PROTEINS: Structure, Function, and Genetics Suppl 5:92–97 (2001)
DOI 10.1002/prot.1169

Sequence-Structure Homology Recognition by Iterative Alignment Refinement and Comparative Modeling

M.G. Williams, H. Shirai, J. Shi, H.G. Nagendra, J. Mueller, K. Mizuguchi, R.N. Miguel, S.C. Lovell,* C.A. Innis, C.M. Deane, L. Chen, N. Campillo, D.F. Burke, T.L. Blundell, and P.I.W. de Bakker
Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom

ABSTRACT **Our approach to fold recognition for the fourth critical assessment of techniques for protein structure prediction (CASP4) experiment involved the use of the FUGUE sequence-structure homology recognition program (http://www-cryst.bioc.cam.ac.uk/fugue), followed by model building. We treat models as hypotheses and examine these to determine whether they explain the available data. Our method depends heavily on environment-specific substitution tables derived from our database of structural alignments of homologous proteins (HOMSTRAD, http://www-cryst.bioc.cam.ac.uk/homstrad/). FUGUE uses these tables to incorporate structural information into profiles created from HOMSTRAD alignments that are matched against a profile created for the target from multiple sequence alignment. In addition, environment-specific substitution tables are used throughout the modeling procedure and as part of the model evaluation. Annotation of sequence alignments with JOY, to reflect local structural features, proved valuable, both for modifying hypotheses, and for rejecting predictions when the expected pattern of conservation is not observed. Our stringency in rejecting incorrect predictions led us to submit a relatively small number of models, including only a low number of false positives, resulting in a high average score. Proteins 2001; Suppl 5:92–97.** © 2002 Wiley-Liss, Inc.

Key words: protein structure; structure prediction; homology recognition; environment-specific substitution tables; model evaluation; alignment databases

INTRODUCTION

Despite the rapid growth of the Protein Data Bank (PDB[1,2]), the rate of increase of sequence data continues to be greater, resulting in an ever-larger number of sequences that have no known three-dimensional (3-D) structure. Sequence-structure homology recognition often allows the identification of relationships between proteins and allows prediction of 3-D structures by comparative modeling.[3,4] In turn, model building allows the examination of the proposed 3-D structure, to determine whether it explains the available data. We treat models as hypotheses that may be accepted, rejected, or modified if they are inconsistent with the sequence or other experimental data.

Fold recognition techniques fall broadly into two categories: those that calculate the compatibility of a given sequence with a 3-D structure and those that aim to determine evolutionary relationships between proteins. Threading dominates the former,[5,6] whereas profile and hidden Markov model techniques, with[7–12] or without[13–17] the incorporation of structural information, are commonly used for the latter. It should be noted that although threading can genuinely be termed "fold recognition," evolutionary relationship techniques by definition operate at the "family" or "superfamily" rather than "fold" level. Proteins in the same family or superfamily are homologous (i.e., derived from a common ancestor); therefore, we use the term "homology recognition" to describe such approaches.

To investigate the relationships between homologous proteins, we have developed databases of structure-based alignments of protein families[18,19] (HOMSTRAD: http://www-cryst.bioc.cam.ac.uk/homstrad/) and superfamilies[20] (CAMPASS: http://www-cryst.bioc.cam.ac.uk/campass/). These databases have allowed the analysis of patterns of sequence conservation in the context of structural environments.

3-D structure places restraints on one-dimensional (1-D) sequences, resulting in non-uniform patterns of variation and conservation within protein families. It has been shown that environment-specific substitution tables, which take account of these differing structural environments,[21–23] offer a more accurate estimate of substitution patterns than traditional substitution tables.[24–27] Many structural features affect substitution patterns, including secondary structure, solvent accessibility, and hydrogen bonds. We use environment-specific substitution tables (derived from HOMSTRAD) throughout the modeling procedure. It is important that they are used in FUGUE,[23] which is a fully automatic program that was developed recently and has participated in CAFASP2[28] (http://www.cs.bgu.ac.il/~dfischer/CAFASP2/).

Grant sponsor: Wellcome Trust; Grant sponsor: BBSRC; Grant sponsor: Pfizer; Grant sponsor: Overseas Research Students Award Scheme; Grant sponsor: Cambridge Overseas Trust; Grant sponsor; Cambridge European Trust and Tanabe Seiyaku Co. Ltd.; Grant sponsor: Ministerio Español de Educatión y Cultura; Grant number: EX 99 09274532.

