Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Structural Biology 150 (2005) 233–244
Journal of Structural Biology
www.elsevier.com/locate/yjsbi

Predicting oligomeric assemblies: *N*-mers a primer

Stephen R. Comeau [a], Carlos J. Camacho [b,*]

[a] *Bioinformatics Graduate Program, Boston University, 44 Cummington St., Boston, MA 02215, USA*
[b] *Department of Computational Biology, University of Pittsburgh, 200 Lothrop St-W1041, Pittsburgh, PA 15261, USA*

Received 23 December 2004, and in revised form 11 March 2005
Available online 6 April 2005

Abstract

Multi-protein complexes play key roles in many biological processes. However, since the structures of these assemblies are hard to resolve experimentally, the detailed mechanism of how they work cooperatively in the cell has remained elusive. Similarly, recent advances on in silico prediction of protein–protein interactions have so far avoided this difficult problem. In this paper, we present a general algorithm to predict molecular assemblies of homo-oligomers. Given the number of *N*-mers and the 3D structure of one monomer, the method samples all the possible symmetries that *N*-mers can be assembled. Based on a scoring function that clusters the low free energy structures at each binding interface, the algorithm predicts the complex structure as well as the symmetry of the protein assembly. The method is quite general and does not involve any free parameters. The algorithm has been implemented as a public server and integrated to the protein–protein complex prediction server *ClusPro*. Using this application, we validated predictions for trimers, tetramers (discriminating between dimer of dimers and 4-fold symmetry structures), pentamers and hexamers (discriminating between trimer of dimers, dimer of trimers, and 6-fold symmetry structures), for a total of 107 assemblies. For 85% of the multimers, the server predicts the complex structure within an average rms deviation of 2 Å from the full crystal. For complexes that involve more than one binding interface, the cluster size at each surface provides a strong indication as to which interface forms first. With improving scoring functions and computer power, our multimer docking approach could be used as a framework to address the more general problem of multi-protein assemblies.

Keywords: Macromolecules; Supramolecular assembly; Docking; Complex prediction; Oligomers

1. Introduction

Multi-protein assemblies play a key role in some of the most complex biological processes like signal transduction, transcriptional regulation, motor proteins, and viral assembly. At the same time, the detailed mechanism of how these assemblies are formed and function is not well understood. The main problem is that it is difficult to make crystals out of protein complexes. Similarly, NMR techniques are limited to small protein, and the resolution of cryo-EM is not yet at an atomic level (Frank, 2002).

Despite the recent advances of traditional protein–protein docking algorithms to predict complex structures (Camacho and Vajda, 2002; Halperin et al., 2002; Smith and Sternberg, 2002), the prediction of multi-interacting complexes has received little or no attention. Indeed, so far the algorithms have been limited to take the three-dimensional coordinates of two proteins and derive a model for the co-crystallized structure. Docking methods will often sample billions of putative complexes, scanning the rotational and translational space between the two proteins (Katchalski-Katzir et al., 1992; Ritchie and Kemp, 2000). Subsequently, these complexes are subjected to various filtering (Camacho et al., 2000; Gabb et al., 1997; Norel et al., 2001; Weng et al., 1996) to eliminate false positives and discriminate the near-native structure. To

* Corresponding author.
E-mail address: ccamacho@pitt.edu (C.J. Camacho).

1047-8477/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2005.03.006

evaluate the current status of the protein–protein docking field, the critical assessment of prediction of interaction (CAPRI) experiment (Janin et al., 2003) came to life. Researchers are given the receptor and the ligand three-dimensional coordinates before the co-crystallized complex is published, and in three to four weeks they are supposed to predict the complex. In just a couple of years, this initiative has led to significant advances in the field, as well as to the independent validation of the different techniques. In the first evaluation meeting (Mendez et al., 2003), Camacho and Gatchell (2003) produced some of the best model structures, appropriately distinguishing between near-native and false positive structures. Based on these promising results, we implemented a rigid body docking algorithm as an automated public server named *ClusPro* (Comeau et al., 2004a); http://structure.pitt.edu. The automated (no human intervention) predictions of *ClusPro* have been further validated in the second round of CAPRI, where out of 10 targets the server predicted good models for 5 of the complex structures (see http://capri.ebi.ac.uk; also, Wodak and Mendez, 2004; and Comeau et al., 2004b).

The main problem predicting multi-protein assemblies is that scoring functions have not been validated for comparing two different interfaces at the same time. In a seminal article, Berchanski and Eisenstein (2003) had a first attempt to this problem by developing a method to predict tetramers with known dimer of dimers (D_2) symmetry. These authors find dimers by constraining the rotational and translational space such that only symmetric dimers are sampled. Then, they developed two different algorithms for generating the second dimer. In the first one, they combine two dimers that have one monomer common between them, yielding a total of three monomers and use geometric operations to generate the last subunit of the complex structure. In the second one, they dock the first dimer against itself to generate the complete tetramer. Applications of this approach are rather limited, since the method requires the final complex to be a D_2 tetramer. A more realistic application requires the prediction of the complex structure without knowing the symmetry. For instance, a gel electrophoresis experiment can easily detect the formation of *N*-mers (see, e.g., Gan et al., 2004), then a natural question to ask is: "what is the structure and symmetry of the *N*-mer assembly?"

In this paper, we present a computational method that, given a protein that forms an *N*-mer, predicts the full structure and symmetry of the protein assembly. The method is validated in 107 homo-oligomer complexes, including dimers, trimers, tetramers, pentamers, and hexamers. Without using any free parameter or adapting our previously developed scoring function for protein–protein docking to any complex structure, the method predicts 88 of the predicted protein assemblies within about 2 Å rms deviation (RMSD) from the crystal structure. For another 14 complexes, a native-like complex is predicted within the best 7 predictions. The method is also implemented as part of our server *ClusPro*, allowing the research community immediate access to this novel technology.

2. Methods

2.1. Validation data set

We validated the method in a set of 107 homo-oligomer structures taken from the Protein Data Bank (PDB; Berman et al., 2000) and listed in Table 1. The set included the group of D_2 tetramers tested by Berchanski and Eisenstein (2003), target 10 from the Critical Assessment of Protein Interactions (CAPRI) experiment, and 90 multimers drawn randomly from the PDB. In total, we consider 40 trimeric structures, 33 tetrameric structures, 17 pentameric structures, and 17 hexameric structures. The set of structures were extracted from the PDB based on a simple search for multimeric complexes (no PDB was eliminated based on poor performance). We also run two hexamers where the structure of the complex is not known, as well as tested the prediction of heptamers for three cases. We found relatively few heptamers, around 10, so we did not include these results in the validation set. The computational cost of predicting the multiple symmetries of octamers and larger assemblies was too high to be implemented on the server.

