Performance of the First Protein Docking Server *ClusPro* in CAPRI Rounds 3–5

Stephen R. Comeau,[1] Sandor Vajda,[2] and Carlos J. Camacho[3*]
[1]*Bioinformatics Graduate Program, Boston University, Boston, Massachusetts*
[2]*Department of Biomedical Engineering, Boston University, Boston*
[3]*Department of Computational Biology, University of Pittsburgh, Pittsburgh, Pennsylvania*

ABSTRACT **To evaluate the current status of the protein–protein docking field, the CAPRI experiment came to life. Researchers are given the receptor and ligand 3-dimensional (3D) coordinates before the cocrystallized complex is published. Human predictions of the complex structure are supposed to be submitted within 3 weeks, whereas the server *ClusPro* has only 24 h and does not make use of any biochemical information. From the 10 targets analyzed in the second evaluation meeting of CAPRI, *ClusPro* was able to predict meaningful models for 5 targets using only empirical free energy estimates. For two of the targets, the server predictions were assessed to be among the best in the field. Namely, for Targets 8 and 12, *ClusPro* predicted the model with the most accurate binding-site interface and the model with the highest percentage of nativelike contacts, among 180 and 230 submissions, respectively. After CAPRI, the server has been further developed to predict oligomeric assemblies, and new tools now allow the user to restrict the search for the complex to specific regions on the protein surface, significantly enhancing the predictive capabilities of the server. The performance of *ClusPro* in CAPRI Rounds 3–5 suggests that clustering the low free energy (i.e., desolvation and electrostatic energy) conformations of a homogeneous conformational sampling of the binding interface is a fast and reliable procedure to detect protein–protein interactions and eliminate false positives. Not including targets that had a significant structural rearrangement upon binding, the success rate of *ClusPro* was found to be around 71%. Proteins 2005;60:239–244.**
© 2005 Wiley-Liss, Inc.

Key words: docking; protein interactions; complex structure; supramolecular assembly

INTRODUCTION

The goal of traditional protein–protein docking algorithms is to take the 3-dimensional (3D) coordinates of 2 independently crystallized proteins that are known to interact and to derive a model for the cocrystallized structure.[1–3] Billions of putative complexes are evaluated by scanning the rotational and translational space between the 2 proteins, often using the Fast Fourier Transform (FFT) technique[4] to expedite the calculation. Then, these putative complexes are subjected to various discrimination techniques in order to eliminate false-positive structures, in search of the high-affinity complex.[5–7]

In order to evaluate the current status of the field, the CAPRI experiment came to life.[8] Computational researchers are given the 3D coordinates of the unbound structures before the cocrystallized complexes are published. The researchers are then given a few weeks to dock the 2 structures together, and can use any information necessary, including biological information and literature searches. 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 Rounds 1 and 2 of CAPRI, Camacho and Gatchell[9] produced some of the best model structures, appropriately distinguishing between near-native and false-positive structures. Based on these promising results, we implemented our filtering and discrimination methods for rigid-body docking algorithm as a public server named *ClusPro*[10] (http://structure.bu.edu). The *ClusPro* Web server is a fully automatic algorithm that rapidly docks, filters, and ranks putative protein complexes within a short amount of time using only the given structures of the component proteins and thermodynamic considerations. One important motivation for developing the *ClusPro* server was to eliminate human intervention, which biases predictions against how well the actual docking algorithms perform. Indeed, in almost all CAPRI targets there was some relevant biochemical information pointing to the binding site, and from the submitted models it is not possible to determine how well a particular method would have performed without human intervention.

We report on the results of the automated server *ClusPro* in Rounds 3–5 of the CAPRI experiment. *ClusPro* was the only automated server participating on this experiment, and the only docking technology that claims finding near-native complex structures using only thermodynamic considerations. *ClusPro* submitted near-native structures for 5 of the 10 targets. In addition, the server submitted a good prediction for a homology modeling target that was

Grant sponsor: National Institutes of Health; Grant number: GM61867 (to Boston University). Grant sponsor: University of Pittsburgh (C. J. Camacho), NSF MCB–0444291.