*Correspondence to: S.C. Lovell, Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge, UK CB2 1GA. E-mail: simon@cryst.bioc.cam.ac.uk

Received 12 March 2001; Accepted 28 June 2001

Quality control and verification are extremely important in fold recognition and model building. Environment-specific substitution tables are a valuable tool for assessing the accuracy of both sequence alignments and models. They can be used to indicate those environments that should result in conserved residues and allow adjustment if such residues are not conserved. Conversely, when models are assessed, substitution tables can indicate environments that should be conserved, for example, the loss of a side-chain to main-chain hydrogen bond is suspicious. Verification of models, followed by their acceptance, rejection, or iteration of the modeling and alignment process is a key feature of our approach. We aim not only to give the best model of a target structure but also, where appropriate, to indicate where a target sequence is not compatible with a given structure. Historically, these negative answers have been lacking in comparative modeling.

In this article we discuss the application of these methods to fold recognition in the CASP4 (http://predictioncenter.llnl.gov/casp4/Casp4.html) experiment. We describe successes and failures and link these to aspects of our approach. In particular, the stringent assessment as part of the modeling cycles resulted in a small number of false positives, and hence a relatively high average score for the models we submitted.

MATERIALS AND METHODS

We used a number of complementary stages. These may be summarized as follows: (a) find candidate hits from our sequence-structure homology recognition program FUGUE[23] and other sources, (b) examine and adjust the sequence alignment, annotated with JOY[29] to ensure optimal alignment of those structural features that are conserved in the family, (c) perform comparative model building, (d) perform critical assessment of the resulting model and, if necessary, (e) iteration. We have found the iteration and critical assessment aspects of this process to be of utmost importance.[4]

Finding Candidate Homologues

FUGUE[23] selects one of 64 different substitution tables, based on the local residue environment. The currently defined environments are solvent accessibility/inaccessibility, local backbone conformation (α-helix/β-sheet/coil/positive ϕ), and presence or absence of hydrogen bonds from the side-chain to the main-chain carbonyl, main-chain amide, or another side chain. These environment-specific substitution tables were derived by using the programs JOY and SUBST (K. Mizuguchi, unpublished. http://www-cryst.bioc.cam.ac.uk/~kenji/subst/) from the quality-assured structure-based alignments in the HOMSTRAD[18,19] database. Profile-profile matching between the target sequence and the HOMSTRAD database generates initial hits for homology recognition and alignments. We also made use of standard sequence analysis tools, such as PSI-BLAST[30] and CLUSTALW,[31] where homologues of known structure were sufficiently closely related to the targets to be recognized by these methods.

Alignment Verification and Adjustment

The alignments produced by FUGUE for the highest-scoring hits were formatted with JOY and analyzed visually to highlight the conservation of structurally important residues. At this stage, the following additional information was included: indications from the literature of residues that are possibly functionally or structurally conserved, both for the target and for the potential parents, results from other CAFASP2 servers, and secondary structure predictions from JPRED2[32] and PSIPRED.[33] Based on these analyses, we validated the choice of the parents and in many cases, manually optimized the alignment. Specifically, if any predicted function of the target sequence is inconsistent with observed functions of the candidate family of proteins, then the prediction was discarded or modified.

Comparative Modeling

Models were constructed with SCORE[34] or MODELLER[35] based on the parent structure(s) and the alignment. In both cases, the structurally variable regions, which normally corresponded to the loops were built (in the case of SCORE) or rebuilt (in the case of MODELLER) by using CODA[36] or SLOOP.[37–39] Side chains were generated by using SCWRL.[40]

Critical Assessment of the Model

Models were validated by PROCHECK,[41] VERIFY3D,[42] PROSA II,[43] and visual inspection by using 3-D graphics software, in some cases after adding hydrogens with REDUCE[44] and calculating interactions with PROBE.[45] It is important that the generated models were realigned to their parent structures by using COMPARER,[46] producing a structure-based alignment that was annotated with JOY. This allowed visual inspection of the conservation of both residues and their structural environments. On the basis of this re-examination, the model was either accepted, rejected, or the alignment was modified and the modeling process repeated. If there was no convergence, or if the process led to a model that does not seem to satisfy the available data, we rejected the hypothesis and chose not to submit a model.

RESULTS

We submitted a number of models to the CASP4 experiment, produced from blind prediction based only on the sequences. Of these, nine were subsequently assigned to the "fold recognition" category and are discussed here.

