
Prediction of structures of multidomain proteins from structures of the individual domains

ANDREW M. WOLLACOTT,^{1,3} ALEXANDRE ZANGHELLINI,^{1,2,3}
PAUL MURPHY,¹ AND DAVID BAKER¹

¹Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA

²Biomolecular Structure and Design, University of Washington, Seattle, Washington 98195, USA

(RECEIVED April 7, 2006; FINAL REVISION October 30, 2006; ACCEPTED October 31, 2006)

Abstract

We describe the development of a method for assembling structures of multidomain proteins from structures of isolated domains. The method consists of an initial low-resolution search in which the conformational space of the domain linker is explored using the Rosetta de novo structure prediction method, followed by a high-resolution search in which all atoms are treated explicitly and backbone and side chain degrees of freedom are simultaneously optimized. The method recapitulates, often with very high accuracy, the structures of existing multidomain proteins.

Keywords: domain assembly; protein–protein docking; protein structure prediction

Proteins are frequently composed of multiple domains (Ponting and Russell 2002; Vogel et al. 2004) that are likely to fold independently (Shen et al. 2005). Determining the structure of multidomain complexes at atomic resolution is critical to understanding the underpinnings of much of biology (Lupas et al. 2001; Aloy and Russell 2006). While structures of single domains can be readily determined through X-ray or NMR techniques, the structures of large multipart proteins are often more difficult to elucidate (Aloy et al. 2003).

There are two general approaches to predicting structures of multidomain proteins from structures of individual domains. First, the domain assembly problem may be treated as a docking problem. For example, Inbar et al. (2005) used rigid body docking methods to predict the structure of the resulting complex. A second approach to domain assembly, which we describe here, is to explicitly

sample the degrees of freedom of the linker rather than the rigid body degrees of freedom of the two domains. Approached in this manner, the domain assembly problem may be viewed as an ab initio prediction problem for a relatively short amino acid sequence with preformed N- and C-terminal structures.

The Rosetta protein modeling method has had success in folding small protein chains ab initio (Bradley et al. 2005), and in protein–protein docking with flexible side chains (Wang et al. 2005). Here we combine these methods to assemble structures of isolated domains into a multidomain complex. The conformation of the linker is explored, keeping the backbone of the individual domains fixed but allowing the side chains in the linker and at the domain interface to sample a full range of rotamer conformations. The lowest energy models found are often very close to the correct structure.

Results and Discussion

Seventy-six two-domain proteins were culled from a nonredundant database of proteins (Berman et al. 2000), as described in Materials and Methods. These proteins contained no cofactors or ligands near the interface of the domains, as the focus was on modeling the interface

³These authors contributed equally to this work.

Reprint requests to: David Baker, Department of Biochemistry Health Sciences Building, Box 357350, Seattle, WA 98195, USA; e-mail: dabaker@u.washington.edu; fax: (206) 685-1792.

Abbreviations: RMSD, root mean square deviation.

Article published online ahead of print. Article and publication date are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.062270707>.

between protein domains only. All systems were first subjected to a low-resolution (side chains represented by centroids) search to generate 5000 candidate decoys, starting from an extended structure of the linker. In this search, the Rosetta de novo fragment assembly method is used to sample the conformational space of the linker; residues in the domains at the N and C termini of the linker interact with each other and with the linker according to the Rosetta low-resolution potential, but no fragment insertions are done within the domains.

The resulting low-resolution models were then subjected to high-resolution refinement using the standard Rosetta Monte Carlo minimization plus side chain repacking protocol (Schueler-Furman et al. 2005). In each attempted move small torsion angle changes are made in the linker and interface side chain conformations are repacked using the Dunbrack backbone-dependent rotamer library (Dunbrack and Cohen 1997) and continuous quasi-Newton optimization (Press et al. 2002) of the linker and side chain degrees of freedom is carried out. The move is accepted or rejected according to the standard Metropolis criterion. Side chain conformations from the native complex were not included in the rotamer search, as this has previously been shown to favor lower RMSD structures (Wang et al. 2005). Using backbone geometries from the native complex may bias the search toward near-native structures; we excluded native side chain information in these studies to reduce this effect.

In most cases, there were only modest changes to the low-resolution structure upon high-resolution refinement, but in some cases, large numbers of clashes generated in the initial side chain grafting caused the domains to separate during refinement and led to large structural changes. Near-native decoys sometimes snapped into a more native-like orientation upon addition and refinement of side chains. Overall, 10.8% of low-resolution decoys with an RMSD <3.0 Å showed a noticeable improvement in RMSD (>0.5 Å difference) after high-resolution refinement. In contrast, only 0.44% of low-resolution decoys with an RMSD >3.0 Å were refined to an RMSD <2.0 Å by the high-resolution refinement protocol; the radius of convergence of the refinement protocol is clearly <3 Å RMSD.

Decoys were ranked based on their interdomain interaction energy to eliminate the effects of energetic differences due to alternative side chain packing away from the interface. For each assembly, the domains were separated, and the energy of each domain was evaluated with side chains in the same conformation as the complex. The interdomain interaction energy was defined as:

$$E_{\text{interaction}} = E_{\text{complex}} - E_{\text{domain1}} - E_{\text{domain2}} - E_{\text{linker}}$$

The domain interaction energy does not include entropic effects associated with complexation, but these are, to a first approximation, independent of the structure of the complex.

Figure 1 shows plots of interdomain interaction energies as a function of RMSD for a number of systems after high-resolution refinement. In many cases, shown in Figure 1A, there is a striking energy funnel with near-native decoys possessing significantly lower energies than higher RMSD models. For comparison purposes, the energies of relaxed native structures (see Materials and Methods) are also plotted. In the majority of cases, the relaxed natives maintained a very low RMSD and had lower interaction energies than the decoys. This further illustrates the deep energy funnel around the native minimum, and suggests that increased sampling could lead to lower RMSD structures.

Predictions of domain structures

We divide our predictions into three major categories: successes, low-resolution failures, and high-resolution failures. Successes are cases where the lowest energy decoys had an RMSD <2 Å after high-resolution refinement. Thirty-eight of the 76 test cases were successful according to this criterion (Table 1). Systems for which no decoys were generated with an RMSD <3 Å after low-resolution refinement were considered failures at the centroid level, representing 13 of the 76 systems (Table 2). The failure to generate a near-native decoy (RMSD <2 Å) (Table 3) or the inability to identify these near-native models from a decoy set (Table 4), were considered failures at the high-resolution level (25 of the systems in this study).

