Biologically enhanced sampling geometric docking and backbone flexibility treatment with multiconformational superposition

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · AUGUST 2005

Impact Factor: 2.63 · DOI: 10.1002/prot.20577 · Source: PubMed

6 AUTHORS, INCLUDING:

Xinqi Gong
Renmin University of China
29 PUBLICATIONS **646** CITATIONS

SEE PROFILE

Biologically Enhanced Sampling Geometric Docking and Backbone Flexibility Treatment With Multiconformational Superposition

Xiao Hui Ma,[†] Chun Hua Li,[†] Long Zhu Shen, Xin Qi Gong, Wei Zu Chen, and Cun Xin Wang[*]
College of Life Science and Bioengineering, Beijing University of Technology, Beijing, China

ABSTRACT **An efficient biologically enhanced sampling geometric docking method is presented based on the FTDock algorithm to predict the protein–protein binding modes. The active site data from different sources, such as biochemical and biophysical experiments or theoretical analyses of sequence data, can be incorporated in the rotation–translation scan. When discretizing a protein onto a 3-dimensional (3D) grid, a zero value is given to grid points outside a sphere centered on the geometric center of specified residues. In this way, docking solutions are biased toward modes where the interface region is inside the sphere. We also adopt a multiconformational superposition scheme to represent backbone flexibility in the proteins. When these procedures were applied to the targets of CAPRI, a larger number of hits and smaller ligand root-mean-square deviations (RMSDs) were obtained at the conformational search stage in all cases, and especially Target 19. With Target 18, only 1 near-native structure was retained by the biologically enhanced sampling geometric docking method, but this number increased to 53 and the least ligand RMSD decreased from 8.1 Å to 2.9 Å after performing multiconformational superposition. These results were obtained after the CAPRI prediction deadlines. Proteins 2005;60:319–323.**

Key words: biologically enhanced sampling; multiconformational superposition; backbone flexibility; active site

INTRODUCTION

The first protein–protein docking algorithm[1] was developed by Janin and Wodak in 1978, and docking methods have greatly advanced since their study stimulated by the CAPRI experiment.[2] In the past few years, many promising docking algorithms have been proposed,[3–6] and successful predictions have been achieved.[7–10] However, the rate of correct prediction is generally higher when biological information is available,[11] for instance, the residues implicated in the activity of the protein. Thus, strategies have been proposed to incorporate biochemical information into the prediction procedure to raise the probability of successful prediction. This can be done during the rotation–translation scan or at the postscan filter to eliminate false-positive solutions. An example of the former choice is the weighted geometric docking method of Ben-Zeev and Eisenstein,[12] which gives more weight during the scan to intermolecular contacts that involve specified residues in one or both of the protein molecules. To illustrate the second choice, the program 3D-Dock of the Sternberg group[3] translates the information on active site residues into distance constraints. The postscan filter requires particular pairs of residues to be close, or it ensures at least that these residues are at the interface. Both methods generally improve the ranking of near-native solutions to some degree. The major problem with the postscan filter is that it cannot help much if the list of solutions produced by the scan contains few near-native modes. This was the reason for our failure in some CAPRI predictions, such as for Targets 13–14 and Targets 17–18.

Handling molecular flexibility during protein–protein docking has been an essential and challenging topic in every round of CAPRI.[13,14] Different methods have been proposed to consider the conformational flexibility of the backbone[15,16] and side-chains.[3,6,17–18] Our work focuses on backbone flexibility, which exerts a greater influence on the quality of the prediction. As a single conformation cannot describe the dynamic properties of proteins, we first generate an ensemble of states for both the receptor and the ligand, and then dock these individual conformational states. This scheme has already been applied to protein–ligand docking in drug design. Shoichet and Kuntz[19] have proposed two related methods, "energy-weighted average" and "geometry-weighted average" for molecular docking, that utilize the information on conformational variability from ensembles of experimental receptor structures. Both methods can accurately reproduce experimentally determined binding orientations in some cases.

