blocking non-CDR residues. The spatial distributions of the antigen are denoted by the location of a single atom in its center. CDRs of the antibody are located near the top of the molecule. The native antigen structure would be located directly above the center of mass of the antibody.*

the antigen binding in the CDRs of antibodies without blocking. Figure 5 shows that in a set of representative results from a Fab-Lysozyme complex, the results of ZDOCK with pair-wise potentials without blocking are much more restricted to the CDR than those of non-pair-wise potential with blocking. The same pattern holds for most other antibody-antigen cases in the benchmark, suggesting that the new potentials, perhaps stemming from the residue preferences in the β-sheet framework regions of antibodies lead to unfavorable interaction energies.

Our benchmark contains one representative camelid antibody-antigen complex - a bound-unbound test case from CAPRI with framework residues participating in recognition. A ZDOCK run with pair-wise potentials and no blocking produced a set of conformations not entirely restricted to the CDRs (Supporting Fig. 2). Indeed, many of the top 10,000 complexes include extensive contacts with the framework regions.

COMPARISON WITH RECENT STUDY

As this manuscript was under review, a new paper was published in this journal by Kozakov *et al.*,[26] which merits its comparison with the work presented here. The unifying theme in both papers is the move toward using pair-wise statistical potentials in the FFT framework. Both manuscripts also acknowledge the importance of reference state definition in a potential. Kozakov *et al.* introduce a new class of docking potentials called DARS (Decoys As Reference State). We believe it is similar in spirit to our surface fraction approach, although very different in implementation. Kozakov *et al.* enumerate a large number of incorrect complex conformations and use the collected interfaces as an average reference state, whereas here we use an implicit mathematical model to achieve a similar goal. They evaluate performance on enzyme-inhibitor and antibody–antigen subsets of the old Benchmark 1[27] and the new (superseding) Benchmark 2.[22] Because in this work we do not use any clustering or post-processing, direct comparison of the results is difficult. We could, however, make a direct comparison between the DARS contact energies presented in Table I of Kozakov *et al.* and our Interface ACE derived using the surface fraction method since both used the 18 ACE atom types. The correlation between the two was 0.75 and highly significant with *P*-value <0.00001. It improves to 0.79 upon removal of one outlier (Supporting Fig. 4). This is remarkable considering that the correlation between our IFACE derived using the surface fraction method and the mole fraction method is only 0.42. This shows that, as we suggested in the IFACE section above, the mathematical derivation of the surface fraction method is approximately equivalent to a reference state obtained from a large number of decoy conformations. It is likely that the remaining difference can be explained by the different sets of protein complexes used in the derivation. The set used by Kozakov *et al.*, while larger, includes many homodimers and obligate heterodimers. The protein complexes of greatest interest to docking are transient interactions between independently folding units. They have been shown to have different evolutionary[8] and physico-chemical[28] properties. The protein complexes used here for derivation of potentials are taken from a curated set of non-redundant transient protein complexes.

DISCUSSION AND CONCLUSIONS

Knowledge-based statistical potentials have been used with some success in various problems in structural computational biology. They have also been used for post-processing and re-scoring of protein–protein docking results. Here, we derive the potentials from a curated high quality non-redundant dataset of transient protein complexes, based on realistic reference state assumptions, and show that directly integrating pair-wise potentials into the FFT docking algorithm results in significant improvement. Integration of residue-level potentials would require substantial increases in computational power but this power is directly proportional to the complexity of protein representation. We then show that with an optimally chosen reduced representation, we can achieve a balance by bringing the computational cost down while keeping the algorithm performance high.

We also examine the performance of the algorithm on a subset of antibody-antigen docking cases and find that our new potentials without any surface blocking perform as well as the old potentials with additional antibody-specific biological information. Furthermore, our camelid example illustrates the power of the pair-wise potential in dealing with antibody cases, where the extent of framework residue interaction with antigen is uncertain.

