Adaptation of a Fast Fourier Transform-Based Docking Algorithm for Protein Design

PO-SSU HUANG,[1] JOHN J. LOVE,[1]* STEPHEN L. MAYO[2]
[1]*Howard Hughes Medical Institute and Division of Biology, California Institute of Technology, Pasadena, California 91125*
[2]*Howard Hughes Medical Institute and Divisions of Biology and Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125*

Received 29 January 2005; Accepted 8 April 2005
DOI 10.1002/jcc.20252
Published online in Wiley InterScience (www.interscience.wiley.com).

Abstract: Designing proteins with novel protein/protein binding properties can be achieved by combining the tools that have been developed independently for protein docking and protein design. We describe here the sequence-independent generation of protein dimer orientations by protein docking for use as scaffolds in protein sequence design algorithms. To dock monomers into sequence-independent dimer conformations, we use a reduced representation in which the side chains are approximated by spheres with atomic radii derived from known C2 symmetry-related homodimers. The interfaces of C2-related homodimers are usually more hydrophobic and protein core-like than the interfaces of heterodimers; we parameterize the radii for docking against this feature to capture and recreate the spatial characteristics of a hydrophobic interface. A fast Fourier transform-based geometric recognition algorithm is used for docking the reduced representation protein models. The resulting docking algorithm successfully predicted the wild-type homodimer orientations in 65 out of 121 dimer test cases. The success rate increases to ~70% for the subset of molecules with large surface area burial in the interface relative to their chain length. Forty-five of the predictions exhibited less than 1 Å C_α RMSD compared to the native X-ray structures. The reduced protein representation therefore appears to be a reasonable approximation and can be used to position protein backbones in plausible orientations for homodimer design.

 J Comput Chem 26: 1222–1232, 2005

Key words: protein docking; protein design; reduced side-chain representation; homodimer; FFT; *de novo* dimer generation

Introduction

Computational protein design using discrete rotamer packing methods has been used to create proteins with novel properties, including enhanced thermostability, catalytic activity, and the ability to bind small-molecules such as TNT.[1–5] In a few cases where the structures of protein/protein dimers are known, protein design methods have been applied successfully to alter binding specificities and to create chimeras with novel functions.[6–8] A prominant goal of protein/protein complex design is to create proteins that target specific sites on molecules of therapeutic or industrial interest. However, compared to the design of small molecule recognition or the redesign of dimers with known structures, the task of creating dimers with novel interfacial geometry from known monomeric structures is a much greater challenge. Depending on the size of the interface, the combinatorial complexity of amino acids involved in a protein/protein interface is likely greater than that of protein/small molecule binding sites. Additionally, because the current molecular mechanics protein design approaches are sensitive to the quality of the structural scaffolds used, having the proper spatial relationship of the subunits in the model dimer is critical, and this requirement increases the difficulty of *de novo* dimer design.

Correspondence to: S. L. Mayo; e-mail: steve@mayo.caltech.edu

*Present address: Department of Chemistry, San Diego State University, 5500 Campanile Dr., San Diego, CA 92182-1030.

Contract/grant sponsor: Howard Hughes Medical Institute

Contract/grant sponsor: Defense Advanced Research Projects Agency

Contract/grant sponsor: Ralph M. Parsons Foundation

Contract/grant sponsor: IBM Shared University Research Grant

Contract/grant sponsor: Army Research Office/Institute for Collaborative Technologies

A properly parameterized docking algorithm can be used to ensure a suitable spatial geometry between the monomeric subunits. For this purpose, we explored the properties of known high-resolution protein homodimer structures. On average, homodimer interfaces are more hydrophobic and bury twice as much surface area as heterodimer interfaces.[9] Their hydrophobic interfaces and symmetry make homodimers especially suitable for molecular mechanics protein design, which has been proved successful in the stabilization of proteins through repacking of their hydrophobic cores.[10] Homodimers are therefore plausible targets not only for parameterizing our docking algorithm so that the computationally docked molecules are likely to have interfaces that accept hydrophobic residues, but also for redesigning the interfaces because we can treat them largely as protein cores.

In this study, we adapted a fast Fourier transform (FFT)-based docking algorithm[11] for protein design. The modified docking algorithm uses a reduced side-chain representation of the molecules to closely approximate the geometry of homodimer interfaces. The reduced representation also provides a background for unbiased interfacial sequence design. Using parameters derived from known homodimer complexes, we tested the performance of our docking algorithm on a set of 121 structures collected by Bahadur et al.[9]

Computational Details and Methods

Reduced Molecular Representation

One of the most important factors in the docking process is the method used to evaluate the fitness of the docked molecules. The fitness function depends on how the molecules are represented in space. For our design purposes, the side chains are explicitly designed in subsequent steps using the ORBIT (optimization of rotamers by iterative techniques) protein design software[1] after the docking process. The docking algorithm must therefore generate a list of plausible dimer orientations based on an approximate molecular representation that includes estimated side chains on a polyalanine backbone to avoid sequence bias before running the sequence design calculation. A crystal structure is used as the scaffold for each of the monomers, with the side-chain atoms beyond C_β deleted. To maintain the overall shape of the surface, the volume originally occupied by a side chain cannot be left empty. Therefore, the most important criterion for our choice of a molecular representation is its ability to approximate the size and shape of an "average" amino acid side chain in a computationally tractable manner.

Conceptually, it is possible to use the original side chains in the docking process, then subsequently replace them during the sequence redesign step. If we use this full side-chain representation, however, we will need a very "soft" scoring function to allow surfaces to overlap, which can potentially lead to backbone clashing. Because the current design algorithm does not allow backbone flexibility, this kind of clashing is strictly prohibited. Moreover, because the side chains on the surface of a protein are usually longer than those found in the core, the use of full surface side chains in the docking process would make the creation of a hydrophobic interface difficult. If the two halves of a dimer are positioned implausibly far away from each other, the design algorithm will converge to a sequence that may not effectively pack the interface. The molecular representation most suitable for our application is therefore one that allows easy "padding" in the side chain voids while maintaining the topology of a potential binding surface. These requirements can be met using spheres of nonphysical sizes to represent atoms on the polyalanine scaffold; some spheres are inflated to make up for the volume of the side chains and hydrogens. Because our polyalanine representation of the backbone does not consider the chemical properties of the side chains, our docking algorithms uses a simple scoring function that only evaluates surface complementarities.