*Correspondence to: Carlos J. Camacho, Department of Computational Biology, University of Pittsburgh, 200 Lothrop Street, Pittsburgh, PA 15261. E-mail: ccamacho@pitt.edu

Received 17 January 2005; Accepted 8 February 2005
DOI: 10.1002/prot.20564

eventually cancelled because human predictors had access to published information. Three of the targets missed by *ClusPro* had significant structural rearrangement upon binding, and required extra information for them to be predicted. A novel target of CAPRI consisted of a trimeric form of the tick-borne encephalitis virus (TBEV) envelope protein.[11] This posed a new challenge to the computational modeling of protein interactions, as 3 monomeric structures needed to be docked together in order to obtain the final outcome. Motivated by this target, we have further developed the server and implemented a new algorithm to predict multimers with different types of symmetry.

METHODS

Rigid-Body Docking

Using the *ClusPro* server, one has the option of selecting DOT[12] or ZDOCK[13] to perform the initial rigid-body docking. Both methods are based on the FFT correlation approach[4] that systematically evaluates a simple grid-based scoring function over billions of relative orientations of the 2 proteins. With DOT we use a shape complementarity score, whereas ZDOCK scoring function includes a combination of shape complementarity, Coulombic electrostatics, and desolvation free energy based on the Zhang et al.[14] atomic contact potential. The latter has already been shown to be important for finding nativelike complex structures.[7] As default, ZDOCK retains 2000 structures. In our methodology, we use these FFT-based tools to rapidly generate a large number of receptor–ligand conformations with good shape complementarity (and in the case of ZDOCK, with relatively favorable electrostatics and desolvation values). Usually the top 20,000 structures are retained for further analysis.

Filtering Using Empirical Free Energy Functions[7]

Since the dominant interactions for protein–protein association are electrostatics and desolvation free energies, we have shown that selecting 2000 structures from the initial 20,000 using empirical free energy estimates yields an appropriate sampling of the free energy landscape. We compute the electrostatic energy using a Coulombic model with a distance-dependent dielectric of $4r$. The desolvation free energy is computed using a knowledge-based atomic contact potential.[14] As default, we retain the 1500 structures with the best electrostatic energies and 500 structures with the best desolvation energies. The reason for retaining more structures with favorable electrostatics is that the electrostatic energy is much more sensitive to small perturbations in the coordinates than the desolvation term, and restricting consideration to 500 structures with the best calculated values of electrostatics generally leads to losing a number of good solutions. Since ZDOCK already screens 2000 structures based on a similar scoring function, we keep them all without rescoring.

Clustering the 2000 Filtered Docked Conformations[9]

The final step of the *ClusPro* server consists in clustering the generated 2000 docked conformations on the basis of their pairwise ligand-binding site root-mean-square deviation (RMSD) values. Note that all conformations have the receptor fix in the origin of the coordinate system. The default clustering radius is 9 Å. To calculate pairwise binding-site RMSD values, we select the residues of the ligand that are within 10 Å of the receptor, and for each docked conformation compute the RMSD of these residues with the same residues in every other docked ligand conformation. The 2000 conformations are then clustered using the binding-site RMSD as the distance measure (i.e., on the basis of the 2000 × 2000 RMSD matrix), using a standard greedy algorithm. The predictions are ranked according to the size of the clusters. It is assumed that larger clusters indicate wider and deeper free energy minima, and that such minima have higher likelihood to correspond to the binding site.[15]

Predicting Assemblies of Homo-*N*-Mers[16]

Target 10 in CAPRI involved the prediction of a homotrimer.[11] Motivated by this target, we developed a general algorithm to predict the assembly of homo-*N*-mers, based solely on the structure of the monomer and the number of monomers *N*. 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 complexes found in the free energy filtered set of 2000 structures. The actual predictions correspond to the structure with best symmetry and no overlaps among the *N* proteins from each ranked cluster. The algorithm has been implemented as part of *ClusPro*. Users upload the structure of a monomer and the numbers of monomers involved in the assembly (between 2 and 7). Note that the actual symmetry does not have to be specified. The method predicts both the symmetry and optimal structure. For instance, in the case of a tetramer, the method would predict if the assembly forms a 4-fold symmetry structure or a dimer of dimers.

