PROTEINS: Structure, Function, and Genetics 39:372–384 (2000)

BiGGER: A New (Soft) Docking Algorithm for Predicting Protein Interactions

P. Nuno Palma,[1,2]* Ludwig Krippahl,[1] John E. Wampler,[3] and José J.G. Moura[1]
[1]*Departamento de Química, Centro Química Fina e Biotecnologia, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal*
[2]*Instituto Superior de Ciências da Saúde, Monte de Caparica, Portugal*
[3]*Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia*

ABSTRACT **A new computationally efficient and automated "soft docking" algorithm is described to assist the prediction of the mode of binding between two proteins, using the three-dimensional structures of the unbound molecules. The method is implemented in a software package called BiGGER (Bimolecular Complex Generation with Global Evaluation and Ranking) and works in two sequential steps: first, the complete 6-dimensional binding spaces of both molecules is systematically searched. A population of candidate protein-protein docked geometries is thus generated and selected on the basis of the geometric complementarity and amino acid pairwise affinities between the two molecular surfaces. Most of the conformational changes observed during protein association are treated in an implicit way and test results are equally satisfactory, regardless of starting from the bound or the unbound forms of known structures of the interacting proteins. In contrast to other methods, the entire molecular surfaces are searched during the simulation, using absolutely no additional information regarding the binding sites. In a second step, an interaction scoring function is used to rank the putative docked structures. The function incorporates interaction terms that are thought to be relevant to the stabilization of protein complexes. These include: geometric complementarity of the surfaces, explicit electrostatic interactions, desolvation energy, and pairwise propensities of the amino acid side chains to contact across the molecular interface. The relative functional contribution of each of these interaction terms to the global scoring function has been empirically adjusted through a neural network optimizer using a learning set of 25 protein-protein complexes of known crystallographic structures. In 22 out of 25 protein-protein complexes tested, near-native docked geometries were found with C^{α} RMS deviations ≤ 4.0 Å from the experimental structures, of which 14 were found within the 20 top ranking solutions. The program works on widely available personal computers and takes 2 to 8 hours of CPU time to run any of the docking tests herein presented. Finally, the value and limitations of the method for the study of macromolecular interactions, not yet revealed by experimental techniques, are discussed. Proteins 2000;39:372–384.**

Key words: protein interactions; protein docking; protein complexes; macromolecular interactions; molecular recognition; molecular surface; algorithms

INTRODUCTION

Experimentally determined docked geometries have been invaluable in expanding our detailed understanding of biochemical processes. And yet the databases of known structures of protein-protein complexes are orders of magnitude smaller than those of experimental information on protein interactions and of structures of individual proteins. Thus, the availability of computational methods, which can reliably predict the way two proteins will most likely interact, generating working models of such molecular complexes, is of unquestionable importance and may provide additional insight into the nature of macromolecular recognition and biochemical mechanisms.

The computational approaches that attempt to find the best matching (either geometric or chemical) between two molecules are generally referred to as molecular "docking" methods. Methods based on a geometric fit are most common and depend on the assumption that the docked proteins are similar in shape to their undocked counterparts. Within the limit of current structural methods, this assumption appears to hold, providing a certain level of "softness" is employed. Softness in this sense is meant to allow for both imprecision in the structures and their adaptive changes to each other.

From a practical point-of-view, the problem of finding an appropriate protein docking solution is generally addressed in two steps: (1) A search over the N dimensional binding and conformational spaces in order to select

Grant sponsor: COST D7, Molecular Recognition, Human Capital and Mobility; Grant number: ERBCHRXCT 940492; Grant sponsor: PRAXIS XXI /Junta Nacional de Investigação Científica e Tecnológica; Grant sponsor: National Institutes of Health, United States; Grant number: GM50736.

*Correspondence to: P. Nuno Palma, Centro de Química Fina e Biotecnologia, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2825-114 Caparica, Portugal. E-mail: palma@dq.fct.unl.pt

Received 27 September 1999; Accepted 20 January 2000

candidate geometries of the complex; (2) Application of a suitable scoring function to distinguish, in a reliable way, near native modes of binding from the other false solutions generated during the initial search. These two different tasks can be summarized in two key words: searching and scoring.

In protein-protein associations, the number of degrees of configurational and conformational freedom can reach immense proportions, making it impracticable to perform a full conformational search with the current computers available. This is the main reason why most protein-protein docking methods treat the globular shape of a protein as a rigid body entity while optimizing the matching of their surfaces.[1–6] Reducing the protein to a non-flexible structure may result in a drastic simplification of the search procedure but limits the application of such methods to situations where very little conformational changes occur to the protein atoms upon molecular association.

In fact, these techniques work generally well when reconstructing the geometry of protein complexes from the co-crystallized structures but may produce unpredictable results when starting from the free component molecules, in particular if their structures differ significantly from that of the complex.

Fortunately, in most of the known protein complexes, the overall structure of the proteins is only very slightly changed, compared to the free forms, with most of the conformational changes confined to the side chain atoms of surface amino acids.[7] So, with slight modifications of the basic rigid body approaches, some techniques may tolerate a limited degree of molecular flexibility by using a "soft" representation of the molecular surface.[1,8–10] Although these methods will improve the chances of success in predicting the structure of protein complexes from those of the unbound components, their general applicability is still to be demonstrated.

The second issue in protein docking is filtering or scoring the putative docked geometries. A complete search procedure may produce an immense number of putative docked geometries, which must then be evaluated according to chemical criteria. Most of the docking algorithms developed so far use the extent of geometric complementarity of the protein surfaces as an initial filter to eliminate a large number of solutions with poor surface matching. As discussed above, the use of such criterion is supported by the observation that most protein complexes exhibit a close geometric match between the molecular surfaces at the binding interface (for example see Hubbard et al.[11]). But the fact that geometric properties are usually easier and faster to compute also gives a practical support for this approach.

It is, however, generally recognized[2] that a criterion based solely on geometric complementarity is far from being sufficient to discriminate between native and non-native docked geometries, except for a very few cases.

The definition of a general form of scoring function, which can reliably distinguish the "true" binding mode from the remaining "false" ones, is a challenging topic of current research. Several such criteria have been implemented with different levels of success, ranging from pure geometric scoring[1,5,6,12–15] to reduced forms of electrostatic interaction potentials[10,16–18] or combinations of the above with subsequent energy refinement.[2,3] Other techniques exploited atomic hydrophobicity[19,20] and Hydrogen bonding[16,21] terms, as well as empirical scoring functions[22] or more complete evaluations of relative free energy of binding.[23]

Regardless the method used for predicting the most favorable interaction mode, none of the individual energetic contributions that have been assessed proves to be sufficient, per se, to distinguish the true solution for all tested complexes. In some cases one single energetic contribution (e.g., electrostatics) is able to select the correct geometry for one complex, but in other cases it leads to false solutions.

One major problem associated with the scoring of candidate-docked solutions is that it must be done exhaustively for a large number of alternative geometries. As a result, the energy functions must be greatly simplified in order to render the computation feasible. The compromise between speed and accuracy thus taken may cause the loss of relevant energy details.

Despite the efforts undertaken by so many authors to develop generic tools for docking proteins, it is fair to say that this goal is still far from being reached. Different complexes may present important differences in the principles of their molecular association[9] and no docking method may be considered generic if not tested against a large number of known protein complexes of different natures.

The present work provides a contribution to this field by proposing a new approach to dock globular proteins. The method tolerates well the conformational flexibility of the amino acid side chains at the protein surface, is exceptionally fast, and uses an empirically optimized scoring function to rank the putative modes of binding. A set of 25 protein-protein complexes with known 3D structure is used to test the algorithm.

MATERIALS AND METHODS

The docking procedure is composed of two modules that work in a sequence. The first module (BoGIE, Boolean Geometric Interaction Evaluation) is a grid-like search algorithm, which seeks to generate a population of docked geometries with maximal surface matching and favorable intermolecular amino acid contacts. In the second module, the putative binding modes thus generated are evaluated according to a set of interaction terms that are thought to be relevant to the stabilization of protein complexes. These terms are finally combined into a global scoring function using weighting factors that have been optimized through a learning process with a neural network algorithm.