Successful predictions

Table 1 lists those proteins for which the structure of the assembly was accurately predicted, in that one of the five lowest energy models has a C $^{\alpha}$ RMSD <2 Å. In these cases, there were generally a significant fraction of near-native decoys, and the energy function was able to accurately identify the low RMSD models. For a large number of these successful predictions (70%), the lowest energy model was within 1 Å RMSD of the native complex. Plots in Figure 1A illustrate the funnel-like energy distribution for successful predictions, while those in Figure 1B show moderate success cases where a funnel distribution was not evident.

In most cases, the very low RMSD structures had native-like side chain packing at the interface, even though crystal structure side chains were not included. Side chain packing at the interface is illustrated in Figure 2. Even though C $^{\alpha}$ RMSDs for near-native decoys were very low (on the order of 0.2 Å), the heavy atom RMSD was generally over 1.0 Å. As expected, surface side chains away from the interface were generally unconstrained and so did not match well with the native

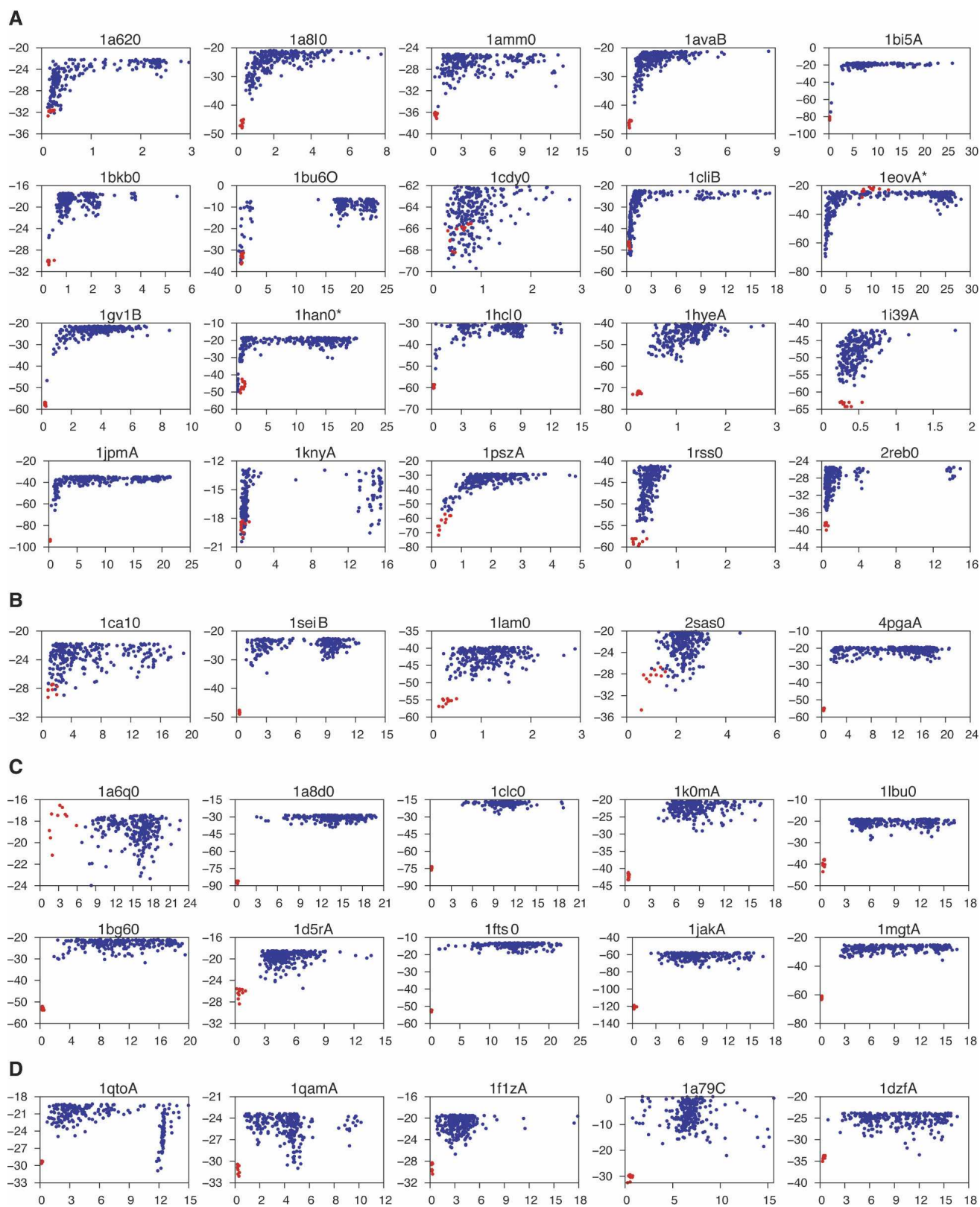


Figure 1. Plots of binding energy (Y-axis) vs. C^α RMSD (X-axis) (Å) for decoys (blue dots), and relaxed natives (red dots) for select complexes. (A) Successes with funnel-like energy distributions. (B) Successes with low-energy and RMSD models but less funnel-like energy distributions. (C) Failures due to insufficient sampling. (D) Failures with low RMSD decoys that were not identified. (*) RMSDs do not include linker regions for these systems.

Table 1. Systems with low-RMSD decoys in the five lowest energy models (after high-resolution refinement)