Structural Correlations

The surface complementarity of the docked molecules can be assessed by the correlation function between the two molecules after discretizing them into 3D grids. Using the Fourier correlation theorem and FFT, a correlation map that shows correlation scores as a function of the relative positions of the molecules can be obtained. If one molecule is assumed to be stationary, then the correlation map depicts the results of moving a mobile molecule against the stationary one.[11] For scoring the correlation function, we adopted a 3D grid scoring scheme similar to the one used by Katchalski-Katzir et al.,[11] where all grid points for the mobile molecule are assigned the value "1," grid points that are not part of the protein are assigned "0," and those for the stationary molecule are assigned "−15" if in the interior and "1" if on the surface (Fig. 1). The core is defined by the space around the protein atoms by some radii, and the surface includes the space between the core and 1.5 Å beyond the core. With the grid points set up this way, we can evaluate the correlation scores between the mobile and stationary molecules by counting the number of overlapping grid points between the two. Because penetration to the core penalizes the score, the use of a small penalty value such as "−15" allows slight penetration while maintaining a high level of correlation on the surface. Therefore, the scoring function is intrinsically "soft" when a small penalty is used.

The scoring function described above, however, is not the most appropriate one for protein design purposes, because it provides no distinction between side-chain and backbone penetrations. Because a reduced representation of the surface side chains is used in our docking protocol, some penetration on the side-chain level is considered favorable, as this will create more surface overlap and possibly make the designed interface more viable. Backbone penetrations, on the other hand, must be prohibited. To account for this, a third category of grid point scores was created. In addition to having values of "0" (for vacuum), "1" (for favorable surface) and "−15" (for unfavorable but allowed penetration), a grid point can also be assigned the value of "−1000" if it falls within 1 Å of an atom center (Fig. 1). Although rarely needed, this third "hard shell" ensures no backbone clashes during docking.

C2 Symmetry-Related Dimers

The use of symmetry offers several important advantages. Most significantly, C2 symmetry can greatly reduce the search space required for the docking process. Using FFT in conjunction with

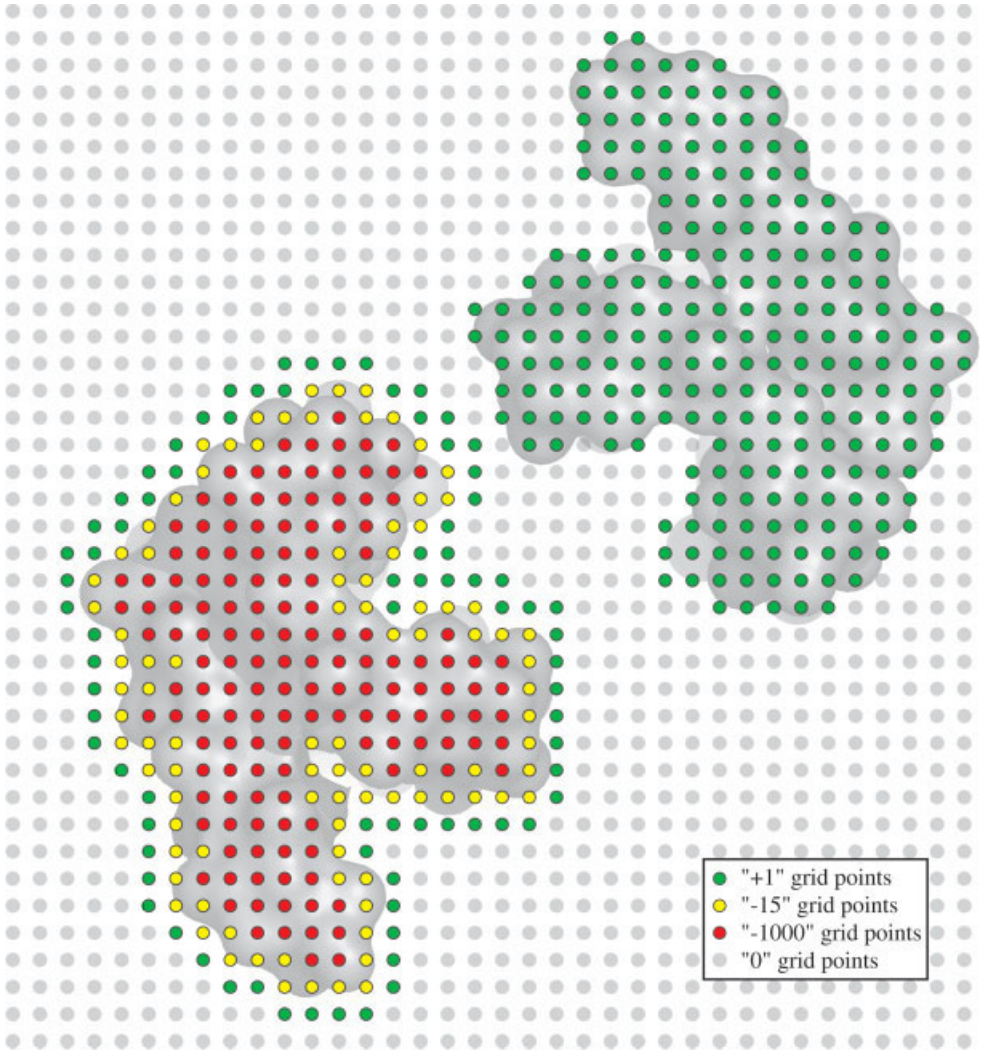

Figure 1. Scoring scheme for discretized molecules. Grid points corresponding to the mobile molecule are assigned uniformly the value of "1"; grid points for the stationary molecule are assigned the values "1", "−15" or "−1000" depending on their locations with respect to the molecule; grid points that are not part of either molecule are assigned the value of "0."

C2 symmetry in the docking stage provides an additional reduction in computational cost.

Some of the reduction in search space results from redundancy associated with C2 symmetry. This can be explained using a coordinate system composed of two symmetry-related coordinate systems. According to Euler's rotation theorem, any rotation can be described by a set of three angles called Euler angles. There are many conventions for the Euler angles; we can depict the concept simply by using the ZXZ convention. In this convention, the three Euler angles, ϕ,θ, ψ, are defined as follows: ϕ is the first rotation ranging from 0 to 2π about the Z-axis, θ is the second rotation ranging from 0 to π about the x′-axis, and ψ is the third rotation ranging from 0 to 2π about the z′-axis (Fig. 2). These three rotations are not commutative, and therefore must be applied in this specific order. The three angles are depicted in Figure 2 with a modified coordinate system to illustrate the search space reduction associated with C2 symmetry. In a Cartesian coordinate system, to thoroughly explore the rotations of a rigid body with respect to the coordinate system (or with respect to another rigid body in the case of docking searches), the rotational space that must be covered is $2\pi \times \pi \times 2\pi$, as defined by the full ranges of the three Euler angles. In Figure 2, the coordinate system shown can be described as a combination of two separate coordinate systems, XYZ and $\xi\eta$Z, related to each other by a twofold (C2) symmetry about the Z-axis. By the definition of C2 symmetry, any point in the XYZ coordinate system corresponds to a point in the $\xi\eta$Z system by a 180° rotation. Due to the 180° rotational symmetry with respect to the Z-axis, the ranges of ϕ and ψ in this XYZ, $\xi\eta$Z combined coordinate system are both reduced by half to $0 = \phi = \pi$, $0 = \psi = \pi$, while the range of θ remains the same. Rotations beyond the range of 0 to π are redundant because the resulting positions can always be folded back to positions within