The assessors gave a score from 0 to 4, with 0 being an incorrect model and 4 representing a model with the correct fold and a good alignment. Additional information was made available to the predictors, which is summarized in Table I. We selected a few of these models to discuss in more detail.

Targets T0100 (Pectin Methylesterase) and T0101 (Pectate Lyase)

Both targets had their folds confidently assigned. However, because of the low sequence identities involved, the

TABLE I. Summary of Results

Target	Name	Assessors' score	Fugue rank	% equiv[a]	% aligned[b]
T0095_1	a-(E)-catenin domain 1	4	1[c]	46[d]	75[d]
T0095_2	a-(E)-catenin domain 2	4	1[c]	46[d]	75[d]
T0096_1	FadR	3	1	88	84
T0100	Pectin methylesterase	3.5	1	48	36
T0101	Pectate lyase	1	1	92	57
T0104	Hypothetical protein HI0065	0	1	38	40
T0108	Family 17 carbohydrate binding	2.5	1[c]	57	57
T0116_4	MutS	2.5	2	64	62
T0127_1	Magnesium cheletase	1.5	1	65	9

[a]Percentage of residues in our model assessed as structurally equivalent in a sequence independent evaluation.
[b]Percentage of structurally equivalent residues that are correctly aligned ("shift 0").
[c]See text for discussion.
[d]These are the combined values for T0095_1 and T0095_2.

alignments that were generated by the fold recognition servers were assumed to be unreliable. Information in the literature acquired from previously determined structures of proteins belonging to the pectate lyase family was used to aid the manual adjustment of the alignment. It is known that the cores of these structures consist of a repeating circular three β-stranded section (assigned the letters *a, b,* and *c* for clarity). In general, the first two β-strands (*a* and *b*) have a short section separating them, containing either an asparagine, serine, cysteine, or a threonine residue, which forms a hydrogen bond "ladder" to the equivalent section in the adjoining repeat, as the JOY annotated alignment illustrates (Fig. 1). The separation between strands *b* and *c* and between *c* and strand *a* of the following repeat are longer and often contain glycine or proline, because these parts need to form a sharp turn to enforce the cyclical nature of the structural motif. Finding the optimal alignment in both cases was greatly facilitated by using the secondary structure prediction programs JPRED2 and PSIPRED. This finding, along with the conserved side-chain to main-chain hydrogen bonding, was used to decide which part of the target formed what type of β-strand in the repeat. For target T0100, apart from a misassignment of a β-strand due to an incorrect secondary structure prediction near the N-terminus, the alignment breaks down only near the C-terminal part of the molecule where two predicted repeat units turn out to have an insertion region containing two antiparallel β-strands [Fig. 2(a)]. This prediction was given an assessors' score of 3.5. In target T0101, the assignment of these repeating motifs is incorrect; short spacers between one of the *b* and *c* strands and between *c* and the following strand *a* have resulted in the motif being misassigned [Fig. 2(b)]. This resulted in an out-of-register alignment for some of the model; this prediction received an assessors' score of 1.

T0104 (Hypothetical Protein HI0065)

Our prediction for this target received the score of zero from the assessors. The top ranking prediction from the automatic FUGUE server, with a Z-score of 4.2 (under 90% confidence) was for the HOMSTRAD family "Kinesin motor, catalytic domain ATPase." In SCOP, the "Motor proteins" family belongs to the "P-loop containing nucleotide triphosphate hydrolases" superfamily. From this superfamily, the structure of human ADP-ribosylation factor 1 (PDB code: 1hur), belonging to the "G proteins" family, was chosen as the template on the basis of the PSIPRED structure prediction server. However, because of the differences in topology between our prediction and the target structure, our model was poor. In fact, T0104 differs in topology from all of the known structures in the "P-loop" superfamily and was classed as a "fold recognition/new fold" target, which is the most difficult class in the "fold recognition" category.

Targets T0095_1 and T0095_2 (Domains One and Two of α-Catenin)

Both targets received the maximum score of 4 from the assessors. FUGUE and other CAFASP servers did not detect any significant hits at the time of release of the target. Subsequently, another domain of α-catenin was released from the PDB, and this was picked up by the 3D-PSSM[11] server. FUGUE also produced a Z-score of >6 (99% confidence) for this structure when HOMSTRAD was updated. The template structure was of a domain upstream from the target sequence, an obvious case of gene duplication. However, this structure exists in its native state as a dimer of four α-helical bundles. Like the target sequence, the structure does exist as a monomer (PDB code: 1dow chain A) when it interacts with a helical fragment of β-catenin. This helps form a second four-helical bundle and stabilizes the structure in its monomeric state, hence the inclusion of this part of the template structure (PDB code:1dow Chain B) in the alignment and model building of target T0095. Once the template had been identified and the role of the extra fragment of structure shown to be essential for the structural stability, it was straightforward to optimize manually the target sequence template structure alignment by using the secondary structure prediction for the target.