Unlike some other techniques,[10,18] the present method searches the complete surface of each molecule, assuming absolutely no additional information regarding the binding sites. Even if this search is performed in discrete steps, the number of degrees of freedom in the 6-dimensional

TABLE I. Protein-Protein Complexes, With Known Structures, Used to Train and Test the Docking Algorithm[†]

Dock name	Description	C^{α} RMS target/probe	pdb files
BOUND			
2SICXX	Subtilisin-inhibitor (wt)	0.00/0.00	2sic
1SBNXX	Subtilisin-eglin c	0.00/0.00	1sbn
1TECXX	Thermitase-eglin c	0.00/0.00	1tec
1ACBXX	α-chymotrypsin-eglin c	0.00/0.00	1acb
3SDHXX	Clam hemoglobin dimer	0.00/0.00	3sdh
2CCPXX	CcP-cytochrome c	0.00/0.00	2pcc
PSEUDO-UNBOUND			
3SDHXF	Clam hemoglobin dimer	0.00/0.76	3sdh
1DXGXF	Desulforedoxin dimer	0.00/0.94	1dxg
6EBXXF	Erabutoxin dimer	0.00/1.08	6ebx
2MIPXF	HIV-2 protease dimer	0.00/1.20	2mip
3HFLXF	HyHel5 Fab-lysozyme	0.00/1.22	1lza, 3hfl
3HFMXF	HyHel10 Fab-lysozyme	0.00/1.25	1lza, 3hfm
1CTAXF	Troponin c dimer	0.00/2.04	1cta
UNBOUND			
2PTCFF	Trypsin-inhibitor	0.63/1.71	2ptn, 4pti, 2ptc
2SICFF	Subtilisin-inhibitor (wt)	0.67/1.27	2stl, 3ssi, 2sic
1BRSFF	Barnase-barstar	0.78/1.62	1a2p, 1a19, 1brs
2PCCFF	CcP-cytochrome c (yeast)	0.85/1.04	1ccp, 1ycc, 2pcc
2PCBFF	CcP-cytochrome c (horse)	0.86/1.31	1ccp, 1hrc, 2pcb
2SNIFF	Subtilisin-chymotrypsin inhibitor	0.87/1.13	1sup, 2ci2, 2sni
1FSSFF	Acetylcholinesterase-fasciculin II	0.91/1.41	2ace, 1fsc, 1fss
1CHOFF	α-chymotrypsin-ovomucoid	0.98/1.52	5cha, 2ovo, 1cho
1MLCFF	D44.1 Fv-lysozyme	1.00/1.28	1mlb, 1lza, 1mlc
1FDLFF	D1.3 Fab-lysozyme	1.08/1.20	1vfa, 1lza, 1fdl
2KAIFF	Kallikrein-trypsin inhibitor	1.21/1.43	2pka, 1bpi, 2kai
1CGIFF	α-chymotrypsinogen-trypsin inhibitor	1.68/2.54	1chg, 1hpt, 1cgi

[†]*Bound, pseudo-unbound,* and *unbound* groups designate, respectively, complexes whose prediction is attempted from the structures of both co-crystallized proteins, the structure of one of the complexed proteins and the free form of the other and from the free form of both proteins. The RMS difference (α carbons only) between the atomic coordinates of the individual proteins (target and probe) in both the free and complexed forms are indicates where pertinent.

binding space has, so far, restricted the use of such real space methods to high-performance machines. The method reported in this work relies on very fast and heuristically optimized Boolean type operations (OR, AND, etc.), which makes it feasible to be implemented on widely available personal computers and to take advantage of local PC networks for parallel processing.

Test Cases and Files

A collection of 25 protein-protein complexes with experimentally (X-ray and NMR) determined structures was used as learning and test sets (Table I). These were chosen from complexes of different types, including several examples of dimeric proteins, protease-(proteic) inhibitor, antibody-antige, and complexes between electron transfer proteins. Furthermore, in order to test the ability of the program to handle the conformational changes that occur upon formation of the complexes, two quality tests were performed: in the first 6 cases the complexes were reconstructed from the structures of the co-crystallized proteins, after being taken apart from the complex and randomly orientated. In these cases, the conformations of the two molecules are already "adapted" to each other and this set of docking simulations is designated as *bound* in Table I. In all other cases, the complexes were predicted using the unbound or native conformations of both interacting proteins (*unbound*) or at least, of one of the two (*pseudo-unbound*). This last category includes all five dimers (monomer one was docked to a copy of itself instead of the other monomer) and the two antibody-antigen complexes 3hfl and 3hfm, in which, the Fab fragments were extracted from the structure of the complexes and docked to the free lysozyme.

Searching and Filtering

The program starts by reading in the atomic coordinates of the two proteins, which are to be docked, in standard PDB format. Hydrogens are added according to the structural information of the isolated residues, provided in the molecular mechanics force field AMBER.[24,25]

Digitizing molecular shape

The first step in BoGIE is the generation of a 3D matrix (volume matrix) composed of small cubic cells of 1Å size, which represents the complex shape of each molecule. Each matrix cell will have a value of 1 (True) if its center

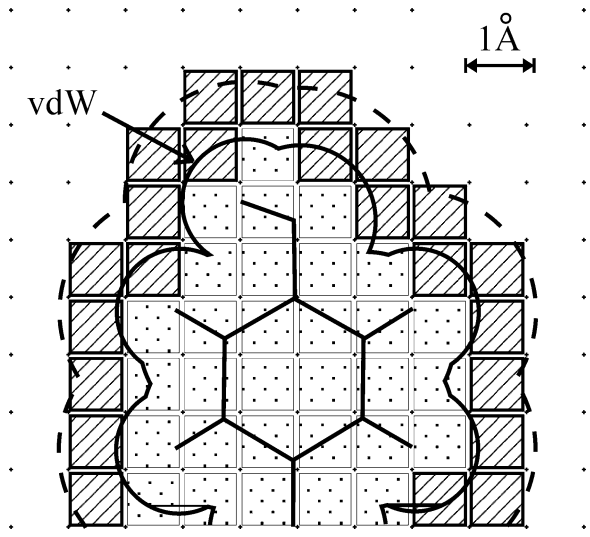

Fig. 1. Schematic representation of a digitized molecular fragment. The side chain of a planar tyrosine residue is represented by its van der Waals surface contour (solid). The molecular shape is placed into a regular Cartesian grid of 1Å unit size and every grid position is assigned to the molecular volume (will have a value of *1*) if its center lies within 1Å (dashed contour) of the van der Waals surface. All other cells represent the exterior of the molecular volume and are assign a value of *0* (False). The outer cubes of this representation (filled with slanted lines) will be subsequently assigned as surface shell and the remaining inner positions will constitute the molecular core (dot-filled cubes).

lies within 1Å of the van der Waals sphere of any protein atom (Fig. 1). All other cells represent the exterior of the molecular volume and are assigned a value of 0 (False). Note: this digitization introduces the first level of "softness" in the algorithm.

Defining Molecular Surface

In this step, two new Boolean matrices are generated for each docking partner. One—the surface matrix—containing a definition of the molecular surface (a hollow shell with the protein shape) and the other—the core matrix—representing the positions belonging to the inner core of the protein.

More precisely, the surface region is defined as the set of volume matrix cells that are occupied by the protein shape (i.e., have a value 1) and have at least one neighbor cell that is empty (i.e., has a value of 0). Computationally, the process of defining the molecular surface is very simple. A copy of the volume matrix is shifted one location in each of the 26 possible directions at a time, and an *Exclusive Or* (XOR) logical operation is performed between the shifted and the original matrices. Since the result of a XOR operation is True only if both operands are different, this will result in a double surface shell that can be trimmed with an And operation with the original volume matrix (Fig. 2). At this point, each protein is represented by two 3D matrices: the surface matrix and the core matrix. This type of representation is very economical, since one unique bit (0 or 1) is needed to define one surface location and one byte may contain 8 surface locations at once. The matrices are also encoded in a compact form, taking advantage of the homogeneity of large regions in the matrices. This encoding improves the performance of the search heuristics and further reduces memory requirements.