In this study, we introduce a method that incorporates external binding site information in the rotation–translation scan performed by the program FTDock,[20] by giving zero value to grid points outside a sphere centered on the

Grant sponsor: Chinese Natural Science Foundation; Grant numbers: 30170230 and 30400087. Grant sponsor: Beijing City Education Committee; Grant numbers: KZ200410005002 and 20040005013.

*Correspondence to: Cun Xin Wang, College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100022, China. E-mail: cxwang@bjut.edu.cn

[†]Both are first authors.

Received 20 January 2005; Accepted 4 March 2005

DOI: 10.1002/prot.20577

geometric center of specified residues in either one or both protein molecules. In this way, we generate a set of solutions biased toward binding modes where the interface region is included in the sphere. We also describe a multiconformational superposition method that accounts for the backbone flexibility of proteins. The application of these 2 approaches to CAPRI targets T08,[21] T11,[22] T12,[22] T14,[23] T15,[24] T16–T18, and T19[25] improves both the number of near-native structures and the ligand root-mean-square deviation (RMSD). What needs to be pointed out is that these results were obtained after the CAPRI prediction deadlines.

MATERIALS AND METHODS

Biologically Enhanced Sampling Geometric Docking

Based on the docking program FTDock,[20] the biologically enhanced sampling geometric docking method incorporates binding-site information into the rotation–translation scan. The protein molecules to be docked are digitalized onto a 3D grid. In regular geometric docking, grid points are given different values that depict the shape of the molecule. On the receptor, surface grid points are given the value 1, those in the interior are given the value −15, and grid points outside the molecule, the value 0. For the ligand, grid points on the surface and in the interior of the molecule are both given the value 1. When some binding-site residues are known on the receptor, we perform biologically enhanced sampling geometric docking by modifying this scheme and setting to zero the values of all grid points outside a sphere centered on the geometric center of the binding-site residues (the average position of all atoms of the binding-site residues). Points inside the sphere keep their values. The sphere radius is selected to cover the binding surface of the receptor. If information on active site residues of the ligand is also available, a similar treatment can be applied to ligand grid points. When a score is calculated by correlating the two grid representations, only the overlaps between grid points inside the spheres can contribute positive values to the score. Terms involving grid points outside the spheres of either the receptor or the ligand make no contribution. In this way, for each orientation of the two molecules, a preference is given to the translations that create more contacts in regions of the protein surface covered by the sphere, and we do not need to constrain the orientations.

Multiconformational Superposition to Represent Backbone Flexibility

An ensemble of conformations can be obtained for each protein through molecular dynamics (MD) simulation. The protein molecule is subject to energy minimization using both steepest descent and conjugate gradient methods, and then solvated by water molecules in a rectangular box. The solvent molecules are minimized and heated up to 300 K with the protein coordinates constrained. Finally, the whole solvated models are heated up to 300 K and subject to a free equilibrium MD simulation for 1 ns. The MD simulations are performed in the constant number of particles, temperature and pressure (NTP) ensemble with the GROMACS program package.[26] From the conformations under the equilibrium state, we select those structures where the backbone RMSD of the binding site residues is above a threshold value, and superimpose them based on backbone atom positions. Each conformation is then projected and digitalized onto a 3D grid based on the regular approach respectively. Finally, a single grid is derived, in which we apply 0 to those grid points with different values and keep the values unchanged for the identical ones through comparing the grid point value of the multigrids. After conformational searching, the docked modes are built using the superimposed structure, which can result in some clashes. We remove the latent clashes by applying energy minimization with the GROMACS program package.