2.2. Flow diagram of the algorithm

Fig. 1 shows a sketch of the algorithm, the steps are as follows:

1. We predict N_{dock} ($\geqslant$20,000) rigid-body docked conformations between two identical structures using the program DOT (Mandell et al., 2001; TenEyck et al., 1995).
2. We filtered the best 500 desolvation and 1500 electrostatic complexes in N_{dock}, and score each of the N_{dock} complexes according to the number of energetically favorable structural neighbors in the set of 2000 filtered structures. This is the same scoring function used for traditional receptor–ligand docking implemented in the server *ClusPro*. It is important to note that none of its original parameters have been modified to dock identical proteins (Camacho et al., 2000; Camacho and Gatchell, 2003; Comeau et al., 2004a). The electrostatics energy is estimated by a Coulombic term with a distance dependent dielectric of $4r$. Desolvation of polar and non-polar groups, including side chain entropy loss, is estimated using an atomic contact potential (Zhang et al., 1997). An implementa-

Table 1
List of proteins in our validation set

PDB code	Protein name
Trimers	
1A3D	Phospholipase A2 from *Naja Naja* Venom
1ALY	Crystal structure of human Cd40 ligand
1BUU	One Ho3+ form of rat mannose-binding protein A
1DBF	Chorismate mutase from *Bacillus subtilis*
1DG6	Crystal structure of Apo2L/Trail
1DUN	Eiav D-utpase native
1E12	Halorhodopsin, a light-driven chloride pump
1EK9	outer membrane protein tolc
1EP9	Human ornithine transcarbamylase
1EUA	Schiff base intermediate in Kdpg aldolase from *E. coli*
1F7L	Holo-acyl protein synthase in complex with coenzyme
1FIM	Macrophage migration inhibitory factor
1FXZ	Soybean proglycinin A1Ab1B homotrimer
1G2O	Purine nucleoside phosphorylase in complex with Inh.
1GR3	Human collagen X Nc1 trimer
1IHC	X-ray structure of gephyrin N-terminal domain
1JS0	Crystal structure of 3D domain-swapped Rnase
1JY8	2C-Methyl-D-erythritol 2,4-cyclodiphosphate synthase
1KCB	Nitrate reductase mutant
1KHT	Adenylate kinase from *Methanococcus voltae*
1KKE	Reovirus attachment protein σ1 trimer
1KRR	Galactoside acetyltransferase in complex with acetyl-CoA
1MIF	Macrophage migration inhibitory factor
1MPM	Maltoporin maltose complex
1N41	Annexin V K27E mutant
1O91	Collagen Viii Nc1 Domain Trimer
1OK8	Postfusion dengue 2 virus envelope glycoprotein
1P32	Human P32, acidic mitochondrial matrix protein
1PLQ	Eukaryotic DNA polymerase processivity factor PCNA
1Q5H	Human Dutp pyrophosphatase complex with Dudp
1QKP	Early intermediate in the bacteriorhodopsin photocycle
1QRF	Carbonic anhydrase
1QY7	Pii protein
1TNF	Tumor necrosis factor-α
1UFY	Chorisate mutase from *Thermus themophilus*
1V6H	Divalent cation tolerance protein cuta1
2PHL	Phaseolin
2XAT	Xenobiotic acetyltransferase
4PNP	Purine nucleoside phosphorylase
C_4 Tetramers	
1A68	Potassium channel Kv1.1
1AK5	Inosine-5′-monophosphate dehydrogenase
1CUK	Ruva protein
1EZJ	Nucleocapsid phosphoprotein
1FQY	Aquaporin-1
1FUA	L-Fuculose-1-phosphate aldolase
1FX8	Glycerol uptake facilitator protein
1HUV	L(+)-Mandelate dehydrogenase
1KBJ	Cytochrome B2 Fmn-binding domain
1L7F	Neuraminidase
1N93	P40 Nucleoprotein
1N9P	G Protein-activated inward rectifier potassium channel 1
1NC7	Hypothetical protein TM1070
1NSC	Neuraminidase
1P7B	Integral membrane channel and cytosolic domains
1PQH	Aspartate 1-decarboxylase
1PVN	Inosine-5′-monophosphate dehydrogenase
Pentamers	
1B0C	Pancreatic trypsin inhibitor
1BZ5	Pancreatic trypsin inhibitor
1C4Q	Shiga-like toxin I subunit B
1DJR	Heat-labile enterotoxin
1FB1	GTP cychlohydrolase I
1JG5	GTP cychlohydrolase I feedback regulatory protein
1JZN	Galactose-specific lectin
1KZ1	6,7-dimethyl-8-ribityllumazine synthase
1LTA	Heat-labile enterotoxin (Lt) complex with galactose
1MSL	Mechanosensitive ion channel
1NLQ	Nucleoplasmin-like protein
1QB5	Heat-labile enterotoxin type Iib B-pentamer
1SAC	Serum amyloid P component
1VPS	Polyomavirus Pv1 pentamer
1XTC	Cholera toxin
2BOS	Shiga-like toxin Iie B subunit
3CHB	Cholera toxin
C_6 Hexamer	
1G41	Heat shock protein Hslu
1G6O	Traffic ATPase
1LJO	Archaeal Sm-like protein Af-Sm2
1NSF	*N*-Ethylmaleimide sensitive factor
1RRE	ATP-dependent protease
D_2 Tetramer	
1A2Z	Pyrrolidone carboxyl peptidase
1A7K	Glyceraldehyde-3-phosphate dehydrogenase
1ADO	Fructose 1,6-bisphosphate aldolase
1BK4	Fructose-1,6-bisphosphatase
1BQ4	Phosphoglycerate mutase 1
1BVQ	4-Hydroxybenzoyl CoA thioesterase
1DGE	Dialkylglycine decarboxylase Q15H mutant
1ENP	Enoyl acyl carrier protein reductase
1FTR	Tetrahydromethanopterin formyltransferase
1GIC	Concanavalin
1HDC	3-α, 20-β-hydroxysteroid dehydrogenase
1RHP	Platelet factor 4
1XVA	Glycine *N*-methyltransferase
1YKF	NADP-dependent alcohol dehydrogenase
4ECA	L-Asparagine amidohydrolase
6PFK	Phosphofructokinase
DC_3 Hexamers	
1BE4	Nucleoside diphosphate kinase isoform B
1EHW	Nucleoside diphosphate kinase
1EKR	Molybdenum cofactor biosynthesis protein C
1I40	Inorganic pyrophosphatase
1JX7	Hypothetical protein Ychn
1LCP	Leucine aminopeptidase
1PJH	Enoyl-CoA isomerase
C_3D Hexamers	
1J2T	Creatinine amidohydrolase
1JE0	5′-Methylthioadenosine phosphorylase
1NNG	Putative acyl-CoA thioester hydrolase
1PJC	L-Alanine dehyrogenase
1R6L	Ribonuclease pH
Heptamers	
1H64	*Pyrococcus abyssi* Sm core
1LOJ	Methanobacterial Sm-like archaeal protein
1WNR	Cpn10

tion of the free energy scoring function can be downloaded from http://structure.pitt.edu (Camacho and Zhang, 2005).