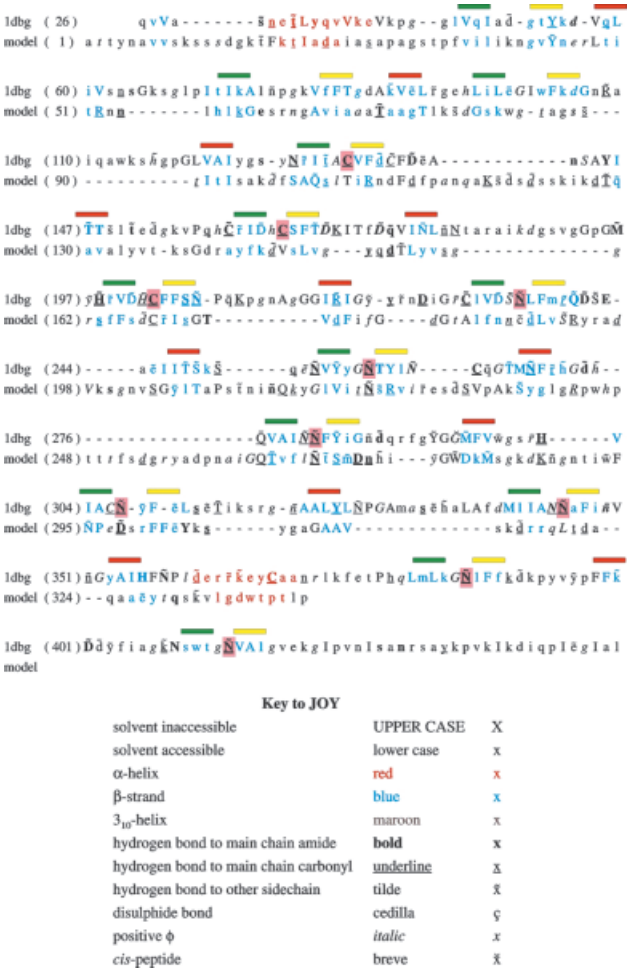

Fig. 1. Sequence alignment of target T0100 and the chosen parent chondroitinase B (PDB code 1dbg). Colored bars indicate β-strands (green for *a* strands, yellow for *b* strands, and red for *c* strands. Residues we judged to be important in adjusting the alignment are highlighted with pink boxes, although these are already given "decorations" by JOY, indicating environments in which residues are usually highly conserved (see key).

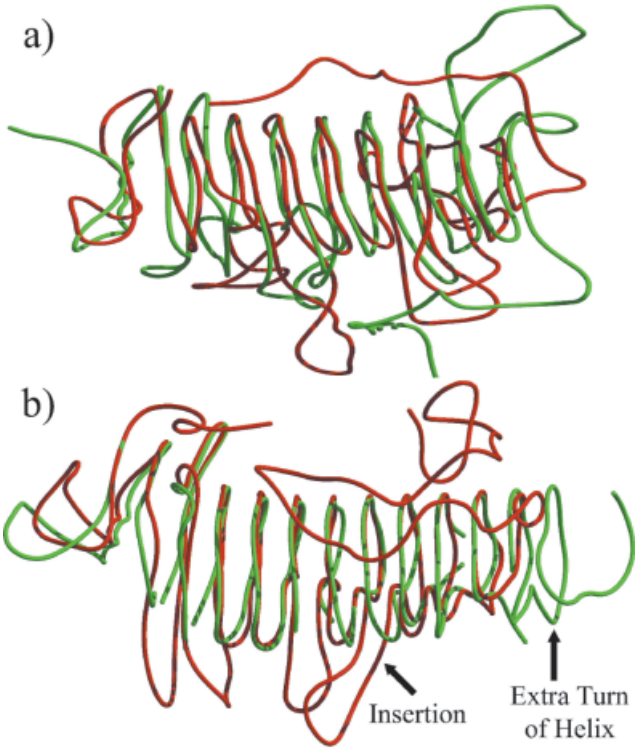

Fig. 2. Comparison of our model (green) and the experimentally determined structure of the target (red) for (**a**) target T0100 and (**b**) T0101. In (b) the position of the misassigned insertion is indicated. This resulted in all residues C-terminal of this position being misaligned, and an extra turn of the β-helix being built.