Scoring Functions

The combined scoring function used here is expressed as

$$Score = SCR \times 0.01 + RP \times 0.45 + ACE \times 0.35 + ELE \times 0.80 + E_{atr} + E_{res} \quad (1)$$

where the *SCR* item is from FTDock and the *RP* item[2] contains the surface complementarity score and residue level pair potentials. *ACE*[27] is desolvation free energy based on the atomic contact energy. Electrostatic contributions are calculated using Coulomb's law with a distance-dependent dielectric constant: $ELE = 332 \times q_i q_j / 4r_{ij}^2$, where r_{ij} = max (r_{ij}, 1.6 Å) to avoid the singularity at atom–atom distance $r_{ij} = 0$. E_{atr} is the attractive portion of a 12-6 Lennard–Jones potential. E_{res} is the repulsive energy between 2 atoms. Both E_{atr} and E_{res} items are described in Kuhlman and Baker.[28] Their weights in the scoring function have been obtained by multiple regression based on the benchmark set.[29]

RESULTS AND DISCUSSION

Biologically Enhanced Sampling Based on Biological Information

We tested the biologically enhanced sampling geometric docking algorithm on nine CAPRI targets (see Table I). Due to T09 monomer's very large conformational changes (backbone RMSD larger than 10 Å) upon binding, the test on it was not done. Additionally, T10 is a symmetric trimer, so a good way to predict its complex structure is to use the symmetry method. Therefore, T10 was also not chosen in the test cases. We made use of the biological information on one protein by selecting 1 to 6 residues possibly at the binding site to constrain the sampling space. For T08, T11, T12, and T14–T18, all information[30–36] was obtained from articles that were published before the release of the crystal structures. For antigen–antibody complex T19, three residues of CDR3 in heavy and light chains were respectively assumed at the binding site. The residue numbers and types for each target are listed in Table I. The radius value of the sphere was set in the range of 15–25 Å. Table I compares the results obtained with the regular geometric docking method and

TABLE 1. Comparison of the Results of Biologically Enhanced and Regular Geometric Docking

	Regular geometric docking		Biologically enhanced geometric docking				
Targets	Hits[a]	L_RMSD of best[b]	Hits[a]	L_RMSD of best[b]	Rank[c]	L_RMSD[d]	Important residues selected
08	1	7.1	15	3.5	39	5.3	A73Asp, A75Asn, A77Val, A92Tyr[30]
11	10	5.5	20	2.3	25	9.3	B11Ser, B12Thr, B18Lys, B45Ser, B46Thr[31]
12	96	0.8	388	1.1	5	3.9	B11Ser, B12Thr, B18Lys, B45Ser, B46Thr[31]
14	0	19.0	3	4.5	1	8.7	B35Lys, B36Val, B37Lys, B38Phe[32]
15	7	8.4	25	7.6	6	9.9	A611His[33]
16	2	9.6	6	6.5	1	8.7	B84Trp, B128Glu, B274Trp[34,35]
17	0	11.7	4	6.6	118	9.6	B119Gln, B121Val, B122Asn, B129Thr, B131Thr, B133Asn[36]
18	0	13.5	53[e]	2.9[e]	15[e]	6.7[e]	B119Gln, B121Val, B122Asn, B129Thr, B131Thr, B133Asn[36]
19	4	6.6	131	3.2	8	7.6	H94Arg, H95Gly, H96Thr, L94Phe, L95Pro, L96Gln

[a]Number of hits in the top 10,000 predictions. Hits are defined as docked structures with ligand RMSD of backbone < 10.0 Å from the crystal complex.
[b]Ligand RMSD (Å) of the best structure in the top 10,000 predictions.
[c]Rank of the first hit in the solution list ranked by the scoring function [Eq. (1)].
[d]Ligand RMSD (Å) of the first hit in the solution list ranked by the scoring function [Eq. (1)].
[e]Results produced by the multiconformational superimposition method.

biologically enhanced sampling. In each case 10,000 docking solutions were retained, and hits were defined as docking solutions where the ligand RMSD is less than 10 Å. On all targets, biologically enhanced sampling geometric docking has many more hits than regular docking. Table I also cites the ligand RMSD values for the best solutions. They decrease in almost all cases, demonstrating that biologically enhanced sampling has the capacity to find more effective candidate solutions than the regular geometric docking procedure.