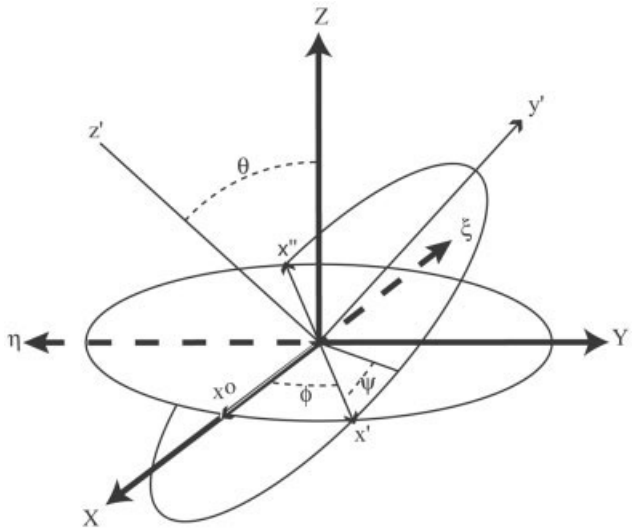

Figure 2. Rotational search space reduction with C2 symmetry related dimers. Two symmetry related coordinate systems are shown: XYZ and $\xi\eta$Z.

the range of 0 to π, as illustrated by vectors x′ and x″. Vector x′ is obtained by rotating vector x° about the Z-axis by ϕ, and by C2 symmetry it corresponds to the vector x″, which can also be obtained by rotating vector x° by $\phi + \pi$. Due to this redundancy, the range of ϕ can be reduced from 2π to π, and this is also true for ψ.

An additional reduction of rotational search space can be achieved when translational searches are performed. This concept is illustrated in Figure 3. To maintain the C2 symmetry, rotations performed on the subunits must be synchronized—the same rotational operation must be performed on both molecules. For clarity, the two molecules in Figure 3 are set at a fixed distance from each other when they are rotated, and the rotations are performed at their respective geometric centers about axes that are parallel to the symmetry axis. One of the properties of cyclic symmetry groups such as C2 is that the subunits are related by rotation about a symmetry axis, and they are always on a plane perpendicular to the symmetry axis. As illustrated in Figure 3, when each of the subunits of the C2 symmetry related dimer is rotated to a new C2 symmetry-related orientation, the rotational steps required to achieve this new orientation can always be replaced by translations on this plane. Therefore, when translational steps are included in a docking search, rotations around the symmetry axis or any axis parallel to the symmetry axis (defined by ϕ) can be eliminated from the searches. FFT replaces explicit translational searches with an efficient computational process that generates the translational correlation map. As a result, to thoroughly search all possible C2-related dimer orientations, we only need to cover $\pi \times \pi$ (the ranges of θ and ψ) when FFT is used; the search space is thus reduced by a factor of 4π.

The computer memory required for discretizing the molecules can also be reduced when docking C2 symmetry related subunits. As described previously, the cyclic symmetry requires the subunits of a C2 related dimer to be on the same 2D plane. This requirement eliminates the need to explore any translations parallel to the symmetry axis. If the Z-axis is used as the symmetry axis for a pair of subunits, only translations along the X- and Y-axis are required to produce dimers with preserved C2 symmetry. The dimension of the arena that is parallel to the symmetry axis can therefore be reduced to the length of the long axis of the molecule instead of three times this length. The FFT implementation used in our docking algorithm, however, requires the number of grid points

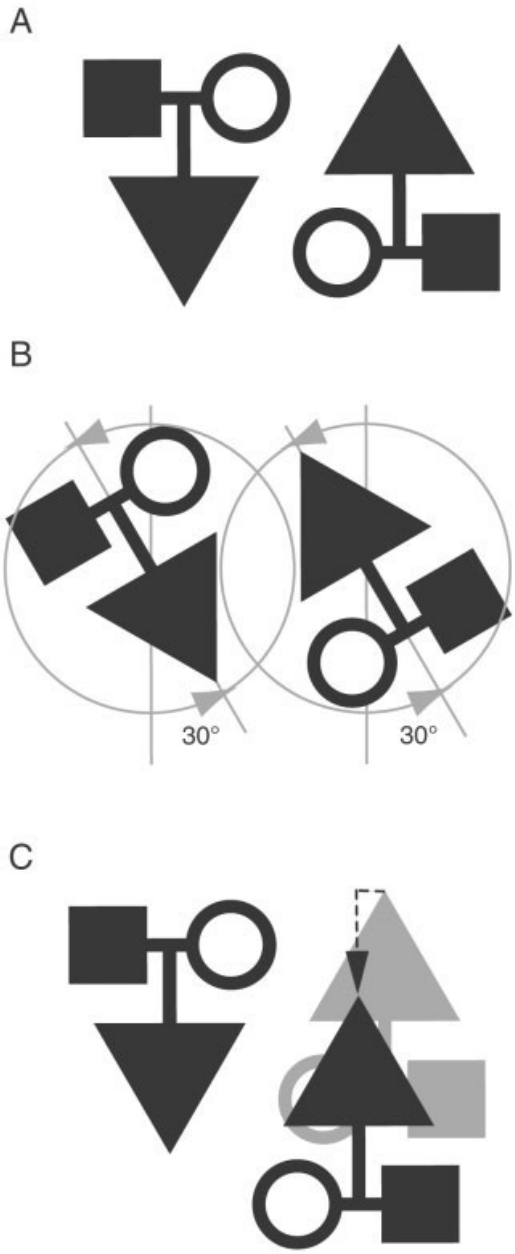

Figure 3. Translational equivalence. (A) A pair of C2 symmetry related monomers in its initial position. (B) Each monomer of the dimer is rotated by 30° to a new C2 symmetry orientation. (C) The same orientation as in B can be obtained by two translational moves.

Table 1. Atomic Radii Determined from Iteration Cycles.