Evaluating Surface Matching

For every relative orientation of the two proteins, the translational interaction space is searched by systematically shifting the matrices defining one molecule (the Probe) relative to the matrices representing the other partner (the Target). This movement is performed in discrete steps of the size of the matrix cells. The extent of surface matching between the two molecules is thus simply evaluated by the number of surface cells of the target molecule overlapping any surface cells of the probe. Computationally, this process is made very fast by applying the AND Boolean operator to the memory registries containing the two surface matrices. A value of *1* is obtained for every position simultaneously occupied by the surfaces of both proteins and *0* elsewhere. At the same time, every solution that results in overlapping core cells of both molecules is immediately discarded, thus rejecting solutions involving unrealistic interpenetration of the docking partners. However, core-surface overlaps are allowed and this represents the second level of "softness" in the algorithm.

Finally, the probe is rotated a certain amount (typically 15°) relative to the target and this process (digitization, translation, surface matching) is repeated until a complete non-redundant search in the 6-dimensional space is performed. It should be pointed that other finer angular steps have been tested, but this did not result in any significant improvement of the results. Since the size of the search angular step greatly influences the number of orientations that have to be assessed and so the computational time, we used 15° in this work.

Treating Molecular Flexibility

Conformational changes that occur during the formation of a protein complex are among the most difficult challenges to rigid body docking methods. While the above algorithm implements some soft character in the digitization of the shapes and the allowed overlap of cores and surfaces, tests suggested more was needed to properly treat molecular flexibility.

Treating this molecular flexibility in an explicit way would be an impracticable computational task. However, the technique used in this work incorporates side chain flexibility, implicitly, in the definition of a soft molecular surface. This is based on the observation that most of the conformational changes occurring upon complex formation are due to flexible amino acid side chains positioned at the molecular surface (Fig. 3). Moreover, not every amino acid shows the same degree of freedom to move. Amongst the protein complexes observed, ARG, LYS, ASP, GLU, and MET present the highest frequency and amplitude of movements between the structures of free and co-crystallized proteins. So, every atom (except carbon β) belonging to the side chains of these amino acids is considered flexible and is allowed to unrealistically penetrate the

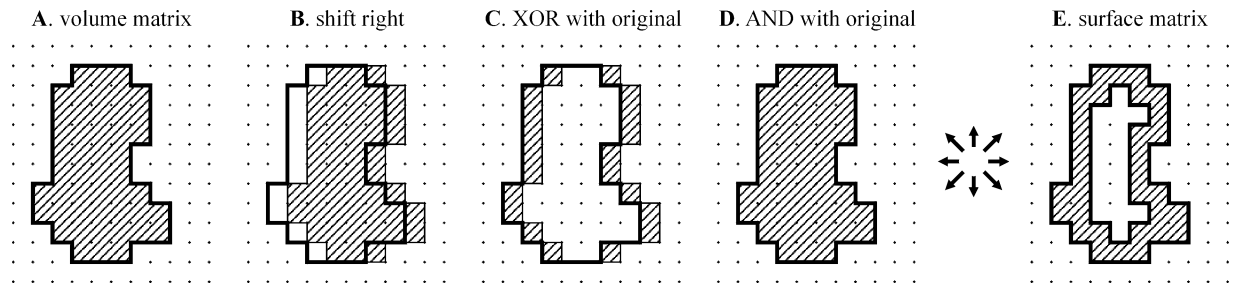

Fig. 2. Schematic representation (two-dimensional) of the surface determination process. **A:** The volume matrix occupied by the protein is outlined with a thick black line. **B:** A copy of the matrix (gray) is shifted one location relative to the original (*heavy contour*). **C:** XOR operation *between both matrices* eliminates (turns to *0*) all *overlapping* locations. Two partial shells are left (*gray cells*), one belonging to the original and the other belonging to the copy (outside the original contour). **D:** The outer shell fragment is eliminated by an AND operation with the original *matrix* (heavy contour). **E:** The previous steps are repeated in all directions (*including diagonals*) and the partial surface shells are *finally* added with OR operations. The resulting surface shell is composed solely of surface locations belonging to the original shape (gray).

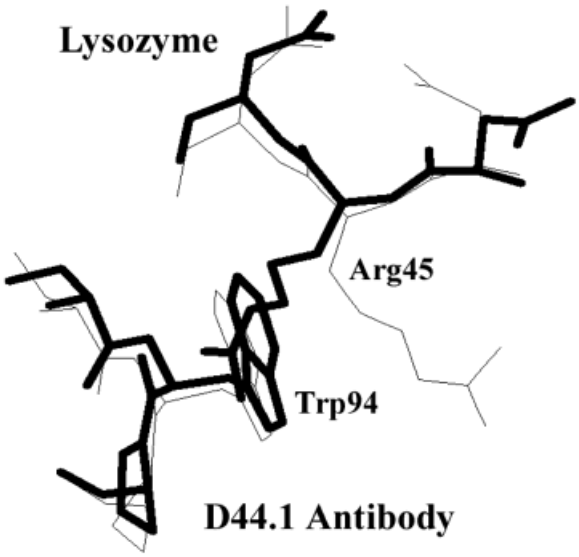

Fig. 3. Detail of the interaction between monoclonal antibody D44.1 Fab fragment (1mlb) and lysozyme (1lza). Thin lines correspond to the X-ray structure of the complex, while thick lines show the conformations of the non-complexed antibody and lysozyme structures, when superposed on that of the crystallographic complex (1mlc). A large conformational difference of Arg45 side chain atoms is revealed, while the relative positions of the main chain atoms are mostly preserved. A major clash between Trp94 of the antibody and Arg45 of lysozyme is expected, if the free structures of the two proteins are docked to each other as rigid bodies.

other molecule during the docking search. The results presented throughout this paper consider soft surface representation for only those five amino acids. However, the list of atoms to treat in this way may easily be edited and extended to other mobile groups, such as carbohydrate adducts found at the surface of some proteins.

This "soft" representation is achieved by simply setting to *0* the core matrix cells assigned to the side chain atoms of those five amino acid types. Having no core cells, such "hollow" side chains can penetrate up to the core of the other protein, without causing forbidden core-core overlaps (Fig. 4). Although not treating conformational flexibility in an explicit way, this approach avoids near-native docked geometries being discarded due to a mispositioned side chain. This "trimming" operation is performed only once, after the generation of the surface and core matrices and no additional information needs to be stored or processed during the search phase.

Filtering

During the search procedure, up to 10^9 different modes of contact between two proteins may be assessed (depending on the molecular size). It is absolutely necessary to drastically reduce this number of putative solutions before they can be evaluated according to any energy terms. A sub-population of 1,000 binding modes is actually kept, which represents a reduction of more than 99.999% of the total solutions sampled.

During this work, two levels of filtering were tested to discard unlikely solutions. The simplest one uses the unique criterion of geometric complementarity. After each evaluation of surface matching, its value is compared with a sorted lookup table containing those of the 1,000 best matching solutions found so far. If its surface matching is poorer than that of the worst solution in the table, then it is discarded. Otherwise, it is saved and the geometry of the worst element in the table is eliminated.

As discussed above, pure geometric surface complementarity may not be sufficient as the sole criterion to safely eliminate unlike binding modes. Within our test cases, we found many situations where the native-like complex has a poor intermolecular surface contact compared to many alternative incorrect solutions. This can cause the native-like solution to be pushed down the table of top scores and eventually, in some cases, to be excluded from the table of retained solutions.