RESULTS

The predictions submitted by the server *ClusPro* are summarized in Table I. We emphasize that these results are fully reproducible by simply uploading the targets to the server. *ClusPro* predicted a near-native complex for 5 of the 10 targets. A sixth target (T17) was also predicted correctly, but the CAPRI management did not evaluate this target. Three of the targets (T9, a dimer; T10, a trimer; and T11, a distant homology model) had significant structural rearrangement upon binding and were not predicted correctly by the server. In the original validation of *ClusPro*,[10] we obtained a success rate of 74% for complexes that do not undergo significant rearrangement upon binding. Removing Targets 9, 10, and 11 from the statistics, the success rate of *ClusPro* is 71% consistent with our earlier claim. In what follows, we discuss each of the targets.

Target 8: Nidogen G3 Domain–Laminin EGF Modules 3–5 Complex[17]

This target was of moderate difficulty,[18] since the binding surface was somewhat polar and not very large.

TABLE I. Nativelike Models Submitted by *ClusPro* to CAPRI Rounds 3–5

Target	*ClusPro* Model No.	f_{nat}[a] ranking[b]	Ligand RMSD [Å]/ranking[b]	Interface RMSD [Å]/ranking[b]	CAPRI score
8	3	0.455/4th	6.29/2nd	0.48/1st	**
12	9	0.927/1st	3.22/11th	0.78/9th	***
13	2	0.129/12th	14.376/11th	2.987/12th	*
15	8	0.5/3rd	6.04/5th	1.83/4th	**
17[c]	1	0.24	17	5.7	(*)
19	2	0.296/10th	6.91/9th	2.48/9th	*

[a] f_{nat} is the ratio of native contacts

[b] Ranking among all groups

[c] *ClusPro* prediction for T17; T16 and T17 were cancelled after submission.

Nevertheless, model 3 [shown in Fig. 1(A)] predicted by the server resulted in the best interface RMSD among all 180 models submitted. It is important to emphasize that this prediction was obtained by simply uploading the receptor and ligand target structures in the server, and it only took a couple of hours to predict.

Target 9: LicT Dimer[19]

This target had a significant change between the unbound and the bound conformation (13 Å RMSD). Since *ClusPro* is a rigid-body approach, it was not able to predict the complex. At a minimum, automating the prediction of these types of targets would require to assume flexibility at the hinges of the structure.[20]

Target 10: TBEV Glycoprotein E Trimer[11]

The TBEV envelope protein had been previously crystallized in a dimeric form,[21] but in acidic pH, the dimers dissociate and form trimers. Prior to the publication of the crystal structure of the trimeric complex, the envelope protein served as one of the targets for the CAPRI experiment, using monomers of the dimeric structure as the given component proteins. Because of this target, we developed a general algorithm to predict the assembly of homo-*N*-mers. By uploading the structure of one monomer and setting the number *N* equal to 3, the server predicts a structure that is 4.1 Å away from the crystal. The prediction is shown in Figure 1(C). In order to obtain this prediction, we also removed the C-terminal domain that was known to differ significantly between the bound and unbound monomers. These 2 structures are 12 Å RMSD apart.

At the time Target 10 was given, the full implementation of the algorithm was not ready, and the scoring function was surface complementarity alone. The structure in Figure 1(C) was then ranked 12th. It was only after the deadline that we implemented the scoring function based on the clustering of low free energy docked conformations (similar to the one used in traditional docking).