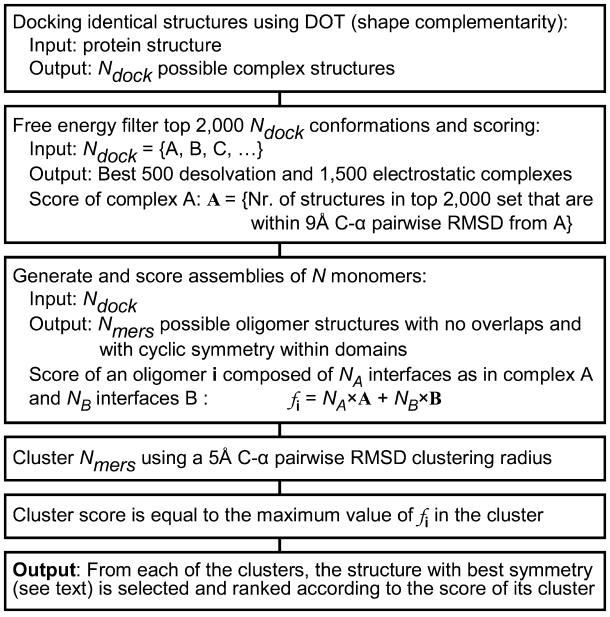

Fig. 1. Flowchart of the algorithm to predict oligomeric assemblies.

3. Depending on the number of monomers (N) forming the complex, we apply the symmetry relationships and build all possible N_{mers} for each docked conformation. N_{mers} have to satisfy two conditions: (a) no significant overlap; and, (b) cyclic symmetry within domains and sub-domains (see below). Based on the score of each docked conformation in Step 2, the score of an N-mer is defined as the sum of the scores of the interfaces forming the oligomer (see Fig. 3).
4. All N_{mers} are clustered using a 5 Å clustering radius.
5. Each of the clusters is assigned a score corresponding to the maximum score obtained by any of its cluster members.
6. One N-mer is predicted for each cluster and ranked according to the score of the cluster. The predicted N-mer is selected based on the symmetry considerations. For C_N complexes, the symmetry is gauged by the RMSD of the first and $(N+1)$th structure, with lower RMSD corresponding to better symmetry. For complexes with more than one interface, we select the individual dimers and trimers forming the assembly based upon the same criterion as all C_N complexes, and then we assess the full symmetry by checking that the distances between the different domains are within a given threshold (see below for details).

2.3. Building N-meric complexes from the set of screened structures

Fig. 2 shows an example of the types of hexameric complexes built using a monomer from PDB code 1G41. The algorithm used to sample the different structural symmetries is as follows:

(a) *N-fold symmetry (C_N)*. Given the value N, that is the number of monomers constituting the final multimer, we filter structures that fit the C_N condition. Namely, for each of the previously screened N_{dock} putative complexes, a third structure is generated by calculating the rotation and translational matrix (M_A^R) needed to move the original structure into the frame of the docked ligand, and then applying $[M_A^R]$ to the docked ligand. This procedure continues until $N-1$ new structures are generated. In total, we have $N+1$ structures: the receptor, the docked ligand, and $N-1$ symmetry generated structures. The complex would correspond to a perfectly symmetric C_N structure if the original receptor and the $(N+1)$th structure superimpose, i.e., their RMSD is zero. In practice, if the RMSD falls below a pre-defined threshold of 8 Å and there are no significant structural overlaps (<3000 Å^3) between the N monomers, we claim that we have built a C_N complex.

(b) *Dimer of dimers (D_2) and dimer of trimers (DC_3)*. If N is not a prime number, then an N-mer can form a complex with symmetry other than C_N. This is the case for tetramers and hexamers. Tetramers can form C_4 or D_2 structures. Hexamers can form C_6, DC_3, or C_3D [see (c) below] structures. Some of these cases are more challenging to predict because contrary to C_N structures that involve only N repeats of one identical interface, other symmetries involve two distinct binding modes, both of which need to be detected in the initial docking step. To build complexes with D_2 symmetry, we use the procedure described in (a) to identify the set of C_2 dimers in the main set of screened complexes. Then, for every pair of dimers, we apply the corresponding rotation and translation transformation that brings one dimer into the frame of reference of the other (this aspect of the method is similar to Berchanski and Eisenstein, 2003) The same procedure is used to build DC_3 complexes, though here we couple all possible C_3 trimers with all dimers. If there are no significant overlaps between the structures, and the distances between the monomers in the first dimer (trimer) and their corresponding monomer in their symmetric dimer (trimer) are between 20% of each other, we claim that we have built a D_2 (DC_3) complex.

(c)*Trimer of dimers (C_3D)*. A slightly different approach is taken to generate structures with C_3D symmetry. Here, only dimers are needed to generate the final hexameric structure. For each predicted dimer of say protein I and II, we build a second dimer between protein II and III. We compute the rotation/translation matrix M_E^R that takes I into II and M_D^R that takes II into III. Then, we apply $[M_E^R]$ to III to build IV, $[M_D^R]$ to IV to build V, $[M_E^R]$ to V to build VI, and $[M_D^R]$ to VI to build I′. If the RMSD between I and I′ falls below the 8Å threshold and there are no significant