PDB	Resolution (Å)	N_{res}	N_{linker}	N_{con}^a	$c^b < 3 \text{ \AA}$	$f^c < 2 \text{ \AA}$	Centroid RMSD (Å) ^d	Low RMSD (Å) ^e	Low energy (Å) ^f
1a620	1.55	125	6	24 (34)	61.04	25.54	0.30	0.12 (0.12)	0.26 (0.16)
1a810	1.90	226	8	51 (72)	20.86	4.90	0.60	0.50 (0.50)	0.83 (0.50)
1amm0	1.20	174	10	23 (42)	69.40	7.38	0.68	0.32 (0.32)	0.66 (0.66)
1aao0	2.40	247	12	67 (84)	27.66	6.30	0.55	0.41 (0.41)	0.43 (0.41)
1avaB	1.90	403	10	48 (72)	52.16	12.67	0.69	0.41 (0.41)	0.48 (0.41)
1b63A	1.90	333	8	36 (58)	73.30	20.32	0.44	0.62 (0.62)	1.37 (1.37)
1bag0	2.50	425	18	43 (78)	15.62	0.69	0.54	0.63 (0.87)	1.44 (0.87)
1bi5A	1.56	389	13	115 (144)	0.66	0.10	0.95	0.39 (0.39)	0.39 (0.39)
1bkb0	1.75	136	6	14 (34)	55.40	39.22	0.68	0.29 (0.29)	0.29 (0.29)
1bu6O	2.37	497	8	115 (139)	3.07	0.75	0.28	0.34 (0.34)	0.65 (0.65)
1ca10	1.90	370	13	44 (75)	17.78	1.64	0.51	0.28 (0.28)	3.02 (0.92)
1cdy0	2.00	178	14	22 (63)	51.31	46.04	0.32	0.28 (0.28)	0.90 (0.35)
1cxB	2.40	353	17	71 (106)	3.00	0.04	1.24	1.29 (1.29)	1.29 (1.29)
1cltB	2.50	325	11	63 (84)	13.80	5.99	0.38	0.23 (0.23)	0.49 (0.23)
1ctu0	2.30	294	20	74 (110)	0.74	0.02	0.98	0.95 (0.95)	1.08 (0.95)
1d09B ^g	2.10	153	21	7 (51)	8.16	5.82	0.91	0.69 (0.26)	0.90 (0.42)
1eovA ^g	2.30	487	20	69 (119)	12.31	1.11	0.72	0.71 (0.71)	0.76 (0.71)
1f3aA	1.90	221	9	47 (67)	99.08	21.60	0.35	0.52 (0.52)	0.93 (0.55)
1gcyA	1.60	415	8	42 (69)	8.67	4.33	1.55	0.31 (0.31)	0.31 (0.31)
1gytB	2.50	290	13	54 (84)	14.84	1.06	0.82	0.35 (0.35)	0.35 (0.35)
1han0 ^g	1.90	288	20	79 (116)	8.01	0.02	1.51	1.57 (0.16)	0.21 (0.16)
1hcl0	1.80	294	7	47 (71)	1.46	11.46	0.61	0.36 (0.36)	0.38 (0.36)
1hyeA	1.90	307	14	65 (102)	54.58	20.20	0.48	0.44 (0.44)	1.11 (0.75)
1i39A	1.95	200	20	27 (81)	99.88	51.81	0.38	0.20 (0.20)	0.44 (0.25)
1jpmA	2.25	359	21	62 (120)	3.91	0.80	1.12	0.45 (0.45)	1.06 (0.45)
1jpnA	1.90	296	20	40 (89)	1.76	1.74	0.46	0.92 (0.92)	0.92 (0.92)
1kbwC	2.40	302	11	84 (108)	73.91	8.47	0.47	0.14 (0.14)	0.20 (0.14)
1knyA	2.50	253	5	36 (47)	5.06	4.50	1.21	0.46 (0.46)	0.55 (0.46)
1ks9A	1.70	291	20	17 (81)	69.40	21.20	0.73	0.35 (0.35)	0.55 (0.55)
1lam0	1.60	484	21	36 (93)	98.70	73.62	0.72	0.24 (0.24)	1.55 (0.36)
1pszA	2.00	286	21	62 (100)	85.07	5.82	0.75	0.35 (0.35)	0.95 (0.35)
1rec0	1.90	185	9	29 (53)	42.42	5.10	0.84	0.72 (0.72)	0.88 (0.21)
1rss0	1.90	135	12	32 (67)	80.82	31.22	0.66	0.15 (0.15)	0.33 (0.21)
1setB	1.90	130	20	21 (59)	16.78	2.06	1.20	0.90 (0.98)	3.05 (0.98)
1svpA	2.00	160	11	37 (65)	36.42	6.72	0.57	0.44 (0.44)	0.81 (0.81)
2reb0	2.30	303	8	24 (44)	22.95	14.13	0.37	0.34 (0.34)	0.54 (0.34)
2sas0	2.40	185	12	29 (58)	99.88	32.58	1.04	0.78 (1.00)	1.96 (1.62)
4pgaA	1.70	330	21	29 (74)	2.04	0.13	1.54	1.52 (1.52)	2.38 (1.99)

^aThe number of contacting residues between the two domains, or between the domains and the linker in parentheses (residues with C^β-C^β distances $< 8 \text{ \AA}$).

^bPercentage of decoys with a C^α RMSD $< 3 \text{ \AA}$ (after low-resolution refinement).

^cPercentage of decoys with a C^α RMSD $< 2 \text{ \AA}$ (after high-resolution refinement).

^dLowest RMSD decoy after low-resolution refinement.

^eLowest RMSD decoy after high-resolution refinement. Lowest RMSD decoy in the top 5% of decoys in parentheses.

^fRMSD of the decoy with the lowest energy, and the lowest RMSD of the five lowest scoring decoys in parentheses.

^gRMSD does not include linker atoms.

Table 2. Systems that failed to yield low RMSD decoys during the low-resolution search

PDB	N_{res}	N_{link}	$N_{\text{con}}^{\text{a}}$	$c\% < 3 \text{ \AA}^{\text{b}}$	$c\% < 5 \text{ \AA}^{\text{c}}$
1c2aA	120	20	10 (64)	0.00	0.28
1cx4A	275	14	28 (55)	0.00	0.24
1ev7A	295	21	20 (84)	0.00	0.24
1fmtA	308	20	17 (60)	0.00	0.00
1i8dB	201	20	35 (67)	0.00	0.00
1j8mF	295	20	36 (79)	0.00	0.06
1pii0	452	12	29 (72)	0.00	0.00
1qcsA	195	5	32 (42)	0.00	29.60
1qlaB	239	8	34 (59)	0.00	0.68
1qovL	281	9	24 (48)	0.00	0.04
1nkr0	195	6	26 (41)	0.00	0.00
1rhs0	293	21	60 (113)	0.00	0.02
1tf4B	605	20	36 (82)	0.00	0.00

^aThe number of contacting residues between the two domains, or between the domains and the linker in parentheses (residues with C^{β} - C^{β} distances $< 8 \text{ \AA}$).

^bPercentage of decoys with a C^{α} RMSD $< 3 \text{ \AA}$.