Therefore, a more complete and combined criterion is used, where indicated, to filter out unlike docked geometries. In this procedure, the list of retained solutions is still sorted by surface matching score and every solution with a lower index of surface matching, than the last element in the list, is immediately discarded, as above. However, the remaining complexes are not automatically kept. They are checked for pairwise amino acid contacts across the molecular surface (see Side Chains Interactions below) and only those possessing favorable net amino acid

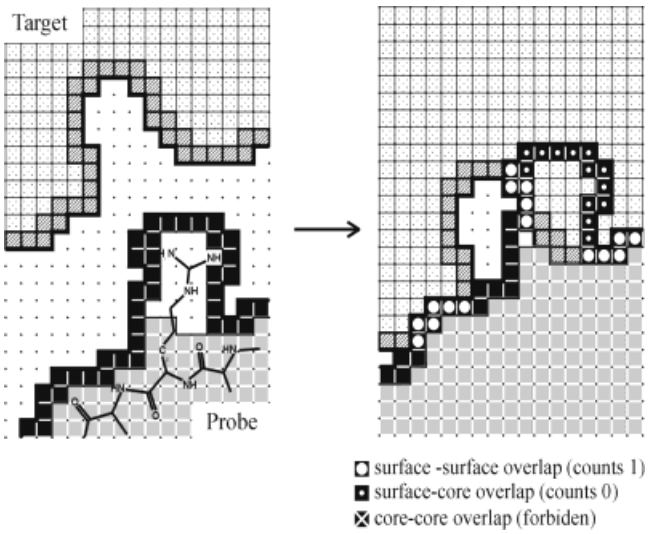

Fig. 4. Schematic representation of the surface matching evaluation. Flexible amino acid side chains are assigned a "soft" shape by eliminating (setting to *0*) every core positions belonging to side chain atoms other than carbon β. This produces a hollow surface shell with the same shape. For every relative position of the two interacting proteins (actually, their digital representations), the extent of their surface matching is evaluated by the number of surface cells of the target molecule overlapping surface cells of the probe. Although allowed to occur, overlaps between surface and core cells do not contribute (they count *0*) for the matching evaluation. Finally, every solution that results in the overlapping of core cells from both molecules is immediately discarded, thus rejecting solutions involving unrealistic interpenetration of the docking partners.

interactions are then inserted in the list of saved solutions and the last element is discarded.

Since this operation is performed for the very small percentage of eligible docked geometries with higher surface matching scores, it does not significantly slow the search and filtering process.

Side Chains Contacts

This is a statistically based interaction potential that takes into account the relative propensity of each pair of amino acids to be in close contact to each other. It's a purely empirical interaction term derived from the observed statistical frequency of naturally occurring contacts between pairs of amino acids in a database of well resolved X-ray protein structures.[26]

The underlying assumption in our implementation of this term is that the molecular packing and the nature of the intermolecular chemical interactions at the interface of two complexed proteins should resemble those observed inside a globular protein structure.

In the current implementation, the probability P_{AB} that amino acid A (or any side chain atom) being in contact with amino acid B, is defined by

$$P_{AB} = \left(\frac{N_{AB}}{N_{AX}} - \frac{N_B}{N_X}\right) \times 100 \qquad (1)$$

where N_{AB} is the statistically observed number of contacts between amino acids A and B, and N_{AX} is the total number of contacts involving amino acid A and any amino acid X. N_B and N_X are, respectively, the number of natural occurrences of amino acid B and the total number of amino acids in the database analyzed.

Hence, only if the percentage of contacts between A and B is higher than the natural frequency of occurrence of B itself, should this interaction represent an actual statistical preference and P_{AB} will be positive.

For example, ALA-ASN interactions represent 5.13% of all interactions involving Alanine. Since this percentage is higher than the natural occurrence of ASN (4.4%), this reveals a slight, but selective preference of ALA for ASN, and this will count as a favorable contact with a score $P_{ALA\text{-}ASN}$.

These values are stored in a lookup table and each time amino acids A and B (belonging to different molecules) are found in contact across the molecular interface, P_{AB} and P_{BA} need just to be summed (note that AB contact may have a different propensity than a BA contact).

The program uses a simple and fast algorithm for evaluating contacts between the side chains of amino acid across the interaction surface. Although the two molecules are docked as rigid bodies, the definition of side chain contacts between proteins takes into account their conformational flexibility in an approximate way. The volume accessible to a particular side chain is taken as a sphere containing all side chain atoms and centered at the geometric center of that group of atoms. Thus, two amino acids belonging to different proteins are considered in contact if the spheres defining their accessible volumes touch or intercept.

Evaluating Interactions

At this stage, the search and filtering algorithm has generated and stored 1,000 putative docked geometries with extended surface matching and overall favorable amino acid contacts across the molecular interface. These two criteria were used as initial filters because their computation can be very fast and performed over the complete set of solutions sampled.

The next module (INTERACT) will evaluate the population of putative solutions according to four different interaction terms pertinent to the process of molecular association (surface matching, side chain contacts, electrostatics, and solvation energy). Finally, they will be properly combined into a global scoring function computed for every docked geometry.

Surface Matching

Geometric complementarity, itself, does not constitute any type of direct driving force for molecular association. Nevertheless, one might think that the more extensive the surface contacts, the greater the probability of other interaction terms, such as van der Waals, hydrogen bonds, and electrostatic interactions, having an effect on the overall stabilization of the complex. So, the values computed and used above to filter the population of putative binding modes will be used now as one of the quantitative scoring criteria.

Side chains contacts

The second interaction term, also described above, evaluates, for each docked geometry, the likeliness of occurrence of every observed contact between pairs of amino acids across the molecular interface. This evaluation was already used in the filtering phase but only to distinguish between favorable and unfavorable interactions. It will now contribute to a more fine quantitative scoring function.

Electrostatics

Although charge-charge interactions play an obviously important role in macromolecular interactions, its accurate evaluation represents a serious challenge. However, if electrostatics are to be evaluated for a large number of different docked geometries, it is imperative to keep the model as simple as possible.

In the present implementation, the program uses a modified Coulombic potential function. The atomic point charges are imported from the molecular mechanics force field Amber 4.1[24,25] and a distance-dependent dielectric function of the type $\epsilon(r_{ij})=r_{ij}$ is used, where r_{ij} is the distance between point-charges i and j.

There is, however, one difficulty associated with the straight use of a point-to-point Coulombic potential. Since conformational flexibility is modeled implicitly, by allowing a limited interpenetration of the grid positions of both molecules, some atoms at the molecular interface may become unrealistically close to each other giving rise to very high interaction energies. Thus, when evaluating the effective electrostatic term between a pair of atoms across the docking interface, a dampening constant c is added to the distance separating both nuclei:

$$V_{elec} = k \cdot \frac{q_i q_j}{(r_{ij} + c)^2} \tag{2}$$

k being a constant including the electric permissivity of *vacuum*. The constant c is set as the minimum distance allowed between two interacting atoms and a value of 1.5 Å is used through this work. A similar "buffered electrostatic potential" has been recently used in the new Merck molecular mechanics force field, MMFF94.[27]

Relative solvation energy

The stabilization of a protein complex relative to the dissociated forms of its components results from the competition between protein-protein and protein-solvent interactions.

Wang and coworkers[28,29] suggested that the relative free energy of solvation of a globular protein be estimated by a simple function of the form

$$\Delta G_{solv} = \sum_i \sigma_i A_i \tag{3}$$

where A_i is the solvent-accessible area of a predefined molecular fragment i and δ_i is a solvation parameter reflecting the relative propensity of its atoms to be solvated in water. In their case, the δ_i values were derived empirically as scoring parameters for evaluating native folds of individual proteins.

For the purpose of comparing and scoring a set of alternative docked geometries between the same two molecules, one simply needs to compute the partial desolvation energy associated with the portion of the molecular surface buried inside the complex interface. This will be designated as the solvent excluded area (SEA). The same set of original molecular fragments and solvation parameters[29] were used but, for reasons of performance, a simplified and faster algorithm is implemented to estimate the atomic solvent excluded areas.

Clustering

Due to the systematic nature of the grid searching process, many of the docked geometries thus generated are similar to each other, differing by just one or two translation or rotation steps. For the sake of identifying the binding sites in the two proteins, such similar solutions may be clustered together and treated as one unique binding mode. Again, this approach fits into the theme of this work to incorporate "softness" into the method. Thus, a cluster of solutions is not sought so much as a group of similar solutions but reveal, in some aspects, the dynamic and variable nature of the docking.