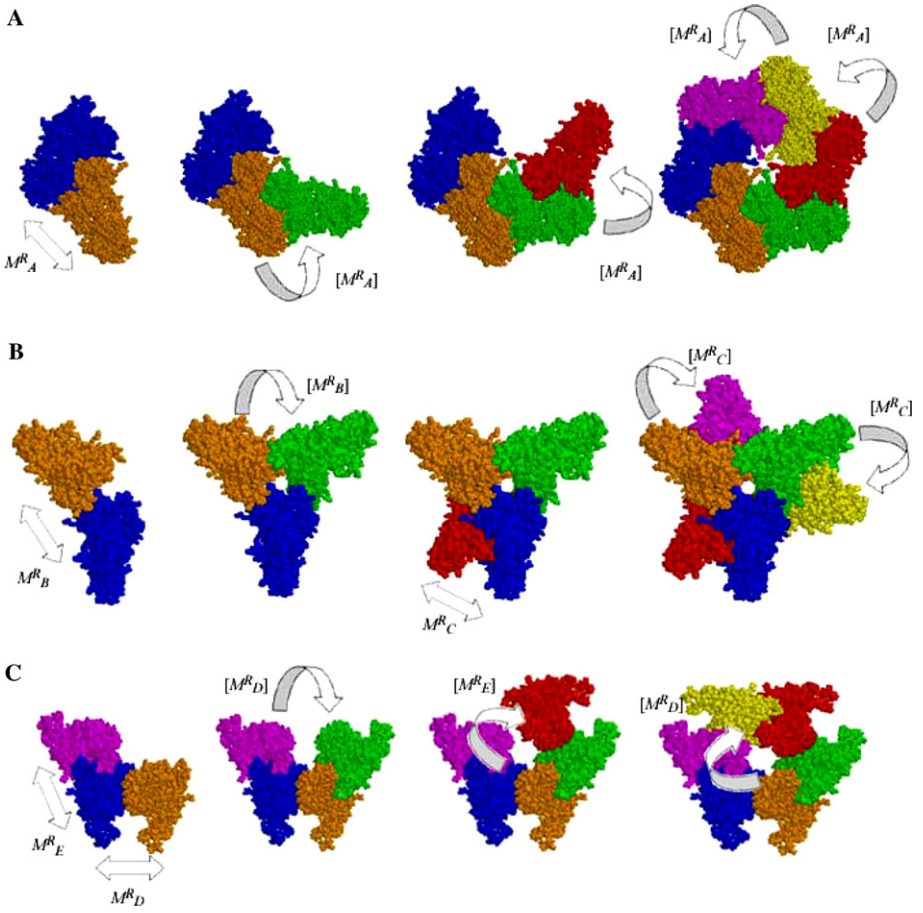

Fig. 2. Prediction of hexameric assemblies of the monomer forming complex 1G41. Three types of symmetries were observed by our method: (A) C_6, (B) DC_3, and (C) C_3D. For each case, we indicate the interface used to compute the rotational/translational matrix as M^R, and show how the matrices are applied to add the missing monomers [M^R].

structural overlaps between the N monomers, we claim that we have built a trimer of dimers complex.

2.4. Selecting the right symmetry of the assembly

For *N*-mers with multiple interfaces, we compute for each interface the largest cluster size of structures found in the set of 2000 filtered structures, and then we add these cluster sizes to obtain the total score for the given assembly. For instance, in Fig. 3 we show the different types of symmetries detected for the hexamer 1G41 (C_6, DC_3, and C_3D), and the total score computed for each complex. We note that this scoring function underestimate some complexes that exhibit more contacts

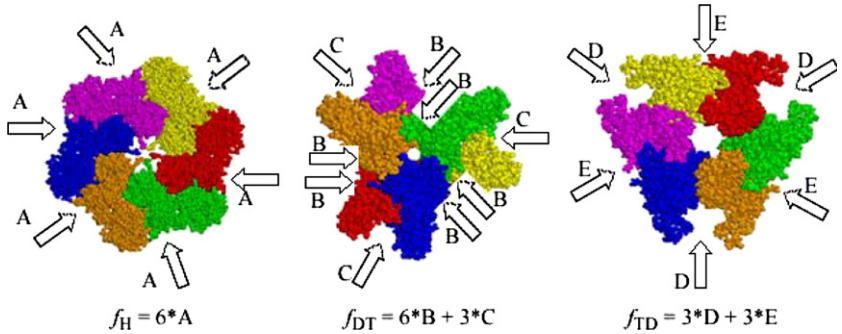

Fig. 3. Scoring function for assemblies with different symmetries. Each interface receives a score equal to the maximum number of low energy conformations at that interface. The total score adds up all the interfaces of the assembly. The score of a 6-fold symmetric structure is equal to 6 times the number of low energy conformations on the interface between two monomers. The score for dimer of trimers have three interfaces (B) on each trimer plus three dimer interfaces (C) between the two trimers for a total score of 6B + 3C. Finally, the trimer of dimers involve 6 interfaces, 3D + 3E.

than those described by the pairwise clusters of low energy conformations (e.g., the case of 1PJC described in Section 3).

2.5. Server output

The above methodology has been fully implemented by the *N*-mer prediction tool of the server *ClusPro*. The final predictions are CHARMm (Brooks et al., 1983) minimized for 300 steps, with an unconstrained backbone to eliminate some of the more significant clashes. The results are then placed on the web, and the user is emailed a link to which they can download the predicted models. The algorithm takes about 3 and 10 h in a single processor to predict dimers and hexamers, respectively. The DOT screening is run in 16 processors and takes less than 30 min.

2.6. Results

Given *N*, the number of monomers forming the oligomer, and the 3D structure from one individual monomer, the method predicts the symmetry of the assembly and the structure of the complex. The method builds *N*-mers of different symmetries, clusters the assemblies using a clustering radius of 5 Å pairwise RMSD, and ranks the clusters according to the number of low free energy structures at each interface. The actual predictions correspond to one structure from each ranked cluster that is selected based on a symmetry criterion (see Section 2).

We validate the method on a comprehensive set of 107 *N*-mers (two more cases were run twice using different chains) using the in silico implementation of the algorithm in the server *ClusPro*. In Table 2 we show a summary of the predictions of the server. For 88 of the 109 proteins run, we successfully predict the native-like assembly, and in 104 cases a native-like structure is predicted within the top seven models. The average RMSD of our predictions is about 2 Å. Predictions of complexes with two different binding interfaces were more difficult, typically resulting in higher RMSDs. We should stress that for almost all the complexes in our validation set we ran a monomer extracted from the co-crystallized structure of the complex. We were not able to find good examples of independently crystallized monomers that would form a multimer other than the envelop protein of the tick-borne encephalitis virus (TBEV; Rey et al., 1995). This protein corresponded to Target 10 in the aforementioned CAPRI experiment.