The largest improvement is obtained with T19, an antigen–antibody complex, where the number of hits increases from 4 to 131 and the ligand RMSD of the best solution decreases from 6.6 Å to 3.2 Å. Three reasons contribute to this improvement according to our analysis. In the first place, the two molecules have rather compact interface regions. It is relatively easy to choose a sphere radius to cover the whole interaction surface of the antibody. The effect is shown in Figure 1. In regular docking, the spheres representing the center of mass of the antigen scatter evenly around the antibody. In biologically enhanced sampling docking, they congregate around the binding site. The second reason is the way the antigen inserts deeply into the V-shaped cleft of the antibody in the crystal structure. This excellent geometric complementarity highly facilitates the geometric filtering. Last, conformation changes are limited in this system. The antibody is "bound" (taken from the crystalline complex) and the antigen's backbone RMSD was less than 2 Å relative to the crystal structure, though homology-modeled from an unbound NMR structure.

The advantage of the biologically enhanced sampling geometric docking method is that because active site data are incorporated during the translation–rotation scan rather than at the filtering stage, conformational searching can be done more effectively. In addition, there are two matters worthy of mention. One is about the selection of the important residues. Performing the biologically enhanced docking algorithm with centrally located residues in the interface will substantially improve docking results versus results obtained with residues on the fringes of the interface. The other is the determination of the radius of the sphere. For too large a radius, the restricting competency would be severely weakened. On the contrary, too small a radius may make the sampling insufficient.

Backbone Flexibility Based on Multiconformational Superimposition

In the case of T18, a xylanase–inhibitor complex, binding induces significant conformational changes of the "thumb" hairpin loop [circled in Fig. 2(a)] near the active site of the receptor (the xylanase). With this target, we had only one hit using biologically enhanced sampling geometric docking. To take the backbone flexibility of this region into account, we employed the multiconformational superposition approach. Two receptor structures were selected to be superimposed. One was the unbound crystal structure, and the other was a conformation taken from MD trajectories of the former, which has the largest backbone RMSD in the "thumb" region [see Fig. 2(a)]. The encouraging outcome of the procedure was 53 hits and the ligand RMSD reduced from 8.1 Å to 2.9 Å for the best solution, shown in Figure 2(b) alongside the crystal structure.

CONCLUSION

We have predicted the binding modes of nine CAPRI targets using the biologically enhanced sampling geometric docking method. This was successful in all cases, yielding many more near-native docking solutions and sharply reducing the ligand RMSD values of the best candidates. In the case of T18, we also considered backbone flexibility for the receptor and upgraded the performance of the docking procedure by applying the multiconformational superposition method.

Two great lessons learned from the recent CAPRI tests are that we should use as much biological information as