As for every individual and dissimilar solution, each cluster is evaluated by the same four interaction terms described above. But in this case, each term is the average of the corresponding values for every solution in the cluster. Averaging these values has the advantage of reducing the associated errors, improving the classification.

The clustering process used is a simple but effective one, in which any given solution is considered as belonging to a given cluster if it is geometrically similar to, at least, one element of that cluster. This simple clustering method could in theory lead to very "elongated" clusters with very dissimilar solutions, but in practice this was not observed. Two docking solutions are considered similar if the RMS deviation of the α-carbons of the probe molecules is less than 2.5Å. Note that the RMS values are computed only for the probes with the target molecules superimposed. If both the target and the probe molecules were considered, the actual RMS deviations would be even smaller (roughly half).

Scoring

At this stage, the four individual interaction terms evaluated for each docking solution (either individual or cluster) must be properly combined into a *global scoring function*. This will be the criterion used to rank docking results and to distinguish the near-native solutions—which resembles the experimentally determined structure of the complex—from the majority of other incorrect solutions.

None of the interaction terms is sufficient per se to be used as a judging criterion. Besides, the theoretical backgrounds behind each of the interaction terms are quite disparate and they are expressed in different units and

scales, so they cannot be directly compared. However, the main assumption behind the scoring function herein described is that a proper combination of the four independently calculated interaction terms should include, in one way or another, most of the relevant details determining the molecular recognition and specificity of non-covalent protein complexes.

The scoring function is empirically defined through a learning process using neural network technology and aiming at predicting the known crystallographic structures of a series of protein complexes.

The complete set of 25 test cases was divided into two subsets. A set of 14 protein complexes (identified with an asterisk [*] symbol in Table II) was used to train the neural network and the remaining ones were subsequently used to test the scoring function optimized during the initial learning phase.

The classification system used is a feed forward neural network with three hidden layers (with 4, 3, and 2 neurons, respectively) and a single output neuron. The input values for the network are the four interaction terms calculated for each solution (or cluster of solutions) and its output is a single scoring value, which is an estimate of the likelihood of that solution representing a near-correct binding mode. In the process of training and testing, each docked geometry was labeled as a near-correct solution if its RMS deviation (α-carbons) from the crystallographic complex was less than 4.0 Å. Otherwise it was labeled as incorrect.

The training phase used a neural network back propagation algorithm[30] and was targeted at maximizing the distinction between near-correct and incorrect structures.

RESULTS AND DISCUSSION

The process of protein docking presents, as mentioned before, two major difficulties: one is the enormous search space that must be assessed, considering not only the relative orientations of the molecules, but also their conformational freedom. The second issue is the adequate scoring of alternative docked geometries.

Treatment of Conformational Flexibility

Table II summarizes the results of all simulations. The first three columns (A, B, and C) show the number of near-correct docked geometries retained by the program after the initial search and filtering phase (a total of 1,000 solutions are retained during each simulation, for subsequent analysis). A given docking solution is considered functionally similar to the crystallographic structure of the complex if the RMS deviation of the main chain atoms is less than 4.0 Å.

The results shown in the first column (Table II, A) were obtained by docking completely rigid representations of the interacting proteins. Within this representation, the amino acid side chains are not allowed for any flexibility (neither explicitly nor implicitly) and the docked geometries are selected solely on the basis of the geometric matching of their surfaces. This approach is equivalent to the classical rigid-body docking methods developed by other authors.[1,2,5,12–15]

TABLE II. Partial and Global Docking Results[†]

	Searching and filtering			Scoring	
Dock name	A	B	C	Rank	RMS
BOUND					
2SICXX	2	3	18	2	3.76
1SBNXX	51	24	4	31	2.23
1TECXX*	8	—	3	77	3.57
1ACBXX*	12	—	5	18	0.61
3SDHXX*	11	23	35	1	3.17
2CCPXX*	4	—	3	1	3.32
PSEUDO-UNBOUND					
3SDHXF	20	6	52	12	2.33
1DXGXF*	81	46	60	2	0.58
6EBXXF	—	—	2	200	3.54
2MIPXF	23	22	22	82	2.21
3HFLXF*	2	1	4	43	3.70
3HFMXF	—	—	11	13	3.89
1CTAXF*	33	15	14	1	1.36
UNBOUND					
2PTCFF*	1	1	4	52	2.73
2SICFF*	—	8	2	15	3.33
1BRSFF*	10	8	28	35	1.89
2PCCFF	—	—	11	50	3.87
2PCBFF*	—	—	2	18	2.36
2SNIFF*	—	3	7	16	1.32
1FSSFF*	—	—	4	11	3.20
1CHOFF	19	12	43	6	2.93
1MLCFF	—	—	—	—	—
1FDLFF	—	—	—	—	—
2KAIFF	—	—	—	—	—
1CGIFF*	7	1	7	9	3.72

[†]*Searching and filtering* displays the number of near-correct docked structures generated and retained, according to different surface definition and filtering criteria (see text for details); A: 'hard surface' and simple surface matching filter; B: 'soft surface' and simple surface matching filter; C: 'soft surface' and combined surface matching and side chain contacts filter. *Scoring* presents the highest ranking position of a near-correct (RMS ≤ 4.0 Å) docked structure and the corresponding RMS deviation from the actual experimental structure. Dock names followed by an asterisk indicate the set of complexes used to train the neural network and thus, to optimize the scoring function. Dashes indicate that no solutions were retained with RMS ≤ 4.0 Å.

It is clear from these results that only a very small percentage of the best matching docking solutions are indeed similar to the crystallographic complexes. In addition, while this methodology works reasonably well when reconstructing the geometry of protein complexes from the structures of the co-crystallized proteins, it starts breaking down in docking simulations where the unbound conformations of the proteins are used. In fact, when considering the *unbound* and *pseudo-unbound* test cases, no near-native solutions were retained (within the set of geometries saved) for about 50% of the docking simulations.

This can be explained by one of the following reasons. As discussed before, simple geometric complementarity between two macromolecules cannot account for the specificity of most complexes. In some of the test cases, the program could generate and retain at least 1,000 non-native docked geometries with tighter surface matching

than the crystallographic complexes themselves. But the most important limitation of this approach is the inability of such rigid body docking methods to handle the structural rearrangements taking place during the process of molecular association. Table I indicates the RMS (all heavy atoms) differences between the starting structures (target and probe) used in the docking simulations and those of the co-crystallized proteins. As these differences grow, the chances of loosing a near-correct solution also increase. Hard-surface rigid body docking should thereby be considered an unreliable general approach for the real-life situations where only the unbound forms of the protein structures are available.

In order to overcome this difficulty without having to consider explicit conformation flexibility, we introduce the concept of soft-surface rigid-body docking, which will be designated by soft-docking hereafter. The definition of the protein's soft-surface representation is described in Materials and Methods, and Table II (column B) presents the searching results for the complete set of docking tests, using the soft-surface representation and a simple geometric filtering as in column A.

The results presented in Table II reveal no apparent gain in using this soft representation of the protein surface. In a couple of situations, some near-correct solutions could be found in cases where hard docking failed before (dock cases 2SICFF and 2SNIFF). But, in contrast, it failed in other simulations where hard docking had produced near-correct solutions (dock cases 1TECXX, 1ACBXX, and 2CCPXX).

Interestingly, the improvements are observed for *unbound* docking cases and the drawbacks occurred all in the *co-crystallized* dockings. In these last cases, the protein structures of the target and probe already reflect the self-adaptation that occurs during complex formation and no improvements should be expected, in fact, from using a soft surface representation.

In contrast, softening the surface representation produces a "fuzzier," less detailed molecular shape. Although this procedure might be essential to capture near-correct solutions in *unbound* docking simulations, it also has the effect of reducing the difference in geometric complementarity between correct and incorrect solutions, thus increasing the error of the method.