2.7. Prediction of N-mer assemblies

Table 3 lists the name of the PDB code of the complex crystal structure, the rank of the predictions with the lowest RMSD with respect to the crystal, and the size of their corresponding cluster of low free energy structural neighbors. From the group of 39 trimers (bound), we have excellent predictions for 37 complexes (average RMSD of 1.7 Å with respect to the crystal). For two cases, PDB codes 1EK9 and 1N41, we were not able to retain any low free energy structure in the set of 2000 filtered structures (thus, their cluster size was 0). These two proteins are membrane proteins, and for 1N41 phospholipid binding is needed for trimerization (Mo et al., 2003). Since our computational free energy estimate does not account for *trans*-membrane protein interactions, it is not surprising that our free energy filtering fails to retain good structures. Besides these complexes, we also have a few predictions with relatively large RMSDs. For instance, 1OK8 was predicted with an RMSD of 5 Å, much larger than the average. In some of these cases (including 1OK8), we had better lower RMSD models within the members of the top cluster; however, we were not able to select the best structure from the cluster using our symmetry selection criterion (see Section 2).

We tested a total of 35 tetramers, and for 28 complexes our method successfully distinguished between the two symmetries, 16 fourfold (C_4) and 12 D_2 assemblies. In 4 cases, we predicted a near-native structure ranked between 2nd and 7th. The average RMSD of

Table 2
Summary of the number of multimeric assemblies tested, and the ranking of the near-native predictions produced by our automated implementation in the server *ClusPro*

N-mer	Total	Symmetry	Rank of best model			RMSD > 10 Å
			1st	2nd–7th	10th–20th	
Trimers	39	C_3	32	5	1	1
Tetramers	35	C_4	16	—	1	—
		D_2	12	4	—	2
Pentamers	17	C_5	14	2	1	—
Hexamers	17	C_6	4	1	—	—
		DC_3	6	1	—	—
		C_3D	3	1	—	1
TBEV*	1	C_3	1	—	—	—
Total	109		88	14	3	4

* Unbound monomer.

Table 3
List of the PDB codes of the complexes tested and rank of the predictions, full C-α RMSD of the best prediction with respect to the crystal structure, and cluster size of the low free energy structural neighbors

PDB code	Rank	RMSD (Å)	Cluster size
Trimers			
1A3D	2 (1)	1.20 (3.62)	23/17
1ALY	1	1.24	24
1BUU	1	2.08	37
1DBF	1	4.65	106
1DG6	1	1.94	89
1DUN	1	0.61	80
1E12	2 (1)	0.55 (2.93)	24/18
1EK9	20	8.7	0
1EP9	1	1.19	40
1EUA	5	0.95	14
1F7L	1	0.54	35
1FIM	4	1.19	24
1FXZ	1	3.16	12
1G2O	1	1.36	53
1GR3	1	1.04	96
1IHC	1	1.13	60
1JS0	1	1.37	112
1JY8	1	1.26	19
1KCB	1	1.41	87
1KHT	1	3.16	25
1KKE	3	4.85	3
1KRR	1	1.79	138
1MIF	1	1.16	165
1MPM	1	1.24	43
1N41	16	11.46	0
1O91	1	0.69	47
1OK8	1	5.02	14
1P32	1	1.22	58
1PLQ	1	2.03	22
1Q5H	1	1.03	50
1QKP	2	1.46	8
1QRF	1	2.01	52
1QY7	1	1.18	88
1TNF	1	2.34	57
1UFY	3	0.5	10
1V6H	1	0.86	151
2PHL	1	1.28	14
2XAT	1	1.4	105
4PNP	1	1.97	171
C_4 Tetramers			
1A68	1	0.66	58
1AK5	1	2.12	42
1CUK	1	1.65	106
1EZJ	1	5.21	10
1FQY	1	2.03	44
1FUA	1	1.84	58
1FX8	1	0.89	37
1HUV	1	1.58	75
1KBJ	1	1.22	34
1L7F	1	2.87	66
1N93	1	1.02	31
1N9P	1	2.01	124
1NC7	1	0.71	74
1NSC	1	0.64	96
1P7B	19	4.06	0
1PQH	1	0.64	72
1PVN	1	2.93	66
C_5 Pentamers			
1B0C	13	7.4	2
1BZ5	6	1.84	9
1C4Q	1	1.15	91
1DJR	1	2.69	53

(*continued on next page*)

Table 3 (*continued*)

PDB code	Rank	RMSD (Å)	Cluster size
1FB1	1	3.76	101
1JG5	1	1.03	116
1JZN	1	1.92	27
1KZ1	1	1.46	46
1LTA	1	5.1	141
1MSL	2	2.62	9
1NLQ	1	1.36	75
1QB5	1	1.16	114
1SAC	1	2.11	29
1VPS	1	1.55	82
1XTC	1	1.83	82
2BOS	1	1.22	62
3CHB	1	1.41	102
C_6 Hexamers			
1G41	1	1	94
1G6O	1	1.16	24
1LJO	1	0.42	84
1NSF	1	0.62	53
1RRE	2	2.3	20

PDB code	Rank	RMSD (Å)	Dimer A (Å)	Dimer B (Å)	Cluster size A	Cluster size B
D_2: Dimer of dimers						
1A2Z	2	2.25	0.67	1.37	16	16
1A7K	1	3.21	1.27	2.28	52	27
1ADO	1	3.42	1.74	2.95	27	26
1BK4	1	2.83	0.73	2.49	54	36
1BQ4_A	11	20.25	1.27[a]	1.91[a]	19	27
1BQ4_D	3	2.81	0.95	2.18	19	2
1BVQ	1	4.13	1.77	3.49	92	19
1DGE	3	0.50	0.38	0.42	43	11
1ENP	1	2.03	0.55	1.83	35	26
1FTR	2	6.63	1	6.87	23	3
1GIC	1	2.95	0.79	2	35	25
1HDC_A	1	3.72	1.43	1.76	69	7
1HDC_D	1	5.62	1.53	2.04	62	20
1RHP	1	3.64	1.69	1.83	103	13
1XVA	1	5.48	2.25	4.87	18	0
1YKF	7	29.31	1.29[a]	2.18[a]	76	29
4ECA	4 (1)	0.98 (3.90)	2.12 (0.78)	2.68 (0.83)	110	16
6PFK	1	2.07	1.02	1.4	81	45

PDB code	Rank	RMSD (Å)	Trimer (Å)	Dimer (Å)	Cluster size trimer	Cluster size dimer
DC_3: Dimer of trimers						
1BE4	1	2.65	1.09	1.6	78	55
1EHW	1	3.37	1.71	1.34	1	58
1EKR	1	1.42	1.13	0.51	1	71
1I40	1	1.69	0.7	1.09	87	23
1JX7	6	5.11	1.23	2.21	14	42
1LCP	1	4.25	0.68	1.61	27	28
1PJH	1	2.43	1.16	2.04	158	31

PDB code	Rank	RMSD (Å)	Dimer A (Å)	Dimer B (Å)	Cluster size A	Cluster size B
C_3D: Trimer of dimers						
1J2T	1	3.98	0.8	1.12	65	44
1JE0	7	19.21	1.08[b]	2.42[b]	73	30
1NNG	2	3.58	0.88	1.65	42	13
1PJC	1	9.06	1.14	4.99	54	0
1R6L	1	5.19	2.82	5.14	77	23

The runs were made using chain A of the complexes and the number of monomers forming the multimers (N). All predictions used the default settings of the server, except for when we used monomer structures extracted from the complex that we only screen 20,000 (instead of 25,000) DOT generated conformations. For complexes with two binding interfaces, we also provide the predictions for each one of them.