Target 11: Cellulosome Cohesin–Dockerin (Homology Model) Complex[22]

The homology model built for dockerin was 5 Å RMSD away from the crystal, and *ClusPro* was not able to find any docked conformation with reasonable affinity between the receptor and the homology model.

Target 12: Cellulosome Cohesin–Dockerin Complex[22]

For Target 12, the bound structure of the dockerin protein was provided. Uploading the target structures "as is" on the server led to the prediction in Figure 1(B). As for Target 8, the *ClusPro* prediction for Target 12 is a high-quality model compared to most manual predictions, having the best ratio of native contacts (see Table I). Since this target was given after Target 11 was closed, we knew that association was mediated by desolvation forces. Thus, for this target only, we ran *ClusPro* using ZDOCK as screening method. ZDOCK has a stronger bias to select hydrophobic contacts than DOT (see Methods section).

Target 13: *Toxoplasma gondii* Surface Antigen 1 (SAG1)–FAB Complex[23]

The server predicted 2 good models of this antibody–antigen complex, Models Nr. 2 and Nr. 4. Model 4, shown in Figure 1(D), has a slightly higher interface RMSD than Model 2 listed in Table I, but the figure is more illustrative, since it shows how the antigen correctly binds to one of the binding grooves and misses the second one.

Target 14: Protein Phosphatase 1–Myosin Phosphatase-Targeting Subunit 1 (MYPT1) Complex[24]

ClusPro was unable to predict this complex because the arm of the MYPTI structure wrapped too tightly around the phosphatase, constraining the complex structure to such an extent that our method failed to have a good enough sampling of the binding area. Thus, the clustering procedure failed to get good predictions.

Target 15: Colicin D Nuclease–Immunity D Complex[25]

Target 15 was a relatively difficult target[18] because the binding site was polar, and the targets had no side-chains. Blind predictions on these types of targets are usually quite challenging. We uploaded the targets after building the side-chains using CHARMM.[26] Without using any information regarding the location of the catalytic site,