The only way to avoid this source of unpredictability is to improve the quality of the criteria used for selecting which structures are to be retained and which should be discarded during the searching and filtering phase. This is a sensitive point since any complication introduced in this filtering criterion may have a dramatic effect on the computational time of this phase. The filter will be applied to an enormous number of docked geometries during the search phase and so it must be very simple and fast to calculate.

The combined selection filter described in Materials and Methods substantially improves the efficiency of the search and filtering process (Table II, column C) without a significant increase in the computation time. For 22 of the 25 test cases, one or more near-correct solutions were identified and retained by this procedure. Besides, this approach also improves the results of docking simulations with unbound proteins without deteriorating the results obtained for the co-crystallized molecules. In view of the total number of configurations generated during the search procedure, the stored population of 1,000 putative docked structures represents a reduction of the searched binding space by a factor of the order of 10^4–10^7.

It is, however, informative to understand why no near-correct structures where retained for the three failed docking cases. Two of these are complexes between antibody fragments and lysozyme (1MLCFF and 1FDLFF) and the third is the complex between the protease kallikrein and trypsin inhibitor (2KAIFF).

When comparing the structures of the monoclonal antibody D44.1 Fab fragment (1mlb) and lysozyme (1lza) with that of the complex between both (1mlc) a significant conformational change is observed on lysozyme, with Pro70 moving as much as 4 Å and dragging the main chain atoms (Fig. 5). This type of change, although possible, is not common and Proline was not included in the list of amino acids to be treated as flexible. As a result, when docking 1mlb to 1lza, the docked geometry that corresponds to the structure of the crystallographic complex imposes a drastic overlap between Pro70 of lysozyme and Tyr50 of the antibody, forcing it to be discarded.

For the other two unsuccessful test cases, 1FDLFF and 2KAIFF, there were apparently no particular steric conflicts that could explain the fact the no near-correct docked structures were retained. We believe that in these cases, the binding space of the interacting proteins offers too many alternative but incorrect solutions with high surface matching and favorable amino acid contacts. Consequently, any near-correct structures that may have been assessed were pushed down the ranking list of selected solutions and finally discarded.

Scoring Putative Complexes

The main goal of a scoring function in the present context is to rank the population of putative docked geometries initially generated, helping the user distinguish between the near-correct structures (or clusters of structures) and the vast envelope of non-native binding modes.

Ideally, such scoring function should rank the near-correct structures at the top, while assigning a lower score and ranking position to every incorrect solution. Besides, the function should be as general as possible in order to be used on a variety of complexes and not only the ones used for its optimization.

Table II (two rightmost columns) presents the ranking position (after scoring) of the first near-correct structure for each of the 25 complexes assessed, followed by the corresponding RMS deviation from the crystallographic complex.

The scoring function has been optimized with a neural network algorithm using a learning or training set comprising the 14 docking cases marked with an asterisk (*). All other complexes were used to test the generality of the

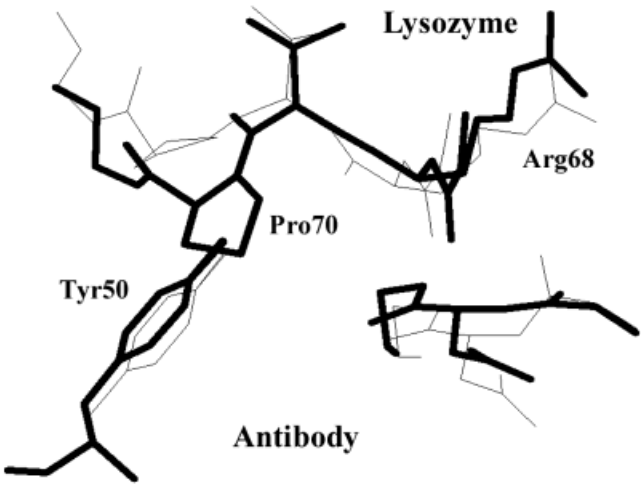

Fig. 5. Detail of the interaction between monoclonal antibody D44.1 Fab fragment (1mlb) and lysozyme (1lza). Thin lines correspond to the X-ray structure of the complex, while thick lines show the conformations of the non-complexed antibody and lysozyme structures, when superposed on that of the crystallographic complex (1mlc). A significant conformational change of Pro70 is shown, including vicinal main chain atoms, as the structure of the free lysozyme forms a complex with the antibody.

evaluation function thereby defined. Note that the three unsuccessful dockings discussed above were naturally excluded from both the training and the test sets.

The average ranking positions of the first near-correct structures for the training and test sets are 21 and 28 (excluding the outlying case of 6EBXXF), respectively. The similarity of the response to both sets of docking cases suggests that the scoring function herein proposed may be applicable to the prediction of other unknown protein complexes. The second requisite for an ideal scoring function, described above, is thus fulfilled.

Figure 6 shows a comparison of the experimentally determined structures of four protein complexes and the best-ranked near-correct predictions reported in Table II. Although the RMS deviation between predicted and X-ray structures can be as high as 3.7 Å, it is clear from these examples that the binding sites are satisfactorily identified, even at this resolution. It must be noted, however, that near-correct structures do not always correspond to the highest scoring solutions and, often, incorrect docked geometries are ranked first.

The first requisite for the ideal scoring function i, therefore, less well accomplished. Even though a near-correct structure is ranked within the top 20 solutions for more than 50% of the complexes predicted, in only three cases does it correspond to a top-ranking solution. At worse, the user may still have to use additional information to select one solution out of a few tens of alternative structures. This might seem disappointing at first, but the fact that a near-correct solution is ranked within the top few tens out of billions of possible binding modes generated is definitely indicative that the algorithm is capturing most of the significant features of the chemical interactions involved in these complexes.

In practical situations, the analysis of the docking results may be simplified if the overall solution pattern is plotted together. One useful way of presenting the results of a docking simulation is shown in Figure 7, where one of the interacting proteins (target) is represented surrounded by small spheres placed at the geometric center of the second molecule (probe), for each of the alternative docked positions generated.

It can be seen that there are two main possible interaction sites for barstar at the surface of barnase. In addition, if the putative solutions are shaded according the corresponding interaction score, the highest scoring solutions (darker spheres) tend to cluster within site A, which corresponds to the crystallographic structure of the complex (the correct position of the inhibitor barstar is shown in light gray). In combination with a second plot with the probe molecule surrounded by spheres showing the relative positions of the target, this type of representation proves to be very helpful to indicate the preferred docking sites in both molecules.

It must be pointed out, however, that the methodology herein described should not be used as a definite technique to pinpoint the native-like structure of an unknown complex. But, when properly combined with experimental information on the target complex, the method proves to work as an invaluable tool to help elucidating the molecular aspects of protein associations. BiGGER has been extensively applied to the study and characterization of several molecular complexes formed between electron transfer partners, including complexes between monohemic *c*-type cytochromes and cytochrome *c* peroxidase[31,32] or between cytochrome c553 and ferredoxin[33] and [Fe]-Hydrogenase.[34] In these studies, partial structural information is obtained from NMR, cross-linking, kinetics and other biochemical experiments and is used to cross-validate the docking results and to help selecting self-consistent docked structures.

Performance

The searching method herein described is based on a real space grid-searching algorithm, in contrast to other docking methods proposed in the literature, based on the Katchalski-Katzir[5] fast Fourier Transform (FFT) algorithm.

The weaker dependence of FFT methods on the size of the space grid, $O(N^3 \times \log_2 N^3)$, makes them more attractive than simple real space matrix methods, which are $O(N^6)$, N being the number of grid nodes. However, the real space search can easily be guided by effective heuristic rules, thus drastically reducing the search space and cutting-off computational time. In the present implementation, the quick identification and ruling out of forbidden (e.g., core-core overlaps) or unfruitful interactions (e.g., where the contact value cannot be greater than the smallest value in the list of retained solutions) results in an algorithm which is only $O(N^{2.8})$ relative to the size of the matrix (estimated from a sample of 13 docking runs, using cubic grids ranging from 24 to 76 grid nodes).