	Ranges of good atomic radii (Å)[a]		
Atom label	1a7w	1c9o	Final values used (Å)
N	1.4–1.6	1.4–1.55	1.4
O	1.3–1.45	1.3–1.35	1.3
C	1.7–1.8	1.7–2.55	1.75
CA	2.15–2.4	2.3–2.4	2.35
CB	2.15	2.2–2.25	2.15

[a]The range is determined from the radii combinations that gave the best 200 docked scores.

along each dimension to be a power of 2, and thus it is convenient to simply reduce the dimension parallel to the symmetry axis by a factor of 2. The number of grid points required for the search becomes half. Assuming 1° is used as the rotational increment, docking a C2-related dimer is 1440 times faster than docking a dimer with no symmetry (4π rotational search space reduction times twofold grid point reduction equals a total reduction of 8π; for 1° rotational increments, this equals 8×180 or 1440-fold reduction).

Determination of Practical Atomic Radii

Because the side chains are truncated in our protein model, we approximate side-chain volumes with spheres centered at the C_β atoms; the projection distance from the C_β atoms (atomic radii) should be chosen to ensure that the resulting dimer interface retains enough space for side-chain placement. We determined an appropriate C_β atomic radius by calculating the surface complementarity scores for two high-resolution crystal structures of protein/protein complexes in the PDB. The molecules were discretized in their native crystal orientations with the side chains truncated, and surface complementarities were calculated. Initially, we tried several radii keeping a uniform radius for all the atoms. However, complexes that had backbone-to-backbone hydrogen bonds always clashed at these points. It was obvious that a uniform radius was inadequate, so we decided to parameterize the radii for each of the five atom types in our polyalanine model. We tested a range of values for amide nitrogen (atom label: N), C_α (atom label: CA), C_β (atom label: CB), carbonyl carbon (atom label: C), and oxygen (atom label: O). The crystal structures of two high-resolution complexes were used: PDB entries 1a7w and 1c9o (1.55- and 1.17-Å resolution, respectively). Combinations of atomic radii within the following ranges were tested: nitrogen between 1.4 and 1.6 Å; oxygen between 1.3 and 1.5 Å; and all carbons (carbonyl carbon, C_α and C_β) between 1.7 and 2.4 Å. A 0.05-Å increment was used to step across each of the ranges. The resulting docking correlation scores were sorted and the radii combinations that gave top ranking scores were analyzed.

The correlation score can be very sensitive to changes in radii for some atom types but not for others. The level of sensitivity is reflected in the "good radii" ranges shown in Table 1; a wider "good radii" range indicates that the correlation score is less sensitive. The values used for subsequent docking tests were determined as follows. If the radii ranges obtained from both structures were about equal, as for atoms N and O, the minimum value of the range was used. If the radii ranges from the two structures matched poorly, as for atoms C, CA, and CB, the mean value of the narrower range was used. The final atomic radii obtained from the parameterizations (Table 1, last column) were used to test the docking algorithm.

Docking Test Cases

We tested the performance of our docking algorithm on a set of 121 structures of the 122 collected by Bahadur et al.[9] One of the structures from the set (1alo) was not used due to its exceptionally large size (907 residues). The structures of the 121 dimers were downloaded in their biological unit coordinates from the PDB.[12] The protein coordinates were processed to resolve any naming and numbering discrepancies, and the subunits were separated into individual files. For each docking calculation, we loaded the coordinates of just one of the subunits, and created the other subunit by duplicating and rotating the loaded coordinates by 180° about the x-axis. Except for orientation, the two subunits were thus identical to each other. The rotational search space was sampled with 1° increments over 180° for both the y- and the z-axes. Depending on subunit size, the number of grids used for the arena was either 128 or 256 in the y and z dimensions, and half this number in the x dimension. All tests were carried out at 1 Å grid spacing.

Docking searches and surface complementarities were calculated for each of the 121 dimers; coordinates for the top 50 orientations were generated for each dimer and compared to the coordinates of the PDB structure. RMSDs were evaluated between all the C_α atoms from both subunits of the docked orientations and their corresponding C_α atoms on the PDB structures. The results are shown in Table 2; RMSDs are only listed if less than 3 Å. From the top 50 docked structures for each dimer, the rank with the lowest RMSD (best match) is listed along with the first rank with a RMSD <3 Å. The buried interface surface area contributed by one of the subunits (reported as B/2) and the ratio of this surface area to the number of residues in the subunit (Area/Res) are also reported. The 121 dimers tested are sorted according to Area/Res in Table 3 along with indications of hits (shown with "+") and misses (shown with "."). A dimer is considered to be a hit if there is at least one docked dimer in its top 50 orientations with an RMSD <3 Å.

Results and Discussion

We achieved 65 successful predictions (hits) out of 121 test cases, slightly above 50%. Although there are no docking benchmarks focusing exclusively on homodimers, our results are comparable to the few homodimer docking cases reported previously, in which three of five test cases were within the top 50 ranked structures.[13] The ranks produced by our simple surface complementarity scoring scheme, however, do not always correlate with the RMSDs of the models. In most cases, the closest match to the wild-type orientation does not receive the highest correlation score, although

one-third of the closest matches are ranked in the top 10 (Table 2, column 4). The best match is predicted as the top rank in four cases (1ajs, 1tox, 1trk, and 1vfr). This is not surprising, because we are using a reduced representation model of the proteins. Because all four of these cases use backbones extensively to achieve binding specificity, it appears that only dimers with these properties are correctly ranked by our scoring scheme. The inability to rank models correctly is a problem that plagues all docking algorithms, and several research groups have developed sophisticated scoring functions that include the properties of the side chains to ameliorate this problem.[13–15]

It should be noted that for our purposes, the ranking does not significantly affect our intended use. We are interested in identifying all distinct dimer orientations for *de novo* protein design, and as long as the interface is plausible it is not necessary to recover the native binding interface. When testing our docking algorithm, however, it is important to recover the native binding interface (regardless of their reported ranks) and to obtain good matches when they are identified because these two factors directly check the validity of our atomic radii. As shown in Table 2, 45 of the 65 successfully docked dimers have RMSDs of less than 1 Å, indicating that our reduced representation protein models are reasonable approximations. We found that we can reliably reproduce most dimers that have backbone-to-backbone interactions across the interface, namely the hydrogen bond pairing between two intermolecular β-strands, despite the absence of an explicit hydrogen bonding term (Fig. 4). An example of a dimer making β-strand contacts across the dimer interface is shown in Figure 4A. The results suggest that our atomic radii capture these hydrogen bonds well. However, for the helical protein shown in Figure 4B, the overall dimer orientation is recovered, but the docked molecule is slightly offset from the native orientation.