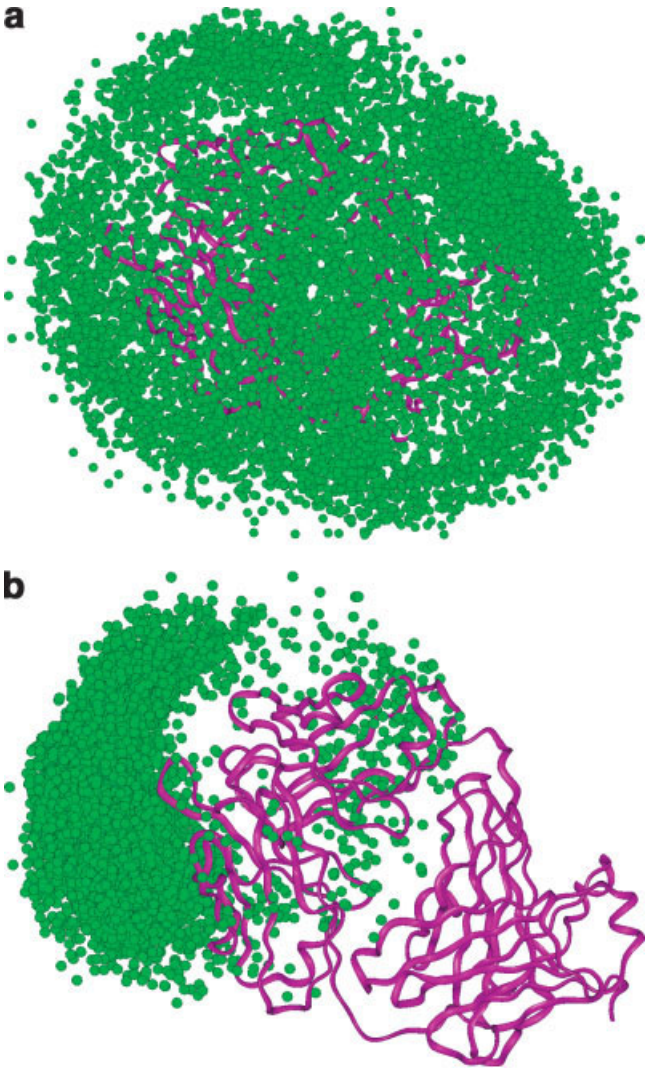

Fig. 1. Position of the geometric center (green sphere) of the antigen relative to the antibody (pink ribbon) in the top 10,000 predictions of Target 19 using the regular geometric docking algorithm (**a**) and the biologically enhanced geometric docking algorithm (**b**).

possible, and that handling protein flexibility is essential to successful prediction. Biological information can be obtained through various kinds of means, like the evolutionary trees method,[37] the sequence space method,[38] the self-organizing map method,[39] and so on. We show here that this information can be used at the sampling stage to yield more satisfactory results than later at the filtering stage (see Table I). On flexibility, the case of T18 illustrates the multiconformational superimposition method. This is only the first step, and further work is currently