An additional advantage of the method herein proposed is the smaller size of the matrices required, as compared to

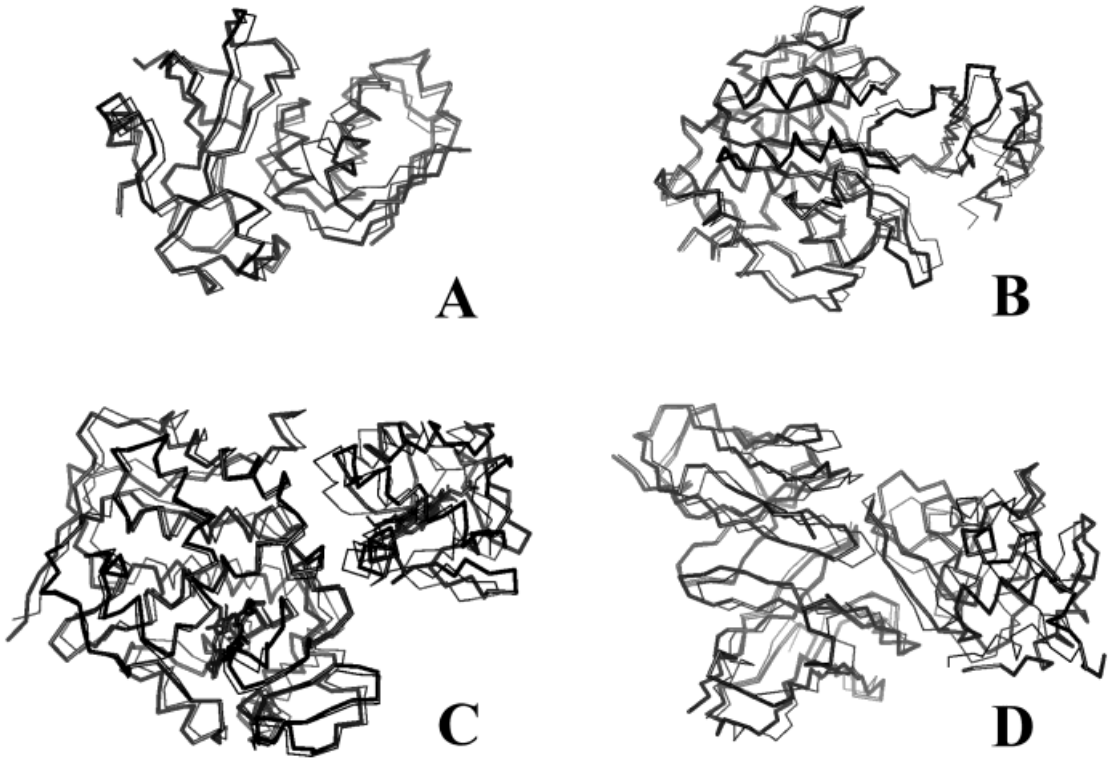

Fig. 6. Superposition of the experimentally determined structures of four protein complexes and the best ranked near-correct predictions reported in Table II. Thick lines: C^{α} trace of experimental structure. Thin lines: C^{α} trace of predicted model. **A:** Dock case 1BRSFF, solution no. 35, C^{α} RMS = 1.89 Å; **B:** dock case 2SNIFF, solution no. 16, C^{α} RMS = 1.32 Å; **C:** dock case 2PCBFF, solution no. 18, C^{α} RMS = 2.36 Å; **D:** dock case 3HFLXF, solution no. 43, C^{α} RMS = 3.70 Å.

FFT algorithms. These can be cut down to fit the width of each molecule in the former method, while a cubic matrix, at least, equal to the sum of the widths of both proteins is required for FFT methods. Furthermore, to be truly effective, FFT requires that the matrix width be a power of 2. As a consequence, the width of a matrix (N) for a real space search is approximately half of that needed for an FFT docking.

Finally, floating point matrices must be used in the FFT algorithm, requiring four-to-eight bytes (depending on the precision used) to describe each cell. The real space method proposed requires about two orders of magnitude less memory of storing space (less than 300 kb for the examples shown here, compared to around 100Mb for FFT with matrices of similar size), thus making it practical to dock even large proteins in virtually any personal computer.

The computation time is also lower than that reported FFT implementations. Gabb et al.[17] reported typical CPU times on the order of 6 hours for a typical FFT docking simulation (on protein complexes similar to the ones discussed in the present work) on an SGI Power Challenge symmetric-array multiprocessor using eight R10000 processors simultaneously. The set of docking simulations described in the present work were run on a single processor desktop PC (Intel Pentium II 450 MHz) and the total CPU times consumed for each simulation ranged from 2 to 8 hours, depending on the size and shape of the proteins docked.

The algorithm is developed for PC platforms, thus taking advantage of the increasing CPU power of PCs and of the widespread use of these inexpensive machines. In addition, BiGGER allows the work load to be easily split among any number of networked machines, which can further reduce the total computation time to a fraction of those values.

CONCLUDING REMARKS

The results presented clearly demonstrate that protein docking methods that are based on rigid body representations of the interacting molecules may easily produce wrong predictions. This is especially true when the molecular structures acquire different conformations as they are in the free (unbound) state or complexed to the other protein. Conformational flexibility, at least at the level of the amino acid side chains, must be considered, in a way or another, in macromolecular docking methods. One simple way of doing so is by increasing the surface thickness, as proposed by Strynadka et al.,[35] for example. However, we did not get satisfactory results with this method (results not shown), as it promotes indiscriminate molecular overlapping and reduces the signal-to-noise ratio. The approach herein proposed proves to be reasonably tolerant to most conformational changes occurring during protein

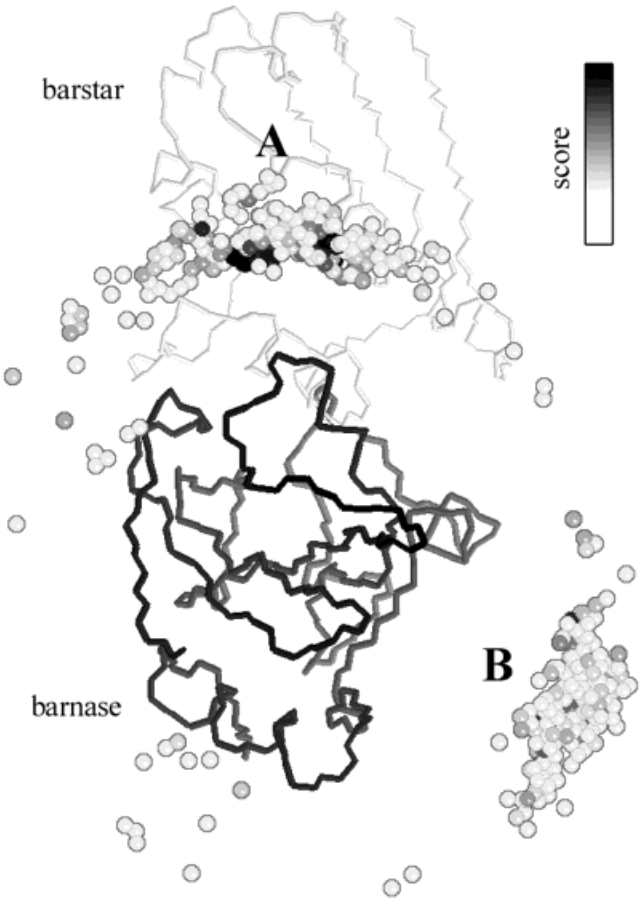

Fig. 7. Global representation of the docking simulation of barnase/barstar complex (1BRSFF). Barnase (target) is shown in dark gray at the center (backbone trace) and the center of mass of the probe molecule (barstar), in each alternative docking position, is represented by a small sphere. The spheres, which represent the top scoring 500 solutions to the docking problem, are coded in shades of gray (darker is stronger) to indicate the relative interaction score of the corresponding solution. Sites **A** and **B** are the two most populated clusters of putative solutions. For comparison, the structure of barstar in the crystallographic complex is also shown, relative to barnase, in light gray (backbone trace).

association, while introducing almost no additional cost to the computational task.