METHODS AND DATA SETS

Datasets

We have previously described a set of non-redundant transient protein–protein complexes.[8] A subset of those complexes, for which interacting partner have been independently crystallized, was used to compile the Protein Docking Benchmark.[22] 63 "rigid-body" cases with little conformational change upon binding were used in this study for validation of algorithm performance. The data set used to derive statistical potentials was obtained by starting with the original transient dataset and removing cases corresponding to the Benchmark cases, thus ensur-

 DOI 10.1002/prot

ing that the "training" and testing sets do not intersect. This resulted in a set of 150 complex interfaces.

Derivation of atom type alphabets

For each atom type alphabet of size k, the Monte Carlo simulation begins by distributing 167 amino acid atoms into k bins (atom types). At every step, one atom is taken from a random bin and transferred into a different random bin, ensuring that no bins are empty. MI is calculated from a set of non-redundant protein complexes. The move is accepted or rejected based on the Metropolis criterion. We ran each simulation for 1 million steps using the modified Lam schedule for simulated annealing.[29] Other details have been reported in previous work (http://www.jsbi.org/journal/IBSB04/IBSB04F021.html).[19]

Coordination numbers

To compute coordination numbers for each atom type, we used a dataset of 808 nonredundant protein structures with sequence identity $< =$ 20% and resolution of 1.8 Å or better from the PISCES database.[30] Using only atoms that were completely solvent-inaccessible (as determined by NACCESS[31]), we computed the number of contacting atoms less than 10% accessible within 6.0 Å that do not belong to chain neighbors. The definition of chain neighbors was taken from Zhang *et al.*, which we believe is appropriate for atom (rather than residue) level applications.[7] All residues were renumbered using the S2C database[32] to ensure correct chain neighbor assignment.

Definition of near-native docking hits

In previous work, a "hit" was usually defined as a near-native docking conformation with interface RMSD (iRMSD) $< =$ 2.5 Å. Over the last several years, the docking community has been using an evolving standard measure of performance, separately defined for "high", "medium", and "acceptable" hit quality used in the evaluation of Critical Assessment of Prediction of Interactions (CAPRI) as described in Mendez *et al.*[33] The measure relies on the combination of RMSD and native contact fraction criteria. To simplify the criteria, we combine the definition of "high" and "medium" quality resulting in a Boolean expression that defines hits as follows:

$$\text{"high"} + \text{"medium" hit} = (\text{iRMSD} \leq 2\text{Å} \cup \text{IRMSD} \leq 5\text{Å}) \wedge (f_{\text{nat}} \geq 0.3) \quad (15)$$

In an effort to conform to an emerging standard we changed our definition of docking "hits" to one that is more strict and also follows the CAPRI criteria, based on medium or better quality hits. We made two modifications to the above definition to make it suitable for this study. First, we added the additional restriction of f_{nonnat} to ensure that none of the potential hits have excessive steric clash (in CAPRI this problem is solved based on averages of all submitted structures). Second, because here we focus on rigid-body docking, interface RMSD of the best possible hit cannot be smaller than that of the superposed unbound complex. Therefore, we allowed iRMSD to be no more than 2 Å greater than that best hit, thus making it independent of the small variations in the extent of conformational change between different benchmark cases. The resulting definition is described by the following Boolean relationship:

$$\text{Docking hit} = (\text{iRMSD} \leq \text{iRMSD}_{\text{superposed unbound complex}} + 2\ \text{Å}) \wedge (f_{\text{nat}} > 0.3) \wedge (f_{\text{nonnat}} < 0.7) \quad (16)$$

Here, iRMSD is defined as Cα RMSD of those residues having at least one atom within 10 Å of the interacting partner. f_{nat} is the fraction of native contacts defined as the number of native residue-residue contacts in the predicted complex divided by the number of contacts in the target complex. $f_{\text{non-nat}}$ is defined as the number of non-native (incorrect) residue–residue contacts in the predicted complex divided by the total number of contacts in that complex. This latter quantity serves as an indication of atomic clash between the interface residues in the predicted complex.