Our successful prediction for dimers with backbone-to-backbone hydrogen bonds across the interface suggests that the radii used for atom N, C_α, C, and O are plausible. However, the task of parameterizing the C_β atoms remains a challenge. The dimer orientation predicted for the 1rfb structure (not shown), for example, showed a less promising C_α RMSD of 2.38 Å when compared to the native structure. Closer examination revealed that the offset is likely due to phenylalanine and tyrosine residues in the interface. Aromatic residues have relatively long flat side chains and cannot be modeled well by spheres. Similarly, the distances between the backbones of the docked 1rpo dimer (shown in Fig. 4B) are too close, again indicating that the C_β radius used for this particular type of dimer is not large enough. Furthermore, because the docking algorithm searches for the highest complementarity between the dimers, interdigitation between the spheres representing side chains on the surface of a helix is preferred over stacking the spheres head-on—as seen with the native 1rpo dimer. This is why we are unable to reproduce helix-to-helix interfaces with high precision. Nevertheless, this preference for interdigitation serendipitously allows the reproduction of native hydrogen bonding patterns in our test cases that form crossinterface hydrogen bonds between β-strands even without the use of an explicit hydrogen bond term in the scoring function.

Although the use of a single, median-sized sphere for all side chains cannot accurately represent both large and small amino acids, the fact that both sides of an interface are "designed" in our method reduces this concern. For any large amino acid selected in the interface, there has to be a complementary small amino acid on the receiving end to avoid clashing. The benefit of creating such "knobs into holes" characteristics in the interface remains an open question as most protein–protein interfaces are relatively flat,[16] but this knobs into holes feature could potentially contribute to higher levels of binding specificity.

Comparing the ratios of interface area to the total number of residues in each monomer (reported as Area/Res in Tables 2 and 3) illustrates another interesting point. These ratios are used as a rough measure of the relative size of the interface in the context of the entire subunit. Sorting the 121 dimers according to this ratio reveals that our docking success rate is much higher for dimers that bury relatively large surface areas with respect to their sizes compared to those that do not. For example, our success rate is 70% for dimers with Area/Res >7.5 Å^2 (60 of the 121 dimers tested) vs. 50% for the entire set. This may be explained by the fact that larger proteins have more competing sites on their surfaces that could provide good docking correlation scores. By reporting only the top 50 docked dimers in our tests, the rank listings may not be deep enough to include the native orientations; these larger proteins are therefore more likely to be "misses." For example, the 1aor dimer crystal structure (not shown) shows a highly complementary interface, but because this protein is relatively large (605 residues), our docking algorithm does not pick up the native orientation in the top 50. Other docking algorithms severely penalize the competing sites ("false positives") by incorporating biochemical data or electrostatic terms in the scoring function, features we cannot include given that our protein model does not include full side chains. The trend observed in Table 3 is not attainable, however, if we simply sort our results by the number of residues. Assuming that typical protein interface sizes fall within a narrow range, the docking success rate is expected to be higher for docking smaller proteins. However, the interface sizes found in our test cases range from 498 to 7149 Å,^{2} which is sufficiently broad that normalizing by protein size (i.e., the number of residues) is necessary to observe the trend reported above.

Even though our docking algorithm was developed to design novel dimers, it contains all the basic components found in docking algorithms that use FFT as a search tool and can be used as a stand-alone algorithm for predicting dimer orientations. However, because of differences in the nature of the problems to be addressed, our algorithm should not be compared to other general docking algorithms. Protein docking is an active area of research, and much progress has been made in developing algorithms suited to this purpose.[14,15,17–38] Although the use of sequence specific reduced representations for side chains have been reported,[18,28] the trend is to model proteins with greater accuracy either through implicit energy terms or explicit simulations. This includes the incorporation of desolvation terms, electrostatic terms, side-chain flexibility, and Monte Carlo simulations, among others. Our algorithm, on the other hand, relies largely on a sequence-independent reduced representation of the protein. Although potentially useful for protein design purposes, this reduce representation greatly diminishes our chances of predicting native orientations, especially if the driving force for association is specific side-chain–side-chain interactions instead of surface geometric complementarity.

Table 2. Docking Results.

			Best match to wild-type[b]		First rank with RMSD <3 Å[b]		Interface[d]	
PDB	Residues	Long axis (Å)[a]	Rank[c]	RMSD (Å)[c]	Rank[c]	RMSD (Å)[c]	B/2 (Å^2)[e]	Area/Res (Å^2)[f]
12as	327	72.3	—	—	—	—	1989	6.1
1a3c	166	55.6	—	—	—	—	853	5.1
1a4i	285	64.9	—	—	—	—	1353	4.7
1a4u	254	63.1	7	0.291	1	0.656	2547	10.0
1aa7	158	47.7	—	—	—	—	1125	7.1
1ad3	446	124.8	4	0.314	1	0.533	3936	8.8
1ade	431	81.4	17	2.452	17	2.452	2708	6.3
1af5	126	60.5	—	—	—	—	856	6.8
1afw	390	72.7	31	0.333	1	0.558	2400	6.2
1ajs	412	100.9	1	1.124	1	1.124	3401	8.3
1amk	250	59.0	50	0.072	1	0.532	1477	5.9
1aor	605	79.3	—	—	—	—	1180	2.0
1aq6	245	60.9	33	0.716	28	1.097	2232	9.1
1auo	218	53.8	—	—	—	—	662	3.0
1b3a	67	104.0	—	—	—	—	763	11.4
1b5e	241	67.3	2	0.156	1	0.156	2581	10.7
1b67	68	60.5	44	0.537	44	0.537	1607	23.6
1b8a	438	107.4	—	—	—	—	4391	10.0
1b8j	448	81.7	22	0.309	1	0.348	3794	8.5
1bam	200	61.7	—	—	—	—	745	3.7
1bbh	131	54.3	—	—	—	—	771	5.9
1bd0	381	95.0	8	0.699	8	0.699	3091	8.1
1bif	432	85.9	—	—	—	—	858	2.0
1biq	339	77.3	15	0.408	1	1.324	3004	8.9
1bis	146	54.9	16	2.563	16	2.563	1495	10.2
1bjw	381	84.5	25	0.501	2	0.795	2938	7.7
1bkp	278	65.8	—	—	—	—	2206	7.9
1bmd	326	65.8	14	0.311	13	0.312	1564	4.8
1brw	433	82.0	—	—	—	—	1083	2.5
1bsl	323	68.7	5	1.38	5	1.38	1918	5.9
1bsr	124	67.8	13	0.975	3	0.997	1888	15.2
1buo	121	86.3	13	0.259	2	0.536	1972	16.3
1bxg	349	67.3	—	—	—	—	1041	3.0
1bxk	341	79.2	—	—	—	—	1286	3.8
1cdc	96	69.5	2	0.621	1	0.638	3918	40.8
1cg2	389	113.4	—	—	—	—	1298	3.3
1chm	401	88.3	23	0.525	1	0.655	3171	7.9
1cmb	104	53.1	6	1.494	1	1.9	1797	17.3
1cnz	363	89.5	11	1.565	11	1.565	2447	6.7
1coz	126	51.5	—	—	—	—	1050	8.3
1csh	435	88.9	34	0.033	1	0.343	5057	11.6
1ctt	294	65.8	45	0.385	1	0.632	1990	6.8
1cvu	551	146.0	14	1.2	3	1.282	2436	4.4
1czj	110	54.4	—	—	—	—	829	7.5
1daa	277	70.2	15	0.448	15	0.448	2193	7.9
1dor	311	75.9	30	0.368	16	1.326	2189	7.0
1dpg	485	102.8	—	—	—	—	2293	4.7
1dqs	381	77.3	4	0.844	3	1.843	1640	4.3
1dxg	36	31.7	21	0.326	21	0.326	729	20.3
1e98	210	55.2	—	—	—	—	770	3.7
1ebh	436	78.8	43	0.799	23	2.481	1784	4.1
1f13	722	126.8	—	—	—	—	2556	3.5
1fip	73	55.2	23	0.55	4	0.658	1836	25.2
1fro	176	68.1	3	0.151	1	0.203	3505	19.9
1gvp	87	54.4	5	1.5	5	1.5	908	10.4
1hhp	99	52.8	2	0.056	2	0.056	1599	16.2
1hjr	158	63.7	—	—	—	—	962	6.1
1hss	111	45.1	44	1.4	11	1.624	1101	9.9
1hxp	340	78.1	22	0.259	1	0.5	3402	10.0
1icw	69	43.4	47	0.898	47	0.898	954	13.8
1imb	273	60.4	36	1.706	9	2.573	1623	5.9
1isa	192	60.1	—	—	—	—	920	4.8
1ivy	452	77.8	—	—	—	—	1601	3.5
1jhg	101	59.1	—	—	—	—	2207	21.9