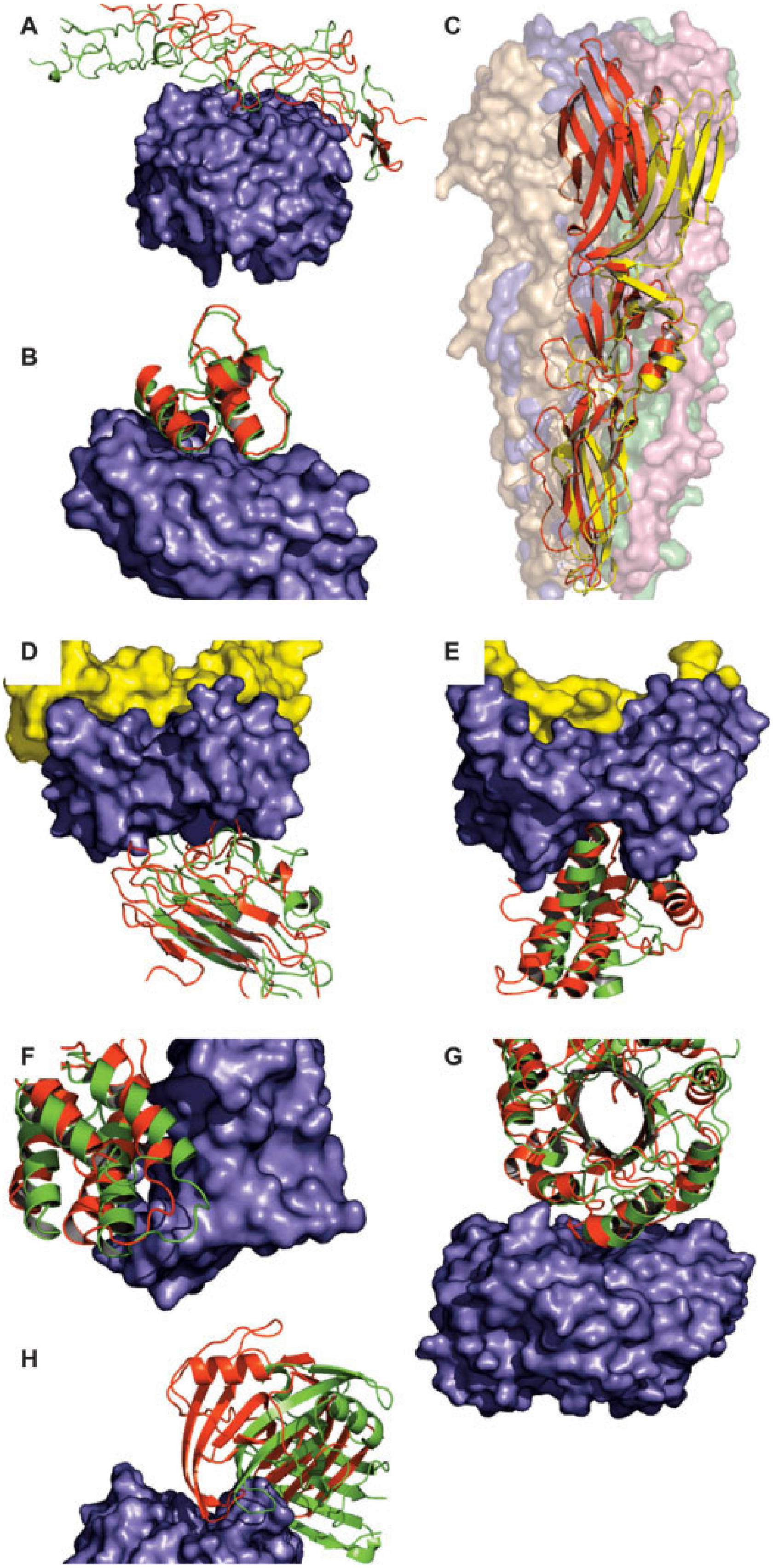

Fig. 1. *ClusPro* predictions for (**A**) Target 8, (**B**) Target 12, (**C**) Target 10 (this prediction was not within the 10 submitted structures); the antibody–antigen complexes are shown in (**D**) Target 13 and (**E**) Target 19; and, the predictions for the homology model targets are shown in (**F**) Target 15, (**G**) Target 16, and (**H**) Target 17. For Targets 8 and 12, the models shown are among the best of all submissions. Shown in red is the bound ligand, and in green, the model. The receptors are shown in cyan as solid models. For the antibody–antigen systems we have colored in yellow residues that were blocked from docking using the utility *block.pl* available from the *ClusPro* website. The figures were made using PyMOL (DeLano, WL. The PyMOL Molecular Graphics System. 2002. Available online at http://www.pymol.org).

ClusPro predicted the model with the third best ratio of native contacts among all the participants [see Fig. 1(F)].

Target 16: Xylanase (*Aspergillus nidulans*)–XIP-1 Inhibitor Complex[27]

The structure of the xylanase was built based on a xylanase from *Penicillium simplicissimum*, a close homolog [Protein Data Bank (PDB) code: 1BG4], using the server Consensus.[28] *ClusPro* did not submit a good structure within the top 10 models. However, the model ranked 22nd had the prediction shown in Figure 1(G). We note that the server has the option of requesting as many as 30 models for a given target. Most of the false positives did not block the active site.