It must be noted that this procedure does not lead to an association model at the atomic level. Candidate models must be subjected to further structure relaxation (via energy minimization or molecular dynamics) to optimize the binding conformation of the amino acids.

It is also clear that the extent of geometric complementarity between two interacting proteins should not take an absolute precedence over other important interaction descriptors, in the process of filtering or scoring putative docked structures. Docking methods that are based on geometric complementarity, as the sole criterion for selecting a population of docked structures, should be considered with extreme care.

The stabilization of macromolecular complexes is the result of a delicate balance between several enthalpic and entropic contributions. Future developments of docking methods should focus on the development and parameterization of new accurate scoring methods to discriminate, in a more reliable way, native-like and wrong docked structures.

We introduced a scoring method that combines, in an empirical way, interaction terms covering most of the chemical features pertinent to molecular interactions. This method represents a new type of approach to the problem of scoring candidate structures of protein complexes and contrasts with energy-based scoring approaches, which seek lowest energy configurations. Since it is not feasible to accurately compute the interaction energy for a large number of putative configurations, these methods have to use coarse approximations of the energy functions, with consequences on accuracy.[23] Additionally, they are based on the assumption that the crystallographic structure always represents the most energetically stable configuration. However, that this might be true for every type of complex is still a questionable matter. In particular, for weak transient complexes such as those occurring between electron transfer proteins, features other than the overall thermodynamic stability might play an important role in molecular recognition and the formation of the complexes. These types of proteins are known to possess large dipole moments and electrostatics should play a major role in guiding their associations. However, the few known structures of electron transfer complexes available show little complementarity and small interface areas,[7] when compared to other types of proteins complexes.

The method described in the present work aims at finding interface properties common to all protein complexes, not necessarily the lowest energy states. Provided that a wide range of protein complexes is used to "teach" an appropriate neural network, as described, the most relevant characteristics that distinguish native docking configurations from other false ones, may be elucidated.

The interaction terms introduced were implemented as "soft" interaction functions, which treat molecular flexibility implicitly. This is essential, since the internal atomic coordinates of the amino acid side chains are not allowed to change during docking and unrealistically close contacts are expected to occur when docking native structures. Yet, the fact that similar docked structures often present unstable values for the interactions terms is indicative that a "softer" form of these functions may have to be developed. These features are currently being investigated in our laboratory and significant improvements of the algorithm and methods will be the subject of future publications.

ACKNOWLEDGMENTS

This work was supported by COST D7, Molecular Recognition, Human Capital and Mobility ERBCHRXCT (J.J.G.M.) 940492 and the National Institute of Health, United States (J.E.W.), grant GM50736.

REFERENCES

1. Jiang F, Kim SH. "Soft docking": matching of molecular surface cubes. J Mol Biol 1991;219:79–102.

2. Shoichet BK, Kuntz ID. Protein docking and complementarity. J Mol Biol 1991; 221:327–346.
3. Cherfils J, Duquerroy S, Janin J. Protein-protein recognition analyzed by docking simulation. Proteins 1991;11:271–280.
4. Bacon DJ, Moult J. Docking by least-squares fitting of molecular surface patterns. J Mol Biol 1992;225:849–858.
5. 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.
6. Helmer-Citterich M, Tramontano A. PUZZLE: a new method for automated protein docking based on surface shape complementarity. J Mol Biol 1994;235:1021–1031.
7. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–2198.
8. Sandak B, Nussinov R, Wolfson HJ. An automated computer vision and robotics-based technique for 3-D flexible biomolecular docking and matching. Comput Appl Biosci 1995;11:87–99.
9. Vakser IA. Protein docking for low-resolution structures. Protein Eng 1995;8:371–377.
10. Walls PH, Sternberg MJ. New algorithm to model protein-protein recognition based on surface complementarity. Applications to antibody-antigen docking. J Mol Biol 1992;228:277–297.
11. Hubbard SJ, Campbell SF, Thornton JM. Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. J Mol Biol 1991;220:507–530.
12. Lin SL, Nussinov R, Fischer D, Wolfson HJ. Molecular surface representations by sparse critical points. Proteins 1994;18:94–101.
13. Norel R, Fischer D, Wolfson HJ, Nussinov R. Molecular surface recognition by a computer vision-based technique. Protein Eng 1994;7:39–46.
14. Norel R, Lin SL, Wolfson HJ, Nussinov R. Shape complementarity at protein-protein interfaces. Biopolymers 1994;34:933–940.
15. 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.
16. Ausiello G, Cesareni G, Helmer-Citterich M. ESCHER: a new docking procedure applied to the reconstruction of protein tertiary structure. Proteins 1997;28:556–567.
17. Gabb HA, Jackson RM, Sternberg MJ. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 1997;272:106–120.
18. Sternberg MJ, Aloy P, Gabb HA, Jackson RM, Moont G, Querol E, Aviles FX. A computational system for modelling flexible protein-protein and protein-DNA docking. Ismb 1998;6:183–192.
19. Meng EC, Kuntz ID, Abraham DJ, Kellogg GE. Evaluating docked complexes with the HINT exponential function and empirical atomic hydrophobicities. J Comput Aided Mol Des 1994;8:299–306.
20. Vakser IA, Aflalo C. Hydrophobic docking: a proposed enhancement to molecular recognition techniques. Proteins 1994;20:320–329.
21. 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.
22. King BL, Vajda S, DeLisi C. Empirical free energy as a target function in docking and design: application to HIV-1 protease inhibitors. FEBS Lett 1996;384:87–91.
23. Jackson RM, Sternberg MJ. A continuum model for protein-protein interactions: application to the docking problem. J Mol Biol 1995;250:258–275.
24. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner P. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. J Am Chem Soc 1984;106:765–784.
25. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KMJ, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins and nucleic acids. J Am Chem Soc 1985;117:5179–5197.
26. Singh J, Thornton JM. Atlas of protein side-chain interactions. Oxford: IRL Press; 1992.
27. Halgren TA. Merck Molecular Force Field. II. MMFF94 van der Waals and Electrostatic Parameters for Intermolecular Interactions. J Comp Chem 1996;17:520–552.
28. Wang Y, Zhang H, Scott RA. A new computational model for protein folding based on atomic solvation. Protein Sci 1995;4:1402–1411.
29. Wang Y, Zhang H, Li W, Scott RA. Discriminating compact nonnative structures from the native structure of globular proteins. Proc Natl Acad Sci USA 1995;92:709–713.
30. Rumelhart DE, Hinton GE, Williams RJ. Learning representations of back-propagation errors. Nature 1986;323:533–536.
31. Pettigrew GW, Gilmour R, Goodhew CF, Hunter DJ, Devreese B, Van Beeumen J, Costa C, Prazeres S, Krippahl L, Palma PN, Moura I, Moura JJ. The surface-charge asymmetry and dimerisation of cytochrome c550 from Paracoccus denitrificans—implications for the interaction with cytochrome c peroxidase. Eur J Biochem 1998;258:559–566.
32. Pettigrew GW, Prazeres S, Costa C, Palma PN, Krippahl L, Moura I, Moura JJG. The structure of an electron transfer complex containing a cytochrome c and a peroxidase. J Biol Chem 1999;274:11383–11389.
33. Morelli X, Dolla A, Czjzek M, Palma PN, Blasco F, Krippahl L, Moura JJG, Guerlesquin F. Heteronuclear NMR and soft docking: an experimental approach for a structural model of the cytochrome c553/ferredoxin complex. Biochemistry 2000; (in press).
34. Morelli X, Czjzek M, Hatchikian CE, Bornet O, Fontecilla-Camps JC, Palma PN, Moura JJG, Guerlesquin F. Structural model of the [Fe]-hydrogenase/cytochrome c553 complex combining TROSY experiments and soft docking calculations. J Biol Chem 2000; (in press).
35. 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. Nat Struct Biol 1996;3:233–239.