^cPercentage of decoys with a C^{α} RMSD $< 5 \text{ \AA}$.

side chain orientations at those positions. However, at the interface, the side chains more closely resembled native structures, as highlighted in Figure 2, A and B.

Rosetta was also able to correctly model proteins with large linkers, especially when those linkers were constrained and structured. This is shown in Figure 2, C and D; the linker forms a large α -helix in 1i39 and a large β -sheet in 1cdy.

For several proteins in Table 1, the RMSD reported did not take into account the linker residues. These systems, 1d09, 1eov, and 1han, contain large unstructured linkers. In several cases, as shown in Figure 3, A and B, the correct relative positions of the domains was predicted for these systems. However, since the linker was generally unconstrained, there was a large degree of flexibility for these linkers. Thus, the linker was poorly predicted even for the cases where the orientation of the two domains

was correctly identified. These cases were considered successful since the packing of the domains was correctly determined.

Failures

Low-resolution failures

Table 2 lists the proteins for which no decoy was found with a C^{α} RMSD $< 3 \text{ \AA}$ in the low-resolution search. Due to the limited radius of convergence of high-resolution refinement, the low-resolution search has to sample close to the native structure for predictions to be successful. These cases were, therefore, considered failures at the early stage of modeling. Further high-resolution refinement was not carried out on these systems in this study to save computer time, as it is unlikely that refinement would convert failures to successes.

The majority of these failures contained large unstructured linkers. The native structures for two of these systems, 1rhs and 1j8m, are shown in Figure 4. In both cases, the linker wraps around one of the domains. In these systems, insufficient sampling is the likely reason for the inability to generate near-native decoys. With few constraints, and an interface far removed from the endpoints of the linker, only a small fraction of decoys might be expected to sample conformational space near the native state. It is possible that by increasing the number of low-resolution models generated, near-native low energy decoys may be found.

It might be expected that native states that contain a large number of interdomain contacts would be more likely to be recapitulated during low-resolution refinement. However, an analysis of the number of contacts between domains, or between the domains and their linker, shows no strong correlation for systems that yielded near-native decoys and those that failed. This further indicates that the centroid-level failures listed in Table 2 suffered from insufficient sampling and not

Table 3. Systems with no low RMSD decoys in the top 5% by energy after high-resolution refinement

PDB	Resolution (\AA)	N_{res}	N_{link}	$N_{\text{con}}^{\text{a}}$	$c\% < 3 \text{ \AA}^{\text{b}}$	$f\% < 2 \text{ \AA}^{\text{c}}$	Centroid RMSD (\AA) ^d	Low RMSD (\AA) ^e	Low energy (\AA) ^f
1a6q0	2.00	363	10	26 (52)	0.06	0.00	2.24	2.84 (6.75)	8.28 (8.28)
1a8d0	1.57	452	21	36 (85)	0.16	0.00	2.33	2.79 (3.07)	13.76 (12.76)
1clc0	1.90	541	13	52 (88)	0.02	0.00	2.99	3.62 (4.60)	9.77 (7.70)
1crzA	1.95	403	7	37 (58)	0.18	0.00	2.90	2.57 (3.14)	8.68 (8.62)
1f5nA	1.70	570	15	72 (108)	0.18	0.02	1.77	1.75 (10.28)	21.69 (20.23)
1k0mA	1.40	235	21	37 (73)	0.04	0.00	2.43	2.62 (4.04)	9.34 (8.31)
1lbu0	1.80	213	21	35 (83)	0.10	0.00	2.35	2.72 (3.58)	6.13 (6.13)

^aThe number of contacting residues between the two domains, or between the domains and the linker in parentheses (residues with C^{β} - C^{β} distances $< 8 \text{ \AA}$).

^bPercentage of decoys with a C^{α} RMSD $< 3 \text{ \AA}$ (after low-resolution refinement).

^cPercentage of decoys with a C^{α} RMSD $< 2 \text{ \AA}$ (after high-resolution refinement).

^dLowest RMSD decoy after low-resolution refinement.

^eLowest RMSD decoy after high-resolution refinement. Lowest RMSD decoy in the top 5% of decoys in parentheses.

^fRMSD of the decoy with the lowest energy, and the lowest RMSD of the five lowest scoring decoys in parentheses.

Table 4. Systems with no low-RMSD decoys in the five lowest energy decoys after high resolution refinement

PDB	Resolution (Å)	N_{res}	N_{linker}	N_{con}^a	$c\% < 3 \text{ \AA}^b$	$f\% < 2 \text{ \AA}^c$	Centroid RMSD (Å) ^d	Low RMSD (Å) ^e	Low energy (Å) ^f
1a79C	2.28	171	13	7 (42)	1.48	0.52	1.38	0.80 (0.80)	10.67 (3.93)
1bg60	1.80	349	21	40 (97)	0.74	0.04	1.38	1.70 (1.91)	14.14 (2.37)
1cvrA	2.00	432	21	33 (100)	0.04	0.00	2.94	2.68 (2.68)	4.29 (4.29)
1d5rA	2.10	307	9	27 (43)	0.46	0.14	1.73	1.19 (2.51)	6.86 (3.05)
1dzfA	1.90	211	17	17 (54)	1.50	0.08	2.00	0.98 (0.98)	12.00 (10.30)
1egaB	2.40	293	20	29 (78)	2.01	0.06	1.45	1.89 (1.96)	19.79 (14.80)
1eudA	2.10	306	16	50 (100)	2.18	0.00	1.02	2.07 (2.42)	5.55 (4.26)
1flzA	2.40	260	20	24 (58)	23.02	1.74	0.87	0.75 (0.75)	3.04 (2.51)
1fts0	2.20	295	14	57 (96)	0.12	0.04	0.81	1.52 (1.52)	15.45 (8.25)
1jakA	1.75	499	20	59 (138)	0.24	0.02	1.74	1.80 (2.37)	13.49 (5.77)
1jgtB	1.95	500	14	82 (118)	0.22	0.00	1.20	2.02 (2.28)	11.49 (2.28)
1mgtA	1.80	169	21	28 (72)	2.76	0.00	1.56	2.08 (2.50)	14.83 (2.96)
1pgs0	1.80	311	8	69 (90)	0.88	0.00	1.18	2.65 (2.71)	17.14 (9.43)
1qamA	2.20	235	11	20 (40)	24.60	6.42	1.19	0.75 (0.78)	5.06 (4.21)
1qfjC	2.20	226	14	38 (69)	2.42	0.02	1.37	1.74 (1.74)	5.59 (3.71)
1qh4B	1.41	380	20	59 (106)	0.30	0.02	2.35	1.79 (1.79)	12.83 (3.31)
1qtoA	1.50	122	12	29 (44)	11.20	1.70	1.42	0.71 (0.71)	11.90 (11.78)
1smd0	1.60	496	6	49 (63)	0.55	0.12	2.30	1.58 (2.05)	4.91 (4.16)

^aThe number of contacting residues between the two domains, or between the domains and the linker in parentheses (residues with C^{β} - C^{β} distances $< 8 \text{ \AA}$).