Statistics

Unless otherwise noted, the calculation of statistical significance for comparisons between potentials was performed with paired-sample Wilcoxon signed rank test.

REFERENCES

1. Smith GR, Sternberg MJE. Prediction of protein-protein interactions by docking methods. Curr Opin Struct Biol 2002;12:28–35.
2. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 2002;47:409–443.
3. 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.
4. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. Proteins 2003;52:80–87.
5. Chen R, Weng Z. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. Proteins 2002;47:281–294.
6. Moont G, Gabb HA, Sternberg MJ. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins 1999;35:364–373.
7. 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.
8. Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci USA 2005;102:10930–10935.

9. Levitt M, Warshel A. Computer simulation of protein folding. Nature 1975;253:694–698.
10. Tanaka S, Scheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 1976;9:945–950.
11. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256:623–644.
12. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
13. Warme PK, Morgan RS. A survey of atomic interactions in 21 proteins. J Mol Biol 1978;118:273–287.
14. Huang ES, Subbiah S, Levitt M. Recognizing native folds by the arrangement of hydrophobic and polar residues. J Mol Biol 1995; 252:709–720.
15. Thomas PD, Dill KA. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257:457–469.
16. Li AJ, Nussinov R. A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. Proteins 1998;32:111–127.
17. Melo F, Feytmans E. Novel knowledge-based mean force potential at atomic level. J Mol Biol 1997;267:207–222.
18. Mitchell JBO, Alex A, Snarey M. SATIS: atom typing from chemical connectivity. J Chem Inf Comput Sci 1999;39:751–757.
19. Mintseris J, Weng Z. Optimizing protein representations with information theory. Genome Inform Ser Workshop Genome Inform 2004;15:160–169.
20. Cline MS, Karplus K, Lathrop RH, Smith TF, Rogers RG, Jr, Haussler D. Information-theoretic dissection of pairwise contact potentials. Proteins 2002;49:7–14.
21. Zhang L, Skolnick J. How do potentials derived from structural databases relate to "true" potentials? Protein Sci 1998;7:112–122.
22. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z. Protein-protein docking benchmark 2.0: an update. Proteins 2005;60:214–216.
23. Desmyter A, Spinelli S, Payan F, Lauwereys M, Wyns L, Muyldermans S, Cambillau C. Three camelid VHH domains in complex with porcine pancreatic α-amylase. Inhibition and versatility of binding topology. J Biol Chem 2002;277:23645–23650.
24. Chen R, Tong W, Mintseris J, Li L, Weng Z. ZDOCK predictions for the CAPRI challenge. Proteins 2003;52:68–73.
25. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. CAPRI: a critical assessment of PRedicted interactions. Proteins 2003;52:2–9.
26. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: an FFT-based protein docking program with pairwise potentials. Proteins 2006;65:392–406.
27. Chen R, Mintseris J, Janin J, Weng Z. A protein-protein docking benchmark. Proteins 2003;52:88–91.
28. Mintseris J, Weng Z. Atomic contact vectors in protein-protein recognition. Proteins 2003;53:629–639.
29. Boyan JA. Learning Evaluation Functions for Global Optimization [PhD thesis]: Carnegie Mellon University; 1998.
30. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
31. Hubbard SJ, Thornton JM. NACCESS. 2.1.1: Department of biochemistry and molecular biology, University College London; 1993.
32. Wang G, Arthur JW, Dunbrack RL. S2C: a database correlating sequence and atomic coordinate numbering in the protein data bank. : http://www.fccc.edu/research/labs/dubrack/s2c; 2002.
33. Mendez R, Leplae R, De Maria L, Wodak SJ. Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins 2003;52:51–67.

 DOI 10.1002/prot