[a] Native-like dimers were discarded because assembly led to structural overlaps.

[b] Native-like dimers were discarded because assembly led to poor symmetry.

all these cases is only 2.5 Å. The prediction ranked poorly for the membrane protein 1P7B, where we again failed to retain any low free energy conformation (see Table 3). For 1BQ4 and 1YKF, we actually found the dimers, however, the D_2 assembly had clashes that failed the overlap criterion and therefore the native-like clusters got discarded.

For 16 of the 17 pentamers, the algorithm predicts a near-native structure with an average RMSD of 2 Å, with 14 of these complexes ranked number one. The worst prediction is for the complex 1B0C (Hamiaux et al., 2000) that forms a pentamer only in acidic pH, whereas here we always assume that binding occurs at neutral pH. Another less than optimal complex is 1LTA, for which we selected the wrong model from the top cluster.

We studied a total of 17 hexamers, for 13 of them we predicted the correct symmetry: 4 sixfold (C_6), 6 dimer of trimers (DC_3), and 3 trimer of dimers (C_3D). We only found 5 hexameric rings in the PDB, typically hexamers are found in the form of dimer of trimers. However, in the 5 cases that we did examine, the predictions had an average RMSD with respect to the crystal of 1.1 Å; all but one was ranked first, with the exception being ranked 2nd, marking a high success rate among these conformations. We also had good predictions for all DC_3 complexes, though the average RMSD, 3 Å, was somewhat higher.

Hexamers exhibiting C_3D symmetry were certainly the most challenging complexes to predict. Three of the five targets predicted a near-native structure ranked 1st, and the fourth was ranked 2nd. Because of poor prediction of 1PJC, the average RMSD of these structures was relatively high 5.5 Å. For 1PJC, the interface between monomer A and D is almost 3900 $Å^2$ and the corresponding cluster size of low energy conformations is 54, whereas the next interface is between the full dimers AD and BF with interfaces of size 1600, 1000, and 1000 $Å^2$ between their monomers DB, DF, and AB, respectively. In other words, the dimers dock at their inter-monomer cleft. Since we identify all our interactions based on two protein interactions, cooperative binding modes involving more than two monomers are difficult to detect. Not surprisingly, the cluster size of the second dimer of 1PJC was 0. The natural conclusion of this result is that the second dimer must dock once the first dimer is formed. This suggestion was checked by first running *ClusPro* to predict a dimer (AD) using monomer A, and then using the predicted dimer (0.73 Å RMSD from the crystal) to predict a trimer, obtaining the hexamer ranked 1st and 2.03 Å away from the crystal. Finally, for 1JE0 the method identified the correct dimers but upon building the trimer of dimers, error propagation led to a final N-mer with poor symmetry between the 1st and $(N+1)$th structure. Hence, the native-like cluster was rejected, predicting a false positive instead.

Finally, we tested the algorithm on three heptamers, predicting high quality models in all cases. Given the small numbers of cases, we do not elaborate on the significance of these results.

2.8. Prediction of multimers based on unbound monomers (Table 4)

The prediction of oligomers based on independently resolved monomers is very difficult. Often the transition from a monomer to a multimer involves structural rearrangements that require special attention. Here, we test three of these cases.

The TBEV envelope protein had been previously crystallized in a dimeric form (Rey et al., 1995), but research has shown that the dimers dissociate and form trimers in acidic pH (Bressanelli et al., 2004). This conformational change enables membrane fusion during cellular infection. Prior to the publication of the crystal structure of the trimeric complex, the envelope protein served as Target 10 for the CAPRI experiment, using monomers of the dimeric structure as the starting three-dimensional coordinates. As a note, the CAPRI organizers announced that there was a significant conformational change in the C terminal domain of the monomer upon binding. Therefore, we removed the C terminal from the analysis. Using as input the first two domains of the unbound dimer (298 residues), the algorithm presented here predicts a model for the trimer that is 4.1 Å from the crystal (PDB code 1URZ). Fig. 4 shows the superposition of the bound structure and the prediction of the server. Most of the RMSD differ-

Table 4
Prediction of multimers based on unbound monomers

PDB unbound	PDB bound	Protein name	Symmetry	Rank	RMSD Å	Cluster size
1SVB	1URZ	Tick-Borne Encephalitis Viral Protein	C_3	1	4.71	3
1IY2	Homolog 1D2N[a]	ATP-Dependent Metalloprotease Ftsh	C_6	10	9.8[b]	2
1JBK	1QVR[a]	Chaperone Clpb	C_6	2	4.96[c]	13

[a] There are no crystal structures for these oligomers.
[b] RMSD between the unit dimer of the predicted C_6 hexamer and the AB dimer in 1D2N.
[c] RMSD between the unit dimer of the predicted C_6 hexamer and the BC dimer in 1QVR.

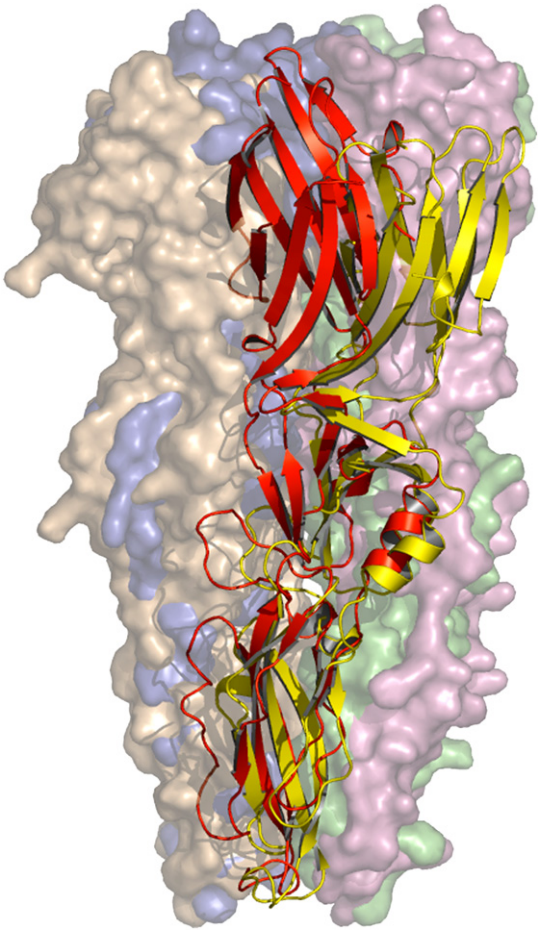

Fig. 4. The bound form of the trimeric TBEV envelope protein (red and solid format) is superposed with the number one model generated by *ClusPro* (green in ribbon format). There is a significant conformational change upon binding as evidenced by the difference between the top domains of the red (bound) and green (unbound) structures.