^bPercentage of decoys with a C^{α} RMSD $< 3 \text{ \AA}$ (after low-resolution refinement).

^cPercentage of decoys with a C^{α} RMSD $< 2 \text{ \AA}$ (after high-resolution refinement).

^dLowest RMSD decoy after low-resolution refinement.

^eLowest RMSD decoy after high-resolution refinement. Lowest RMSD decoy in the top 5% of decoys in parentheses.

^fRMSD of the decoy with the lowest energy, and the lowest RMSD of the five lowest scoring decoys in parentheses.

deficiencies in the energy function. Indeed, as shown in Table 2, with the exception of 1qcs, these proteins did not have sufficient sampling within 5 \AA of the native structure.

Several systems in Table 2 contained short linkers and yet did not yield low RMSD decoys, such as 1qcs and 1nkr. In these cases, the problem may be that the linker was so short that it became overly restrictive. The bond lengths and angles in the linker are kept fixed during refinement, and with only the torsional angles variable with a short linker near-native complexes may not be conformationally accessible. It is possible, therefore, that increasing the size of these linkers may actually improve the likelihood of generating low RMSD decoys.

In order to test this hypothesis, the linker of 1qcs was extended by one residue on either side so that the linker was seven residues in length instead of five. The low-resolution search was repeated with this new linker, yielding 7% of decoys with an RMSD $< 3 \text{ \AA}$ (data not shown). Using this new definition of the linker, the assembly procedure would be considered a success after low-resolution refinement. This result indicates that if the linker is too small the conformational space available to sample may become too restrictive to obtain near-native decoys, especially when there is a high shape complementarity between the domains. Increasing the size of the linker can allow for better sampling of the conformational space near the native state.

High-resolution failures

Table 3 lists the systems, after high-resolution refinement, for which no decoy $< 3 \text{ \AA}$ was present in the lowest 5% interdomain interaction energy subset of the population. As with the centroid-level failures, these systems are hampered by insufficient sampling. Only a small fraction of near-native low-resolution decoys were created for these complexes; consequently, after high-resolution refinement, there remained very few low RMSD decoys. Only one of these systems (1f5n) had a decoy with an RMSD $< 2 \text{ \AA}$, and that model did not score favorably. Plots in Figure 1C illustrate high-resolution failures that exhibited insufficient sampling. With the exception of 1a6q, the relaxed natives had significantly lower interaction energies, further indicating that this was a sampling problem.

Table 4 lists additional failures after high-resolution refinement in which near-native decoys (C^{α} RMSD $< 2 \text{ \AA}$) were present in the lowest energy 5% of structures but were not the five lowest energy models. Overall, there were only a small fraction of decoys sampled for these systems that were considered near-native. For 1flz, 1a79, and 1dzf, which contain a significant number of lower RMSD decoys, there appears to be a discrete bottleneck to achieving the low energy native minimum as the relaxed native structures are considerably lower in energy than the decoys. Thus, for the majority of these systems, the problem again appears to be insufficient sampling. Notable

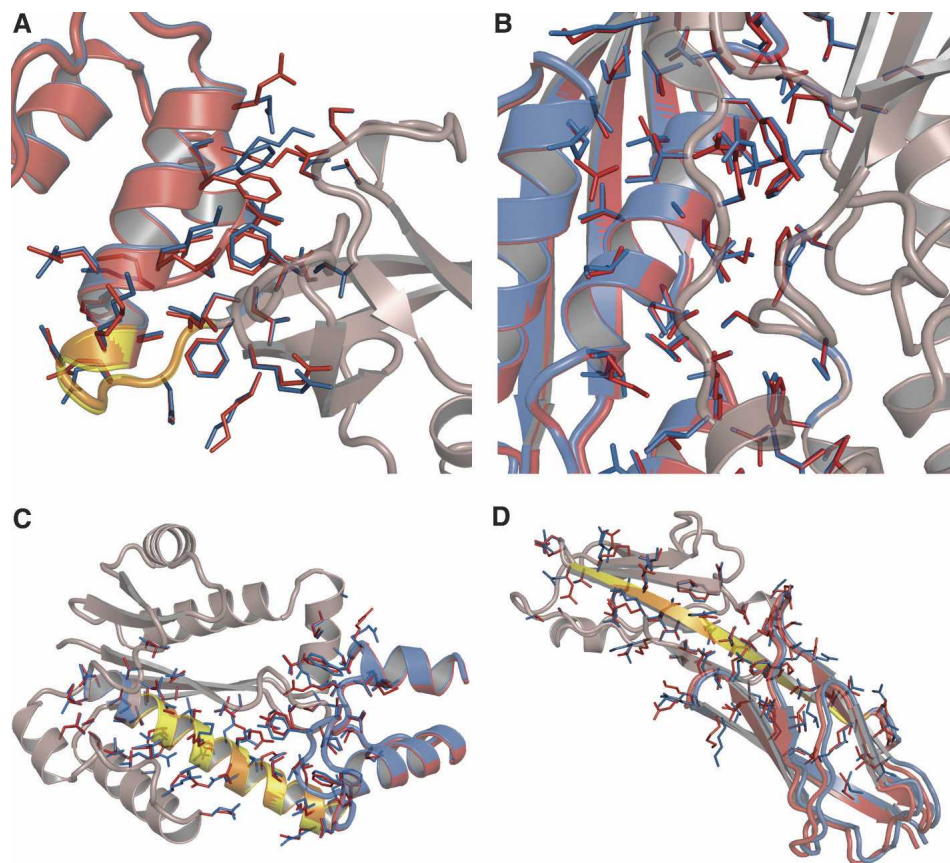


Figure 2. Examples of accurate predictions with domain orientation and side chain packing close to the native structure. (A) 1a62 (C^α RMSD = 0.12 Å; heavy-atom RMSD = 1.05 Å). (B) 1cli (C^α RMSD = 0.23 Å; heavy-atom RMSD = 1.27 Å). (C) 1i39 (C^α RMSD = 0.20 Å; heavy-atom RMSD = 1.20 Å). (D) 1cdy (C^α RMSD = 0.32 Å; heavy-atom RMSD = 1.40 Å). The native structures are in red, the native linker in yellow, the decoy in blue, and the decoy linker in orange. Structures were superimposed onto only one domain.