Target 17: Xylanase (*Penicillium funiculosum*)–XIP-1 Inhibitor Complex[27]

The structure of the xylanase was built based on the homolog xylanase from *A. niger* (PDB code: 1UKR), using the server Consensus[28] and side-chains were modeled using molecular dynamics (MD).[29] The top model by *ClusPro* is shown in Figure 1(H). *ClusPro* readily identified the binding region and correct contacts. The model is rotated around the main contacts by 40° or so. The reason for this rotation is that the homology model was missing the first 2 β-sheets, leaving exposed hydrophobes at the N-terminal. The interaction between these residues and the inhibitor rotated the structure in order to have a larger contact area at the interface than that observed in the crystal.

Target 18: Xylanase (*A. niger*)–(*Triticum aestivum*) Xylanase Inhibitor Complex[30]

ClusPro did not submit a good prediction for this target. The reason for this failure might be the fact that His374 in the inhibitor is likely protonated.

Target 19: Homology Model of Ovine PrPc–FAB Complex[31]

The homology model for the ovine prion was constructed using the server Consensus, and then CHARMM was used to build the missing side-chains. As shown in Figure 1(E), Model 2 from *ClusPro* was very close to the antibody–antigen complex structure.

DISCUSSION

The results of Rounds 3–5 of CAPRI demonstrated that the server *ClusPro* is a fast and reliable predictor of protein–protein complexes, provided that the complex does not undergo a significant structural transformation upon binding. We have shown that for 5 of the 7 such targets in Rounds 3–5, our algorithm identified a native-like complex structure within the best 10 models, a 71% success rate. The robustness of the method is manifested on the fact that *ClusPro* produced meaningful models for all the targets the homology modeling targets with a close homolog in the PDB. Namely, for Targets 15, 16, 17, and 19, the server predicted native-like conformations within the top 10 models; for Target 16, a high-quality model was ranked 22. Target 11 was also a homology-modeling target but the template structure was different from the target.

The server has also been further developed to allow modeling of multimeric assemblies. Given the number of monomers forming a multimeric complex and the structure of one monomer, the method predicts the symmetry and structure of the complex. The method was designed to scan all possible interactions, and select the models with the broadest free energy funnels that also satisfy the symmetry constraints without steric overlaps.

ClusPro has been validated as a technology capable of predicting protein complexes based solely in thermodynamic free energy estimates.

REFERENCES

1. Camacho CJ, Vajda S. Protein–protein association kinetics and protein docking. Curr Opin Struct Biol 2002;12:36–40.
2. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 2002;47:409–443.
3. Smith GR, Sternberg MJE. Prediction of protein–protein interactions by docking methods. Curr Opin Struct Biol 2002;12:28–35.
4. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem A, Aflalo C, Vakser IA. Molecular surface recognition—determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci USA 1992;89:2195–2199.
5. Weng Z, Vajda S, DeLisi C. Prediction of protein complexes using empirical free energy functions. Protein Sci 1996;5:614–626.
6. Jackson RM, Gabb HA, Sternberg MJE. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. J Mol Biol 1998;276:265–285.
7. Camacho CJ, Gatchell DW, Kimura SR, Vajda S. Scoring docked conformations generated by rigid-body protein–protein docking. Proteins 2000;40:525–537.
8. Janin J, Henrick K, Moult J, Ten Eyck L, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. CAPRI: A Critical Assessment of PRedicted Interactions. Proteins 2003;52:2–9.
9. Camacho CJ, Gatchell DW. Successful discrimination of protein interactions. Proteins 2003;52:92–97.
10. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. *ClusPro*: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics 2004;20:45–50.
11. Bressanelli S, Stiasny K, Allison SL, Stura EA, Duquerroy S, Lescar J, Heinz FX, Rey FA. Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation. EMBO J 2004;4:728–738.
12. Ten Eyck LF, Mandell J, Roberts VA, Pique ME. Surveying molecular interactions with DOT. In Hayes A, Simmons M, editors. Proceedings of the 1995 ACM/IEEE Supercomputing Conference. New York: ACM Press; 1995.
13. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein docking algorithm. Proteins 2003;52:82–87.
14. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 1997;267:707–726.
15. Camacho CJ, Weng ZP, Vajda S, DeLisi C. Free energy landscapes of encounter complexes in protein–protein association. Biophys J 1999;76:1166–1178.
16. Comeau SR, Camacho CJ. Predicting oligomeric assemblies: N-mers a primer. J Struc Biol (in press).
17. Takagi J, Yang Y, Liu JH, Wang JH, Springer TA. Complex between nidogen and laminin fragments reveals a paradigmatic beta-propeller interface. Nature 2003;424:969–974.
18. Vajda S, Camacho CJ. Protein–protein docking: is the glass half full or half empty? Trends Biochem Sci 2004;22:110–116.
19. Graille M, Zhou CZ, Receveur V, Collinet B, Declerck N, van Tilbeurgh H. Structure of the native and inactive LicT PRD from *B. subtilis*. J Biol Chem 2005;280:14780–14789.
20. Schneidman-Duhovny D, Inbar Y, Polak V, Shatsky M, Benyamini H, Barzilay A, Dror O, Haspel N, Nussinov R, Wolfson HJ. Taking geometry to its edge: Fast unbound rigid (and hinge-bent) docking. Proteins 2003;52:107–112.