(*continued*)

Table 2. (*Continued*)

			Best match to wild-type[b]		First rank with RMSD <3 Å[b]		Interface[d]	
PDB	Residues	Long axis (Å)[a]	Rank[c]	RMSD (Å)[c]	Rank[c]	RMSD (Å)[c]	B/2 (Å^2)[e]	Area/Res (Å^2)[f]
1jsg	111	60.2	—	—	—	—	794	7.2
1kba	66	47.7	28	2.286	28	2.286	498	7.5
1kpf	126	45.8	5	0.31	3	0.358	1867	14.8
1lyn	125	61.9	—	—	—	—	948	7.6
1m6p	146	53.3	37	1.143	9	1.282	1025	7.0
1mkb	171	58.0	2	0.46	1	0.799	1605	9.4
1mor	366	70.0	50	0.513	21	1.401	2540	6.9
1nox	200	72.3	10	0.397	1	0.741	3033	15.2
1nse	416	84.0	—	—	—	—	2736	6.6
1nsy	271	68.9	19	0.486	19	0.486	2592	9.6
1oac	719	109.3	—	—	—	—	7149	9.9
1opy	123	52.5	28	0.249	23	0.58	1048	8.5
1pgt	209	58.3	15	2.808	15	2.808	1238	5.9
1pre	449	131.0	29	0.844	17	1.272	2300	5.1
1qfh	212	98.5	—	—	—	—	2264	10.7
1qhi	304	68.5	23	0.713	8	2.149	1714	5.6
1qr2	230	75.8	—	—	—	—	1947	8.5
1r2f	283	73.4	48	2.601	36	2.635	1746	6.2
1reg	122	105.3	—	—	—	—	659	5.4
1rfb	119	63.7	9	1.28	1	1.752	2650	22.3
1rpo	61	53.4	44	0.704	1	1.084	1405	23.0
1ses	421	136.3	—	—	—	—	2211	5.3
1slt	133	42.7	—	—	—	—	536	4.0
1smn	241	54.8	26	0.999	1	1.552	866	3.6
1smt	98	77.2	34	0.267	1	0.3	1970	20.1
1sox	463	81.3	—	—	—	—	1404	3.0
1tcl	175	57.7	—	—	—	—	1540	8.8
1tox	515	94.7	1	0.026	1	0.026	3721	7.2
1trk	678	105.6	1	0.371	1	0.371	4476	6.6
1uby	348	78.1	12	2.564	6	2.576	2168	6.2
1utg	70	46.6	44	1.263	9	2.391	1485	21.2
1vfr	217	72.6	1	0.368	1	0.368	3431	15.8
1vok	192	73.4	—	—	—	—	1577	8.2
1wtl	108	58.6	—	—	—	—	698	6.5
1xso	149	47.0	—	—	—	—	662	4.4
2arc	161	56.0	—	—	—	—	765	4.8
2ccy	127	53.9	—	—	—	—	792	6.2
2hdh	286	74.4	—	—	—	—	1524	5.3
2ilk	155	78.0	8	0.314	1	0.745	4542	29.3
2lig	157	89.1	—	—	—	—	1686	10.7
2mcg	215	77.7	—	—	—	—	1646	7.7
2nac	374	75.5	—	—	—	—	3789	10.1
2ohx	374	77.8	—	—	—	—	1718	4.6
2spc	106	118.3	—	—	—	—	2508	23.7
2sqc	623	84.8	—	—	—	—	809	1.3
2tct	198	77.1	45	1.198	45	1.198	2675	13.5
2tgi	112	75.6	—	—	—	—	1262	11.3
3dap	320	79.9	—	—	—	—	2661	8.3
3grs	461	79.1	—	—	—	—	3302	7.2
3sdh	145	49.7	—	—	—	—	873	6.0
3ssi	108	56.6	36	0.824	22	2.878	866	8.0
4cha	239	55.5	—	—	—	—	1026	4.3
4kbp	424	79.4	24	2.793	24	2.793	1478	3.5
5csm	250	69.6	25	2.13	9	2.962	2007	8.0
5rub	436	86.0	—	—	—	—	2859	6.6
8prk	282	56.8	—	—	—	—	969	3.4
9wga	170	62.3	2	0.132	1	0.139	2293	13.5

[a]The long axis of the molecule is determined by two times the distance from its geometric center to the farthest atom.
[b]Only the molecules from the highest 50 correlation scores are considered.
[c]— means no match within 3 Å RMSD from the top 50 ranked molecules.
[d]Per subunit.
[e]Data in this column are taken from ref. 9.
[f]The interface area contributed by each subunit divided by the number of residues per subunit.