exceptions in Table 4 are 1qam and 1qto. Here the problem is most likely due to the inability of the energy function to discriminate near-native models from high RMSD decoys (Fig. 1D). As shown in Figure 5A, the low energy 1qto decoy is stabilized by a nonnative strand pairing between the two domains, while maintaining an equivalent number of contacts as in the native structure. For 1qam (Fig. 5B), the lowest energy decoy is stabilized by an increased number of contacts between the two domains. Thus, this remains a scoring problem as the energy function is unable to discriminate native-like complexes from incorrect complexes with larger numbers of contacts. This may require improved measures of packing.

Improvements with high-resolution refinement

In the protocol described here, models are first created using a low-resolution approach where the protein is modeled at the centroid level. This allows for a more

rapid sampling of the conformational degrees of freedom of the linker region. Subsequent high-resolution allows small changes in the linker region to optimize the details of side chain packing at the interface so that the best models can be identified using a physically realistic atomic level energy function. Although the changes are often small, they can dramatically reduce the energy; without backbone refinement, many models contain significant interatomic clashes. While the primary purpose of the high-resolution refinement is to improve recognition of the best models based on the all-atom energy function, in many cases the refinement protocol improves model quality. As shown in Tables 1, 3, and 4, 65% of the systems had a lower RMSD decoy after high-resolution refinement than after low-resolution refinement.

Figure 6 illustrates the different stages in the model generation process. Figure 6A shows the centroid-level energy distribution for 1cli after the low-resolution de novo buildup of the linker. While many near-native models are produced, the scoring function is unable to

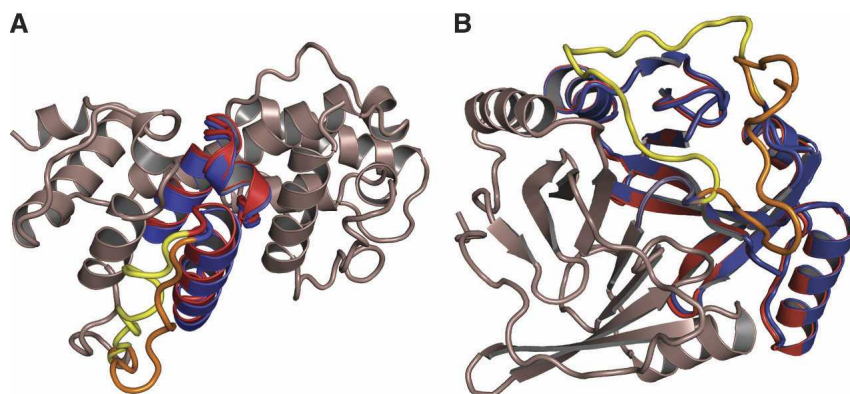


Figure 3. Correct prediction of relative positions of the domains but not the structure of linker. (A) 1aoa (C^α RMSD = 1.04 Å; C^α RMSD neglecting linker = 0.22 Å). (B) 1han (C^α RMSD = 5.07 Å; C^α RMSD neglecting linker = 0.24 Å). The native structures are in red, the native linker in yellow, the decoy in blue, and the decoy linker in orange. Structures were superimposed onto only one domain.

discriminate near-native decoys from structurally divergent models. Figure 6B shows the all-atom energy distribution after side chains are grafted onto low-resolution models and optimally repacked, keeping the linker fixed. This leads to an abundance of models with large numbers of atomic clashes, as the relative orientations of the domains cannot accommodate the native sequence. By allowing the linker region to relax, a more dramatic energy funnel is obtained, as shown in Figure 6C, allowing for the identification of near-native decoys using the scoring metric. Figure 6D summarizes the changes in structure that occur upon high-resolution refinement. In the majority of cases, the structures diverge from the native model due to clashes introduced by the side chain grafting, but a subset of the lower RMSD structures become more native-like.

Overall, the Rosetta domain assembly protocol appears to be quite successful at predicting the structure of two-

domain complexes, and the methodology can be readily extended to multidomain assemblies.

Conclusion

The Rosetta domain assembly method is successful in predicting and identifying near-native complexes for domain assembly problems in 50% of cases studied here. By explicitly modeling the polypeptide linker that tethers both domains, the conformational space available for the docking of each domain is reduced, and thus treating domain assembly as a linker folding problem can be more powerful than restricted docking methods. Most of the failures with this method appear to be due to insufficient sampling; with increased computational resources, the success rate should increase considerably. The high accuracy of many of the lowest energy models (Fig. 2) illustrate recent progress in high-resolution modeling

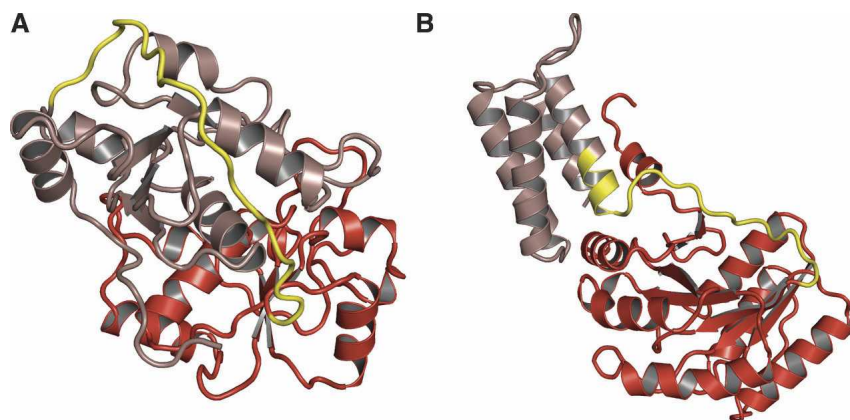


Figure 4. Challenging complexes where sampling was a problem, with long linkers stretching around a domain. Native structures for 1rhs (A) and 1j8m (B).