ence stems from a slight rotation about the primary axis of the monomer. This occurs because the monomers are also optimizing contacts between the middle domain during the initial docking step, and those contacts do not exist in the bound form. Nonetheless, the automated prediction by *ClusPro* is still better than the majority of predicted models made by human experts, and again we emphasize that the near-native prediction is ranked 1st. For CAPRI, we had not implemented the discrimination by clustering low free energy structures, and the purely geometric criterion used by the server at the time ranked the near-native prediction 12th.

We also obtained encouraging predictions for two more hexamers, the protease FtsH domain (Krzywda et al., 2002) and the chaperone Clpb (Lee et al., 2003), whose hexameric crystal structures are not known. The first predicted cyclic hexamer of the FtsH domain ranked 10th. The model has a similar interface than the FtsH homolog 1D2N, but the overall structure appears buckled inside out. The main problems here are that the monomer structure has a flexible hinge that we have kept as in the monomer structure 1IY2 and a missing loop was "copied" from the dimer structure; a small relaxation of these constraints can easily improve the predictions. Finally, the first cyclic ring predicted for the Clpb hexamer ranked 2nd. The dimer structure of this prediction has less than 5 Å RMSD from the dimer interaction of the 1QVR trimer. It is important to note that both structures have highly flexible regions, and important structural motifs are missing from the monomer structures. The above notwithstanding the models suggested by our method are useful starting points for a more elaborate flexible refinement.

2.9. Cluster size suggests order of assembly process

For complexes that involve more than one interface, i.e., D_2, DC_3, and C_3D, the cluster size of each interface listed in Table 3 provides an indication of which interface forms first. This is consistent with the underlying principle that cluster size correlates with the lowest free energy minimum, the larger the cluster size of one interface is the more likely that it forms first. We have already discussed at length the example of 1PJC, a similar observation applies to all cases where the cluster size ratio is larger than 10. Namely, we always find that the large cluster corresponds to the largest interface (that also happens to be more stable according to our free energy scoring function), whereas the interface with fewer cluster members has a smaller pairwise interface. Actually, the second domain always involves interactions with more than one monomer of the substructure that has the largest cluster. The implication is that the latter domain should bind first in order for the second domain to find the appropriate substrate. The latter can readily be seen in the cases involving the complexes 1FTR, 1XVA, 1EHW, 4ECA, and 1EKR. We do not have experimental confirmation for this suggestion. However, it makes sense that the binding pocket should form prior to docking.

3. Discussion

The present study provides a first attempt to predict multimeric proteins, without assuming the symmetry of the complex. As far as we know, this is the first paper done in the field of protein–protein docking that includes multiple binding modes. The only published method to date deals with D_2 tetrameric structures (Berchanski and Eisenstein, 2003). Here, we present an algorithm that successfully docks and discriminates a wide variety of multimeric structures, ranging from complexes consisting of two subunits to complexes consisting of six subunits. The method consists in first docking two identical monomer structures to then scan

the rotational and translational space in which the resulting docked conformations can further assemble. To select the best symmetry and complex structure, we score the resulting assemblies based on a free energy filtering and clustering.

We have shown that our algorithm consistently identifies the native multimer complex as the best predicted model, with an average RMSD between the native complex and predicted model of close to 2Å. In few cases, the algorithm identifies the correct subunits, but fails to predict the correct assembly due to small errors in the prediction of the subunits.

For the subset of 16 D_2 structures analyzed by Berchanski and Eisenstein (2003), we showed comparable results. Two important distinctions in our approach are that first, we do not restrict the sampling of conformations to symmetric dimers, and second, our algorithm identifies the correct symmetry of the assembly (i.e., C_4 or D_2).

Although most of our targets used monomer structures extracted from the complex, the subunits of the co-crystallized complexes have slight perturbations amongst themselves such that the selection of a different subunit may lead to different results. Eisenstein's group has argued that these differences introduce some aspects of unbound docking to this problem. In fact, they had the most difficulties with 1BQ4 and 1HDC. In both of these cases, they did not have good predictions using chain D as the starting monomer. We used our algorithm on both chain A and chain D for these two structures. Ironically, we had success using the D subunit in 1BQ4, while obtaining a poor prediction using chain A. For 1HDC, however, the results for chain D were slightly less accurate than using the A subunit. This shows that the trends observed by Eisenstein are valid, and that each monomer can yield somewhat different results. There was only one case in which the true unbound form was tested, the tick-borne encephalitis virus protein, and the server predicts a native-like complex for this target.

We have observed that most hexamers seem to have a dimer of trimer symmetry. This is also the symmetry that has the largest number of contacts (i.e., pairwise protein–protein interactions), and therefore they are the most packed complexes. Namely, three interfaces for each trimer and three in between the trimers, adding to a total of nine interfaces (see Fig. 3). If this observation holds true in a more exhaustive analysis of all multicomponents assemblies, one might argue that optimal packing could have played a role in the evolution of these large assemblies.

4. Final remarks

We present a general method for the prediction of multimeric assemblies. Given the number of monomers forming an oligomeric complex and the structure of one monomer, the method predicts the symmetry and structure of the complex. The method was designed as a framework to scan all possible interactions. In particular, predictions are selected based on a scoring function that depends only on a homogeneous sampling of low free energy conformations of pairwise protein–protein encounter complexes. In fact, an improved scoring function and computer power could in principle be applied to predict the complex without knowing the number of *N*-mers.

One important limitation of the algorithm is the lack of sampling of interactions between more than two proteins. Thus, predictions of complexes that have proteins with important contacts at the cleft of two other proteins are significantly less accurate than complexes with mostly pairwise interactions.