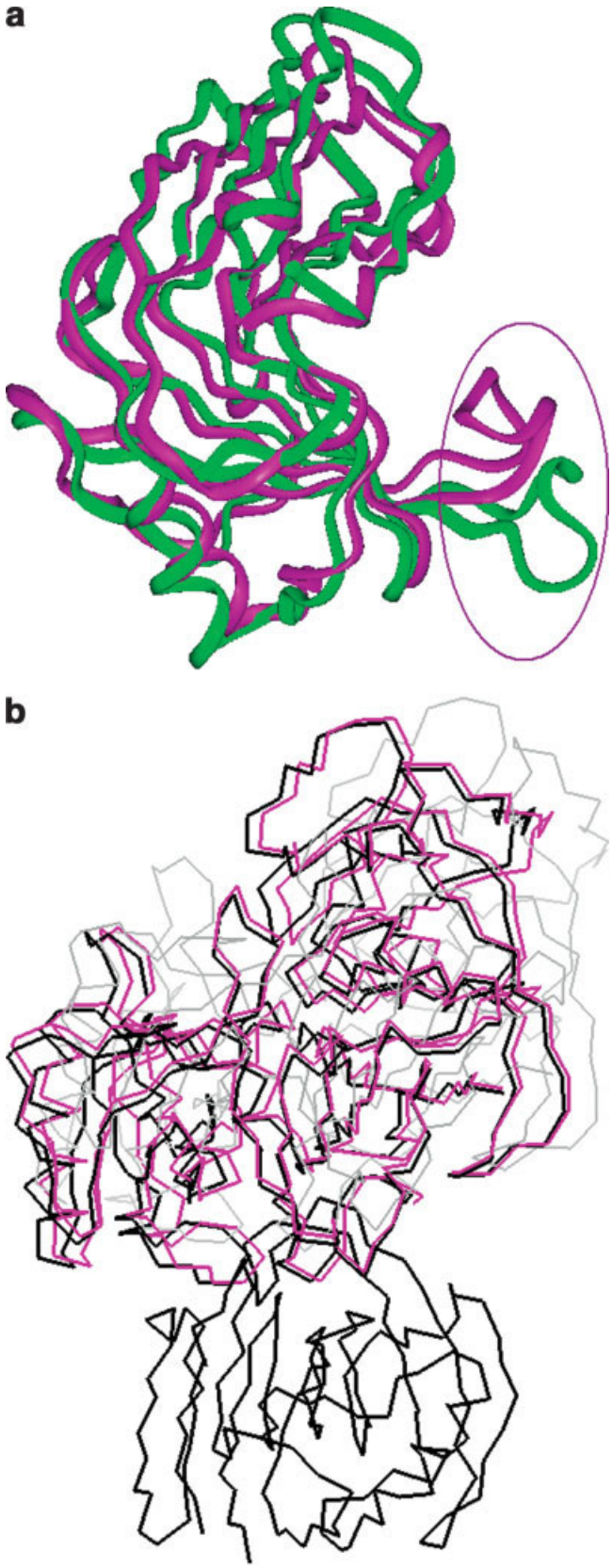

Fig. 2. (**a**) Superposition of the unbound crystal structure (pink ribbon) of the receptor xylanase of Target 18, with a conformation taken from the MD trajectory (green ribbon). Evident structural discrepancy is highlighted with the purple oval. (**b**) The crystal structure of Target 18 (black line) is superimposed with positions of the ligand obtained by biologically enhanced sampling geometric docking (gray), or by the multiconformational superposition method (pink line).

under way to represent the flexibility of the protein main-chain and side-chains.

ACKNOWLEDGMENTS

We thank the organizers and the evaluation team of CAPRI for their effort, the experimental groups for providing targets for this prediction exercise, and Drs. Ben-Zeev (Rehovot, Israel) and Yao-Qi Zhou (Buffalo, NY) for helpful discussion. We also thank Professor J. Janin for comments on the manuscript.