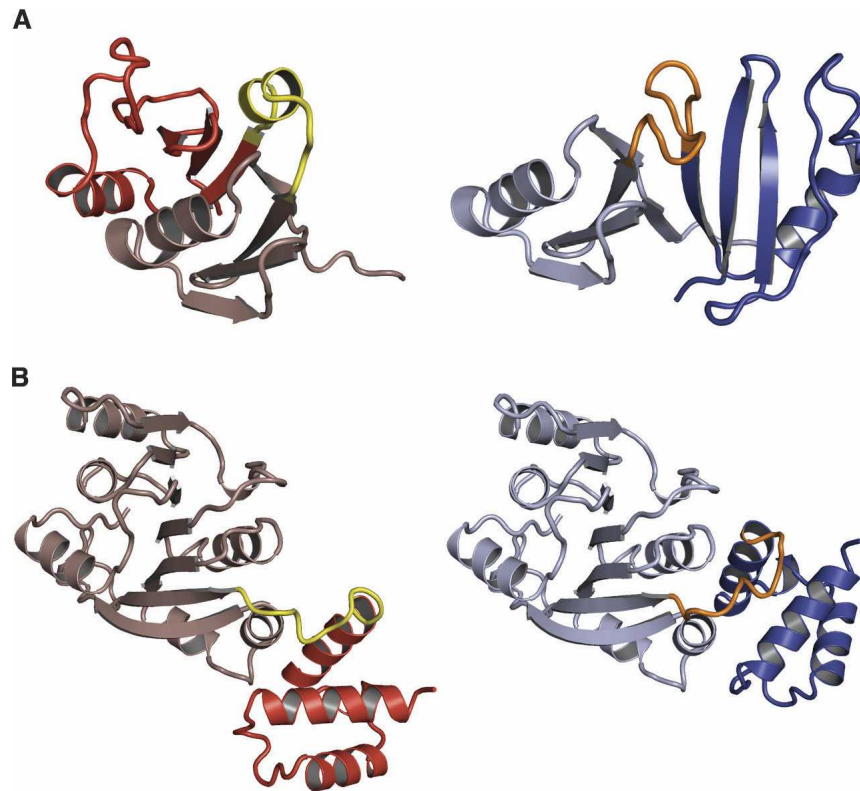


Figure 5. Low-energy high RMSD decoys for several complexes. (A) 1qto (C^α RMSD = 11.90 Å) and (B) 1qam (C^α RMSD = 5.06 Å). The native structures are on the left and decoys on the right.

(Schueler-Furman et al. 2005). To better treat problems in which the domain structures are produced by homology modeling, the next step would be to incorporate domain flexibility, particularly in loops, into the assembly protocol.

Materials and methods

Data set and linker definition

A nonredundant subset of the Protein Data Bank (Berman et al. 2000) was used to build the data set used in this study. The initial data set was obtained through the PISCES server (Wang and Dunbrack 2003) with the following parameters: the cutoff for redundancy at the sequence level was set at <40% sequence identity, and X-ray structures with a resolution of ≤ 2.5 Å were chosen. This culling reduced the data set to 601 structures.

An automatic domain parsing and linker definition procedure was implemented. Three independent domain prediction algorithms were used, Dali (Holm and Sander 1998), CATH (Orengo et al. 1997), and Taylor's (Taylor 1999). The sequence residues that define the linker between the two domains were determined for each method, and the consensus of the three methods was chosen as the linker region. A secondary structure assignment was determined for each protein (Kabsch and Sander 1983),

and the linker region was extended on both sides up to the boundaries of contiguous secondary structure segments. The linker size was limited to 21 residues. This method allows for a systematic and automatic definition of linker regions.

Proteins in our data set were further filtered to consider only two-domain proteins that were not part of oligomeric complexes. Also, structures containing ligand groups or metal cofactors near the interface of the two domains were removed from the benchmark set. This yielded a total of 76 structures, listed in Tables 1–4.

Low-resolution domain assembly

A low-resolution search was performed for each system using a centroid representation of the protein (backbone and C^β atoms only) in a manner similar to the Rosetta ab initio folding method (Bradley et al. 2005). The linker between each domain was initially set in an extended conformation ($\phi = -150^\circ$, $\psi = 150^\circ$, $\omega = 180^\circ$), with ideal bond lengths and angles. The Rosetta de novo fragment assembly protocol was then used to build 5000 alternative conformations of the linker (Rohl et al. 2004) while the two flanking domains were held internally rigid. The centroid-level energy function used to guide the Monte Carlo sampling through alternative conformations favors burial of hydrophobic residues, pairing of β -strands, and penalizes clashes between backbone atoms and the residue centroids.

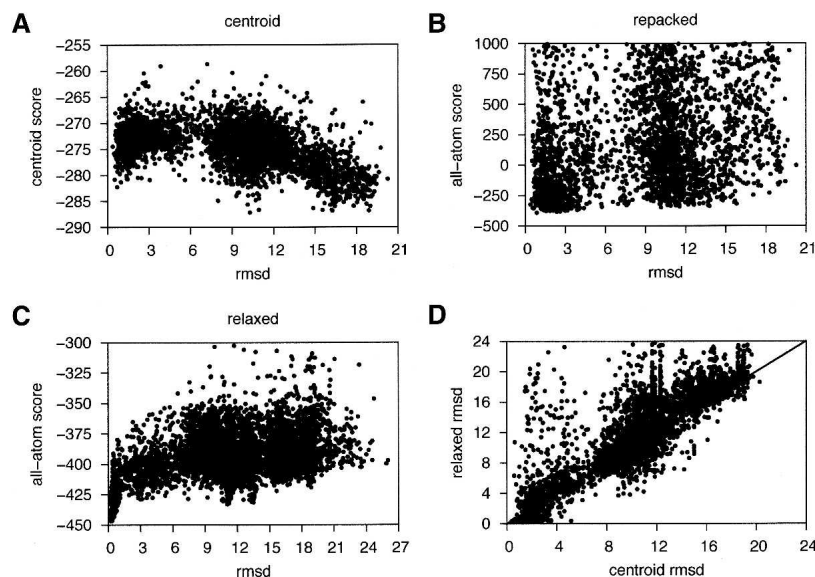


Figure 6. Energy distributions at successive steps in the prediction protocol. (A–C) Plots of energy vs. RMSD to native for a population of models generated for Icli (A) after de novo modeling of the linker using the centroid-based energy function (B) after grafting and repacking interface side chains onto models in A and C after high-resolution refinement of models in B. (D) Comparison of model RMSDs to native before and after high-resolution refinement. For this system, high-resolution refinement significantly increased the extent that the best models could be recognized based on their energies (cf. panels A and B to panel C), and improved the quality of many of the models (D).