Although there is still a long way to go to predict multi-protein assemblies, the prediction of multimers and their symmetry could prove useful for a large set of oligomeric complexes. Along these lines, we use our server to predict two hexamer complexes that have not yet been resolved experimentally. Moreover, in silico predictions could also be used in combination with experimental techniques like cryo-EM imaging to refine structural models (Frank, 2002; Chui et al., 2002). With improving scoring functions and computer power, the multimer docking approach presented here might be used as a framework to address the more general problem of multi-protein assemblies.

The algorithm described here has been implemented as part of the public server *ClusPro* (http://structure.pitt.edu), and it is freely available to the research community.

Acknowledgments

The research on this paper was started while C.J.C was at the Faculty of the Department of Biomedical Engineering at Boston University. S.C. is supported by NIH Grant GM61867. C.J.C. is grateful for the support of the University of Pittsburgh and NSF MCB-0444291.

References

Berchanski, A., Eisenstein, M., 2003. Construction of molecular assemblies via docking: Modeling of tetramers with D_2 symmetry. Proteins 53, 817–829.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Res. 28, 235–242.

Bressanelli, S., Stiasny, K., Allison, S.L., Stura, E.A., Duquerroy, S., Lescar, J., Heinz, F.X., Rey, F.A., 2004. Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation. EMBO J. 4, 728–738.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M., 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. 4, 187–217.

Camacho, C.J., Gatchell, D.W., Kimura, S.R., Vajda, S., 2000. Scoring docked conformations generated by rigid-body protein–protein docking. Proteins 40, 525–537.

Camacho, C.J., Vajda, S., 2002. Protein–protein association kinetics and protein docking. Curr. Opin. Struc. Biol. 12, 36–40.

Camacho, C.J., Gatchell, D.W., 2003. Successful discrimination of protein interactions. Proteins 52, 92–97.

Camacho, C.J., Zhang, C., 2005. FastContact: Rapid estimate of contact and binding free energies. Bioinformatics (e-pub available).

Chui, W., Baker, M.L., Jiang, W., Zhou, Z.H., 2002. Deriving folds of macromolecular complexes through electron cryomicroscopy and bioinformatics approaches. Curr. Opin. Struct. Biol. 12, 263–269.

Comeau, S.R., Gatchell, D.W., Vajda, S., Camacho, C.J., 2004a. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics 20, 45–50.

Comeau, S.R., Gatchell, D.W., Vajda, S., Camacho, C.J., 2004b. ClusPro: a fully automated algorithm for protein–protein docking. Nucleic Acids Res. 32, W96–W99.

Frank, J., 2002. Single-particle imaging of macromolecules by cryo-electron microscopy. Annu. Rev. Biophys. Biomol. Struct. 31, 303–319.

Gabb, H.A., Jackson, R.M., Sternberg, M.J.E., 1997. Modelling protein docking using shape complementarity, electrostatics, and biochemical information. J. Mol. Biol. 272, 106–120.

Gan, L., Conway, J.F., Firek, B.A., Cheng, N., Hendrix, R.W., Steven, A.C., Johnson, J.E., Duda, 2004. Mol. Cell 14, 559–569.

Halperin, I., Ma, B., Wolfson, H., Nussinov, R., 2002. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 47, 409–443.

Hamiaux, C., Perez, J., Prange, T., Veesler, S., Ries-Kautt, M., Vachette, P., 2000. The BPTI decamer observed in acidic pH crystal forms pre-exists as a stable species in solution. J. Mol. Biol. 297, 697–712.

Janin, J., Henrick, K., Moult, J., Ten Eyck, L., Sternberg, M.J., Vajda, S., Vakser, I., Wodak, S.J., 2003. CAPRI: a critical assessment of predicted interactions. Proteins 52, 2–9.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A., Aflalo, C., Vakser, I.A., 1992. Molecular surface recognition—determination of geometric fit between proteins and their ligands by correlation techniques. Proc. Natl. Acad. Sci. USA 89, 2195–2199.

Krzywda, S., Brzozowski, A.M., Verma, C., Karata, K., Ogura, T., Wilkinson, A.J., 2002. The crystal structure of the AAA domain of the ATP-dependent protease FtsH of *Escherichia coli* at 1.5 Å resolution. Structure 10, 1073–1083.

Lee, S., Sowa, M.E., Watanabe, Y.H., Sigler, P.B., Chiu, W., Yoshida, M., Tsai, F.T., 2003. The structure of ClpB: a molecular chaperone that rescues proteins from an aggregated state. Cell 115, 229–240.

Mandell, J.G., Roberts, V.A., Pique, M.E., Kotlovyi, V., Mitchell, J.C., Nelson, E., Tsigelny, I., TenEyck, L.F., 2001. Protein docking using continuum electrostatics and geometric fit. Protein Eng. 14, 105–113.

Mendez, R., Leplae, R., De Maria, L., Wodak, S.J., 2003. Assessment of blind predictions of protein–protein interactions: Current status of docking methods. Proteins 52, 51–67.

Mo, Y., Campos, B., Mealy, T.R., Commodore, L., Head, J.F., Dedman, J.R., Seaton, B.A., 2003. Interfacial basic cluster in annexin V couples phospholipid binding and trimer formation on membrane surfaces. J. Biol. Chem. 278, 2437–2443.

Norel, R., Sheinerman, F., Petrey, D., Honig, B., 2001. Electrostatic contributions to protein–protein interactions: Fast energetic filters for docking and their physical basis. Prot. Sci. 10, 47–61.

Rey, F.A., Heinz, F.X., Mandl, C., Kunz, C., Harrison, S.C., 1995. The envelope glycoprotein from tick-borne encephalitis virus at 2 Å resolution. Nature 375, 275–276.

Ritchie, D.W., Kemp, G.J.L., 2000. Protein docking using spherical polar Fourier correlations. Proteins 39, 178–194.

Smith, G.R., Sternberg, M.J.E., 2002. Prediction of protein–protein interactions by docking methods. Curr. Opin. Struct. Biol. 12, 28–35.

TenEyck, L.F., Mandell, J., Roberts, V.A., Pique, M.E., 1995. Surveying molecular interactions with DOT. In: Hayes, A., Simmons, M., (Eds.), Proceedings of the 1995 ACM/IEEE Supercomputing Conference. ACM Press, New York.

Weng, Z., Vajda, S., DeLisi, C., 1996. Prediction of protein complexes using empirical free energy functions. Protein Science 5, 614–626.

Wodak, S.J., Mendez, R., 2004. Prediction of protein–protein interactions: the CAPRI experiment, its evaluation and implications. Curr. Opin. Struc. Biol. 14, 242–249.

Zhang, C., Cornette, J.L., DeLisi, C., 1997. Determination of atomic desolvation energies from the structures of crystallized proteins. J. Mol. Biol. 267, 707–726.