21. Rey FA, Heinz FX, Mandl C, Kunz C, Harrison SC. The envelope glycoprotein from tick-borne encephalitis virus at 2 Å resolution. Nature 1995;375:275–276.
22. Carvalho AL, Dias FM, Prates JA, Nagy T, Gilbert HJ, Davies GJ, Ferreira LM, Romao MJ, Fontes CM. Cellulosome assembly revealed by the crystal structure of the cohesin–dockerin complex. Proc Natl Acad Sci USA 2003;100:13809–13814.
23. Graille M, Stura E, Bossus M, Muller BH, Letourneur O, Battail-Poirot N, Sibaï G, Gauthier M, Rolland D, Le Du MH, Ducancel F. Structure of the immunodominant epitope displayed by the surface antigen 1 (SAG1) of *Toxoplasma gondii* complexed to a monoclonal antibody. 2005. Submitted for publication.
24. Terrak M, Kerff F, Langsetmo K, Tao T, Dominguez R. Structural basis of protein phosphatase 1 regulation. Nature 2004;429:780–784.
25. Graille M, Mora L, Buckingham RH, Van Tilbeurgh H, De Zamaroczy M. Structural inhibition of the colicin D tRNase by the tRNA-mimicking immunity protein. EMBO J 2004;23:1474–1482.
26. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4: 187–217.
27. Payan E, Leone P, Porciero S, Furniss C, Tahir T, Williamson G, Durand A, Manzanares P, Gilbert H, Juge N, Roussel A. The dual nature of the wheat xylanase protein inhibitor Xip-I: structural basis for the inhibition of family 10 and family 11 xylanases. J Biol Chem 2004;279:36029–36037.
28. Prasad JC, Comeau SR, Vajda S, Camacho CJ. Consensus alignment for reliable framework prediction in homology modeling. Bioinformatics 2003;19:1682–1691.
29. Rajamani D, Thiel S, Vajda S, Camacho CJ. Anchor residues in protein–protein interactions. Proc Natl Acad Sci USA 2004;101: 11287–11292.
30. Sansen S, De Ranter CJ, Gebruers K, Brijs K, Courtin CM, Delcour JA, Rabijns A. Structural basis for inhibition of *Aspergillus niger* xylanase by *Triticum aestivum* xylanase inhibitor-I. J Biol Chem 2004;279:36022–36028.
31. Eghiaian F, Grosclaude J, Lesceu S, Debey P, Doublet B, Treguer E, Rezaei H, Knossow M. Insight into the PrPC→PrPSc conversion from the structures of antibody-bound ovine prion scrapie-susceptibility variants. Proc Natl Acad Sci USA 2004;101:10254–10259.