The C^α RMSD to the native structure was calculated for each conformation (decoy) generated in this de novo buildup process, and if any decoys were found with an RMSD <3 Å, then the high-resolution refinement protocol was applied to all decoy models for that protein. Otherwise, the system was considered a failure at the low-resolution level, since it is unlikely that low RMSD models would be produced by high-resolution refinement given the limited radius of convergence. The centroid score (Rohl et al. 2004), evaluated for the entire structure, could not reliably identify low RMSD domain assemblies, so it was not used to reduce the number of models that were subjected to high-resolution refinement.

High-resolution refinement

Side chains were grafted onto the centroid-level models using the following protocol. First, for each system, the individual domains in the native structure were separated and the side chains repacked using a Monte Carlo sampling protocol (Kuhlman and Baker 2000) and the Dunbrack backbone-dependent rotamer library (native side chain conformations were not used as they could unfairly bias refinements toward the native structure). The repacked side chain orientations were then grafted onto each of the centroid-level decoys, and the residues near the domain–domain interface (with C^β – C^β distances <8 Å across the interface) were then repacked in each decoy. The rationale behind this approach was to reduce energy differences due to rotamer packing differences far from the domain interface.

Following side chain grafting and repacking, the linker was relaxed using the Monte Carlo Minimization (MCM) protocol developed for high-resolution refinement of protein models in Rosetta (Bradley et al. 2005; Misura and Baker 2005). Each step

in the MCM protocol consists first of small random perturbations that are applied to the backbone ϕ and ψ dihedral angles at random positions in the linker. The side chains in the linker as well as at the interface are then optimized using the greedy “rotamer-trials” algorithm (Rohl et al. 2004), and subsequently the backbone degrees of freedom of the linker and the side chain chi angles of all residues were minimized using a quasi-Newton algorithm. The move is then accepted or rejected based on the standard Metropolis criterion. Full combinatorial side chain repacking of the linker and interface between domains was carried out after every 25 MCM cycles. The repulsive Lennard-Jones term was initially damped and then ramped up over the first 10 cycles to more smoothly transition from the centroid to all-atom representations of the protein chain.

The interdomain interaction energy was computed for each decoy by subtracting the energy of individual domains and the linker from the energy of the assembly. For comparison purposes, the interdomain interaction energies of the X-ray structure and relaxed natives were also calculated. Relaxed native structures in this context were generated from a centroid model of the complex that contained the linker in the same orientation as the native state. Starting with this model would be equivalent to obtaining a decoy of near 0 Å RMSD after a low-resolution search.

In several cases, the domains in the relaxed natives appear to drift apart, leading to large deviations from the X-ray structure. The Ieov system demonstrates this effect as shown in Figure 1, where the decoys exhibit a funnel shape while the relaxed natives have a high RMSD. The reason for the large deviations in the near-native structures is that the linkers were idealized before relaxation; bond lengths and angles replaced by ideal values. For tight complexes, idealization of only the linker can lead to small backbone clashes that can cause the complex to drift apart upon relaxation.

Acknowledgments

The authors thank Keith Laidig for maintaining the computational resources used in this work. We also acknowledge David Kim and Carol Rohl for the initial culling and analysis of the data set. This work was supported by the HHMI.

References

- Aloy, P. and Russell, R.B. 2006. Structural systems biology: Modelling protein interactions. *Nat. Rev. Mol. Cell Biol.* **7**: 188–197.
- Aloy, P., Ceulemans, H., Stark, A., and Russell, R.B. 2003. The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.* **332**: 989–998.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28**: 235–242.
- Bradley, P., Misura, K.M., and Baker, D. 2005. Toward high-resolution de novo structure prediction for small proteins. *Science* **309**: 1868–1871.
- Dunbrack Jr., R.L. and Cohen, F.E. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**: 1661–1681.
- Holm, L. and Sander, C. 1998. Dictionary of recurrent domains in protein structures. *Proteins* **33**: 88–96.
- Inbar, Y., Benyamini, H., Nussinov, R., and Wolfson, H.J. 2005. Combinatorial docking approach for structure prediction of large proteins and multi-molecular assemblies. *Phys. Biol.* **2**: S156–S165.
- Kabsch, W. and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**: 2577–2637.
- Kuhlman, B. and Baker, D. 2000. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci.* **97**: 10383–10388.
- Lupas, A.N., Ponting, C.P., and Russell, R.B. 2001. On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* **134**: 191–203.
- Misura, K.M. and Baker, D. 2005. Progress and challenges in high-resolution refinement of protein structure models. *Proteins* **59**: 15–29.
- Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH—A hierarchic classification of protein domain structures. *Structure* **5**: 1093–1108.
- Ponting, C.P. and Russell, R.R. 2002. The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.* **31**: 45–71.
- Press, W., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. 2002. *Numerical recipes in C++*, pp. 424–430. Cambridge University Press, Cambridge, MA.
- Rohl, C.A., Strauss, C.E., Misura, K.M., and Baker, D. 2004. Protein structure prediction using Rosetta. *Methods Enzymol.* **383**: 66–93.
- Schueler-Furman, O., Wang, C., Bradley, P., Misura, K., and Baker, D. 2005. Progress in modeling of protein structures and interactions. *Science* **310**: 638–642.
- Shen, M.Y., Davis, F.P., and Salí, A. 2005. The optimal size of a globular protein domain: A simple sphere-packing model. *Chem. Phys. Lett.* **405**: 224–228.
- Taylor, W.R. 1999. Protein structural domain identification. *Protein Eng.* **12**: 203–216.
- Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C., and Teichmann, S.A. 2004. Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.* **14**: 208–216.
- Wang, C., Schueler-Furman, O., and Baker, D. 2005. Improved side-chain modeling for protein–protein docking. *Protein Sci.* **14**: 1328–1339.
- Wang, G. and Dunbrack Jr., R.L. 2003. PISCES: A protein sequence culling server. *Bioinformatics* **19**: 1589–1591.