Table 3. Area/Residue Ratios and Docking Hits.[a]

PDB	Area/Res (Å^2)[b]	Hit[c]	PDB	Area/Res (Å^2)[b]	Hit[c]	PDB	Area/Res (Å^2)[b]	Hit[c]
2sqc	1.3	.	1hjr	6.1	.	1aq6	9.1	+
1aor	2.0	.	1afw	6.2	+	1mkb	9.4	+
1bif	2.0	.	1r2f	6.2	+	1nsy	9.6	+
1brw	2.5	.	1uby	6.2	+	1hss	9.9	+
1bxg	3.0	.	2ccy	6.2	.	1oac	9.9	.
1sox	3.0	.	1ade	6.3	+	1hxp	10.0	+
1auo	3.0	.	1wtl	6.5	.	1b8a	10.0	.
1cg2	3.3	.	5rub	6.6	.	1a4u	10.0	+
8prk	3.4	.	1nse	6.6	.	2nac	10.1	.
4kbp	3.5	+	1trk	6.6	+	1bis	10.2	+
1fl3	3.5	.	1cnz	6.7	+	1gvp	10.4	+
1ivy	3.5	.	1ctt	6.8	+	1qfh	10.7	.
1smn	3.6	+	1af5	6.8	.	1b5e	10.7	+
1e98	3.7	.	1mor	6.9	+	2lig	10.7	.
1bam	3.7	.	1m6p	7.0	+	2tgi	11.3	.
1bxk	3.8	.	1dor	7.0	+	1b3a	11.4	.
1slt	4.0	.	1aa7	7.1	.	1csh	11.6	+
1ebh	4.1	+	1jsg	7.2	.	9wga	13.5	+
4cha	4.3	.	3grs	7.2	.	2tct	13.5	+
1dqs	4.3	+	1tox	7.2	+	1icw	13.8	+
1cvu	4.4	+	1czj	7.5	.	1kpf	14.8	+
1xso	4.4	.	1kba	7.5	+	1nox	15.2	+
2ohx	4.6	.	1lyn	7.6	.	1bsr	15.2	+
1dpg	4.7	.	2mcg	7.7	.	1vfr	15.8	+
1a4i	4.7	.	1bjw	7.7	+	1hhp	16.2	+
2arc	4.8	.	1chm	7.9	+	1buo	16.3	+
1isa	4.8	.	1daa	7.9	+	1cmb	17.3	+
1bmd	4.8	+	1bkp	7.9	.	1fro	19.9	+
1pre	5.1	+	3ssi	8.0	+	1smt	20.1	+
1a3c	5.1	.	5csm	8.0	+	1dxg	20.3	+
1ses	5.3	.	1bd0	8.1	+	1utg	21.2	+
2hdh	5.3	.	1vok	8.2	.	1jhg	21.9	.
1reg	5.4	.	1ajs	8.3	+	1rfb	22.3	+
1qhi	5.6	+	3dap	8.3	.	1rpo	23.0	+
1bbh	5.9	.	1coz	8.3	.	1b67	23.6	+
1amk	5.9	+	1qr2	8.5	.	2spc	23.7	.
1pgt	5.9	+	1b8j	8.5	+	1fip	25.2	+
1bsl	5.9	+	1opy	8.5	+	2ilk	29.3	+
1imb	5.9	+	1tc1	8.8	.	1cdc	40.8	+
3sdh	6.0	.	1ad3	8.8	+			
12as	6.1	.	1biq	8.9	+			

[a]Sorted by the Area/Res ratio of each PDB entry in ascending order.
[b]The interface area contributed by each subunit divided by the number of residues per subunit.
[c]Only the top 50 correlation score ranked dimers are considered.
"+" means there is at least one docked dimer with an RMSD less than 3 Å to the native structure. "." means there is no match.

In summary, our docking algorithm performs reasonably well in docking a test set of 121 homodimers. For matches that fall below the 3-Å threshold, the average best RMSD is 0.87 Å, indicating that the atomic radii used for the backbone and the reduced representation side chains on the C_β atoms are reasonable. There is about a 50% success rate in finding the native orientations in the entire docking set. The success rate is significantly higher for dimers that form relatively large interfaces with respect to their amino-acid chain length.

Conclusions

We have described the development of a docking algorithm that generates dimer orientations for protein design purposes. To position the backbones correctly for protein design and to avoid bias toward wild-type sequences, the wild-type side chains are not considered in the docking process. The strategies employed include the use of 3D grids to represent the protein molecules, and spheres to approximate the side chains. The spheres are defined by