REFERENCES

1. Wodak SJ, Janin J. Computer analysis of protein–protein interaction. J Mol Biol 1978;124:323–342.
2. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJE, Vajda S, Vakser I, Wodak SJ. CAPRI: a critical assessment of predicted interactions. Proteins 2003;52:2–9.
3. Gabb HA, Jackson RM, Sternberg MJE. Modelling protein docking using shape complimentarity: electrostatics and biochemical information. J Mol Biol 1997;272:106–120.
4. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. Proteins 2003;52:80–87.
5. Abagyan RA, Totrov MM, Kuznetsov DA. ICM: a new method for structure modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comp Chem 1994;15:488–506.
6. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 2003;331:281–299.
7. Smith GR, Sternberg MJE. Evaluation of the 3D-DOCK protein docking suite in rounds 1 and 2 of the CAPRI blind trial. Proteins 2003;52:74–79.
8. Chen R, Tong WW, Mintseris J, Li L, Weng Z. ZDOCK predictions for the CAPRI challenge. Proteins 2003;52:68–73.
9. Fernández-Recio F, Totrov M, Abagyan R. ICM-DISCO docking by global energy optimization with fully flexible side-chains. Proteins 2003;52:113–117.
10. Gray JJ, Moughon SE, Kortemme T, Schueler-Furman O, Misura KMS, Morozov AV, Baker D. Protein–protein docking prediction for the CAPRI experiment. Proteins 2003;52:118–122.
11. Méndez R, Leplae R, Maria LD, Wodak SJ. Assessment of blind predictions of protein–protein interactions: current status of docking methods. Proteins 2003;52:51–67.
12. Ben-Zeev E, Eisenstein M. Weighted geometric docking: incorporating external information in the rotation–translation scan. Proteins 2003;52:41–46.
13. Ehrlich LP, Nilges M, Wade RC. The impact of protein flexibility on protein–protein docking. Proteins 2005;58:126–133.
14. Gunasekaran K, Buyong Ma, Nussinov R. Is allostery an intrinsic property of all dynamic proteins? Proteins 2004;57:433–443.
15. Abagyan R, Totrov M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 1994;235:983–1002.
16. da Silva RA, Degrève L, Calin A. LMProt: an efficient algorithm for Monte Carlo sampling of protein conformational space. J Biophys 2004;87:1567–1577.
17. Li CH, Ma XH, Chen WZ, Wang CX. A protein–protein docking algorithm dependent on the type of the complexes. Protein Eng 2003;16:256–269.
18. Heifetz A, Eisenstein M. Effect of local shape modifications of molecular surfaces on rigid-body protein–protein docking. Protein Eng 2003;16:179–185.
19. Shoichet BK, Kuntz ID. Protein docking and complementarity. J Mol Biol 1991;221:327–346.
20. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Nat Acad Sci USA 1992;89:2195–2199.
21. Takagi J, Yang YT, Liu JH, Wang JH, Springer TA. Complex between nidogen and laminin fragments reveals a paradigmatic beta-propeller interface. Nature 2003;424:969–974.
22. Carvalho AL, Dias FMV, Prates JAM, Ferreira LMA, Gilbert HJ, Davies GJ, Romao MJ, Fontes CMGA. Cellulosome assembly revealed by the crystal structure of the cohesin–dockerin complex. Proc Nat Acad Sci USA 2003;100:13809–13814.
23. Kerff F, Terrak M, Dominguez R. Protein phosphatase 1 targeting and regulation as revealed by the structure of a PP1–MYPT1 complex. to be published.
24. Graille M, Mora L, Buckingham RH, Van Tilbeurgh H, De Zamaroczy M. Structure inhibition of the colicin D tRNase by the tRNA-mimicking immunity protein. EMBO J 2004;23:1474–1482.
25. 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 Nat Acad Sci USA 2004;101:10254–10259.
26. Lindahl E, Hess B, van der Spoel D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. J. Mol Mod 2001;7:306–317.
27. 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.
28. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Nat Acad Sci USA 2000;97:10383–10388.
29. Chen R, Mintseris J, Janin J, Weng Z. A protein–protein docking benchmark. Proteins 2003;52:88–91.
30. Stetefeld J, Mayer U, Timpl R, Huber R. Crystal structure of three consecutive laminin-type epidermal growth factor-like (LE) modules of laminin γ1 chain harboring the nidogen binding site. J Mol Biol 1996;257:644–657.
31. Lytle BL, Volkman BF, Westler WM, Heckman MP, David Wu JH. Solution structure of a type I dockerin domain, a novel prokaryotic, extracellular calcium-binding domain. J Mol Biol 2001;307:745–753.
32. Egloff MP, Johnson DF, Moorhead G, Cohen PTW, Cohen P, Barford D. Structural basis for the recognition of regulatory subunits by the catalytic subunit of protein phosphatase 1. EMBO J 1997;16:1876–1887.
33. Masaki H, Ogawa T.The modes of action of colicins E5 and D, and related cytotoxic tRNases. Biochimie 2002;84:433–438.
34. Schmidt A, Gübitz GM, Kratky C. Xylan binding subsite mapping in the xylanase from *Penicillium simplicissimum* using xylooligosaccharides as cryo-protectant. Biochemistry 1999;38:2403–2412.
35. Flatman R, Russell Mclauchlan W, Juge N, Furniss C, Berrin JG, Hughes RK, Manzanares P, Ladbury JE, Obrien R, Willamson G. Interactions defining the specificity between fungal xylanases and the xylanase-inhibiting protein XIP-1 from wheat. Biochemistry 2002;365:773–781.
36. Tahir TA, Berrin JG, Flatman R, Roussel A, Roepstorff P, Williamson G, Juge N. Specific characterization of substrate and inhibitor binding sites of a glycosyl hydrolase family 11 xylanase from *Aspergillus niger*. J Biol Chem 2002;46:44035–44043.
37. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
38. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Biol 1995;2:171–178.
39. Andrade MA, Casari G, Sander C, Valencia A. Classification of protein families and detection of the determinant residues with an improved self-organizing map. Biol Cybern 1997;76:441–450.