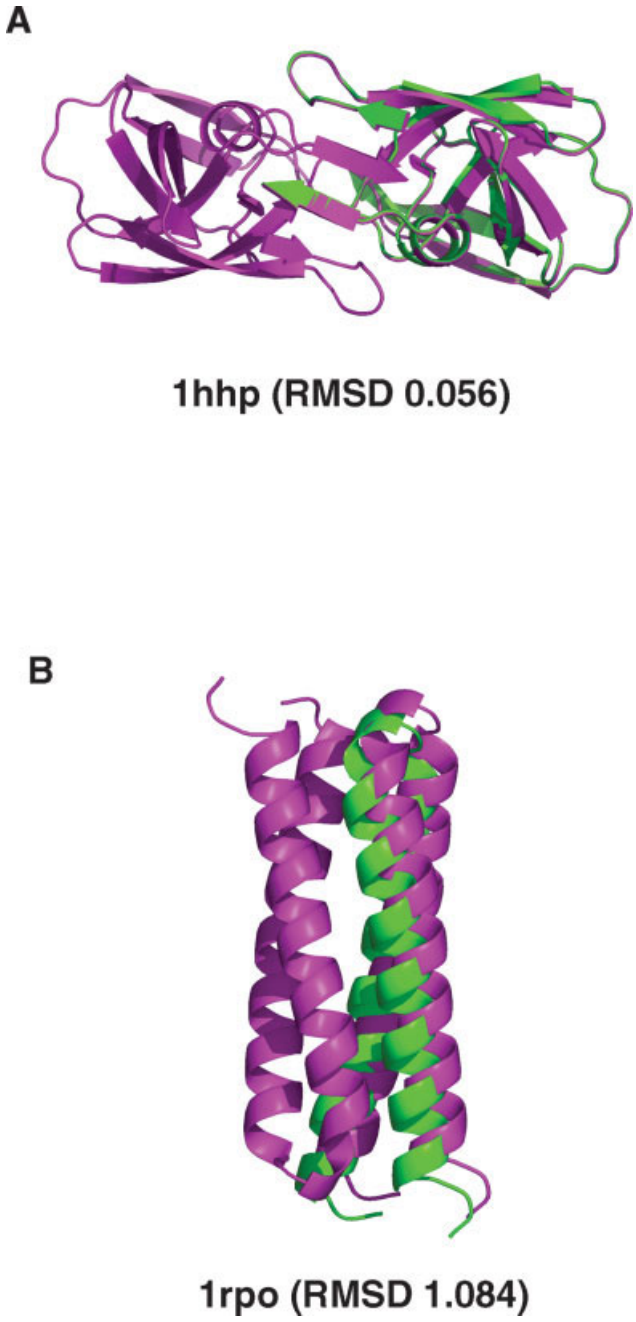

Figure 4. Sample docking results. Native orientations are shown in magenta and docked orientations are shown in green. For each of the dimers, one of the monomers of the docked structure was superimposed on the corresponding native monomer (shown here on the left); the overlapping or offset portions of the docked structure relative to the native structure can be seen by the amount of green displayed. C_α RMSDs are indicated in paranthesis next to the PDB codes. (A) Proteins that use backbone hydrogen bonds in the interface to form crossdimer β-sheets. (B) Helical proteins that primarily use side-chain/side-chain interactions to form the interface.

atomic radii, which are determined using known high-resolution dimer structures. Established FFT correlation methods are employed to efficiently cover all translational dimensions and search through all six degrees of freedom, and surface shape complementarities are used to score the fitness of the docked structures. Because C2 symmetry related homodimers tend to bury more surface area and use more hydrophobic amino acids in the interface, their interfaces are more protein core-like and should be modeled well by our protein design algorithms. We therefore parameterized our docking algorithm using C2 symmetry-related homodimers and used dimers of this type as test cases. Imposing C2 symmetry also allowed us to make modifications that significantly improve computational efficiency. The resulting docking algorithm performed reasonably well in the 121 test cases used to validate the reduced representation protein model. These results suggest that the reduced protein side-chain representation employed by our algorithm is a reasonable estimate, and the shapes defined by this representation can be used to position protein backbones to form plausible dimer orientations for protein design.

Acknowledgments

The authors would like to thank Marie Ary, Christina L. Vizcarra, and Benjamin D. Allen for editing and reviewing the manuscript.

References

1. Dahiyat, B. I.; Mayo, S. L. Science 1997, 278, 82.
2. Malakauskas, S. M.; Mayo, S. L. Nat Struct Biol 1998, 5, 470.
3. Bolon, D. N.; Mayo, S. L. Proc Natl Acad Sci USA 2001, 98, 14274.
4. Looger, L. L.; Dwyer, M. A.; Smith, J. J.; Hellinga, H. W. Nature 2003, 423, 185.
5. Dwyer, M. A.; Looger, L. L.; Hellinga, H. W. Science 2004, 304, 1967.
6. Bolon, D. N.; Wah, D. A.; Hersch, G. L.; Baker, T. A.; Sauer, R. T. Mol Cell 2004, 13, 443.
7. Kortemme, T.; Joachimiak, L. A.; Bullock, A. N.; Schuler, A. D.; Stoddard, B. L.; Baker, D. Nat Struct Mol Biol 2004, 11, 371.
8. Chevalier, B. S.; Kortemme, T.; Chadsey, M. S.; Baker, D.; Monnat, R. J.; Stoddard, B. L. Mol Cell 2002, 10, 895.
9. Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin, J. Proteins 2003, 53, 708.
10. Dahiyat, B. I.; Mayo, S. L. Proc Natl Acad Sci USA 1997, 94, 10172.
11. Katchalskikatzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. Proc Natl Acad Sci USA 1992, 89, 2195.
12. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235.
13. Chen, R.; Weng, Z. P. Proteins 2002, 47, 281.
14. Heifetz, A.; Katchalski–Katzir, E.; Eisenstein, M. Protein Sci 2002, 11, 571–587.
15. Murphy, J.; Gatchell, D. W.; Prasad, J. C.; Vajda, S. Proteins 2003, 53, 840.
16. Peters, K. P.; Fauck, J.; Frommel, C. J Mol Biol 1996, 256, 201–213.
17. Fernandez–Recio, J.; Totrov, M.; Abagyan, R. J Mol Biol 2004, 335, 843.
18. Zacharias, M. Protein Sci 2003, 12, 1271–1282.

19. Mendez, R.; Leplae, R.; De Maria, L.; Wodak, S. J. Proteins 2003, 52, 51.
20. Li, L.; Chen, R.; Weng, Z. P. Proteins 2003, 53, 693.
21. Heifetz, A.; Eisenstein, M. Protein Eng 2003, 16, 179.
22. Fernandez–Recio, J.; Totrov, M.; Abagyan, R. Proteins 2003, 52, 113.
23. Chen, R.; Weng, Z. P. Proteins 2003, 51, 397.
24. Chen, R.; Li, L.; Weng, Z. P. Proteins 2003, 52, 80.
25. Lorber, D. M.; Udo, M. K.; Shoichet, B. K. Protein Sci 2002, 11, 1393.
26. Camacho, C. J.; Gatchell, D. W.; Kimura, S. R.; Vajda, S. Proteins 2000, 40, 525.
27. Ewing, T. J. A.; Kuntz, I. D. J Comput Chem 1997, 18, 1175.
28. Gray, J. J.; Moughon, S.; Wang, C.; Schueler–Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. J Mol Biol 2003, 331, 281.
29. Duan, Y. H.; Reddy, B. V. B.; Kaznessis, Y. N. Protein Sci 2005, 14, 316.
30. Lee, K.; Czaplewski, C.; Kim, S. Y.; Lee, J. J Comput Chem 2005, 26, 78.
31. Shatsky, M.; Dror, O.; Schneidman–Duhovny, D.; Nussinov, R.; Wolfson, H. J. Nucleic Acids Res 2004, 32, W503.
32. Berchanski, A.; Shapira, B.; Eisenstein, M. Proteins Struct Funct Bioinformat 2004, 56, 130.
33. Smith, G. R.; Sternberg, M. J. E.; Bates, P. A. Biophys J 2004, 86, 413A.
34. Comeau, S. R.; Gatchell, D. W.; Vajda, S.; Camacho, C. J. Bioinformatics 2004, 20, 45.
35. Ritchie, D. W. Proteins Struct Funct Genet 2003, 52, 98.
36. Dominguez, C.; Boelens, R.; Bonvin, A. J Am Chem Soc 2003, 125, 1731.
37. Aloy, P.; Querol, E.; Aviles, F. X.; Sternberg, M. J. E. J Mol Biol 2001, 311, 395.
38. Berchanski, A.; Eisenstein, M. Proteins 2003, 53, 817.