Target T0108 (Family 17 Carbohydrate Binding-Module)

Our model for this target had a reasonable score of 2.5 (only one other competitor in CASP4 received a higher score of 3). FUGUE initially produced insignificant hits (Z-scores < 3.0). The fact that the target is a C-terminal carbohydrate-binding module of a cellulase enzyme prompted us to perform further sequence analyses, including a PSI-BLAST search using the full-length sequence, and examination of the collected homologues. The homologous sequences were realigned with CLUSTALW and resubmitted to FUGUE. This time the search produced a reasonably high Z-score (4.12) for the top hit, corresponding to the HOMSTRAD family of CBD_6 (cellulose binding domain). The functional similarity was obvious and encouraging. An alignment generated by FUGUE between the target sequence and the structurally aligned members of the family (PDB codes: 1ulo and 1cx1a) was largely consistent with the PSIPRED and JPRED2 secondary structure predictions. Several major modifications had to be made manually, however, to optimize the gap positions and consistency with the secondary structure prediction. Annotation with JOY aided this adjustment.

Based on this alignment, an initial model was built with MODELLER. Examination of a superposition of the model and the templates suggested several additional restraints on hydrogen bonds. The alignment was re-examined by using JOY and slightly altered. The model quality was assessed with PROSA II to identify problematic regions. Several rounds of model building led to a model with reasonable quality (PROSA Z-score $= -6.09$). This result was judged to support our prediction.

DISCUSSION

We submitted relatively few models, but we had quite a high success rate with those that we did submit. This resulted in our total score being respectable, but our average score being high. We believe that this is a strength of our method; not only are models proposed, they are stringently assessed and, if necessary, rejected.

An important aspect of all types of protein structure prediction is the assessment of confidence of the prediction. This aspect is sometimes ignored completely or not considered explicitly. For example, in CASP, only two levels of confidence are implicitly used: "confident," and a

prediction is submitted, or "not confident" and therefore no submission is made. More tentative suggestions can be made by submitting more than one model, but these are not usually assessed because of the increase in workload that this would entail for the assessors. It is our future aim to make quantitative analyses of the confidence of all aspects of our predictions, both for comparative modeling and sequence-structure homology recognition, to allow users to determine accurately the quality of the models we produce.

The list of potential templates and the initial alignments produced by FUGUE, using local structural information, are essential for developing initial hypotheses. Annotating protein sequence alignments with structural information, using JOY, makes structurally important features within a family of proteins easy to identify, and their conservation is a good indication of a correct prediction. Furthermore, identification and conservation of functionally important residues improves the confidence in a correct prediction and an accurate alignment. Indeed, some templates were identified on the basis of their function, which in some cases strongly suggests a given fold. However, as two protein structures diverge, the percentage of residues that can be classified as being structurally equivalent drops. The sequence identity of those structurally aligned residues also drops, giving a sequence identity over all residues as low as 10%. In regions where structures have diverged, a sequence alignment, which represents expected local equivalences, may have little meaning. In such a case, it is not possible to infer confidently structural environments from one family member to another.

This raises interesting questions about the limitation of fold/homology recognition and comparative modeling. Sippl maintains a view of protein structures as discrete but overlapping fold space, that is, partial similarity to two known structures produces a similar but distinct protein fold.[47] Current methods essentially rely on identifying homology between the target and one or more parents and copying those regions of high sequence similarity. In many cases, however, the structures have diverged sufficiently that the closest parent has, in fact, a similar but distinct fold. This problem of structural divergence goes to the heart of structure prediction from evolutionary relationships. Knowledge of the likelihood of similarity of local environments for individual amino acids (e.g., as determined by the SCORE program) is highly advantageous. Assessment of confidence on a residue-by-residue basis would also identify the structurally divergent regions, although prediction of the conformation of these divergent regions is currently very inaccurate. These inaccuracies may be due to rigid body movements of domains or secondary structural elements,[48] or more localized backbone conformational changes.

FUGUE performed well in CAFASP2,[28] where it operated without human intervention. However, the consensus of servers, which outperformed all individual servers, was still well down the field compared with the human participants. Clearly, human intervention is adding a great deal of value. This can come in a number of forms, such as knowledge of the literature and homologous families and superfamilies. This information is of enormous help and indeed many people correctly assigned the fold of pectin methylesterase and pectate lyase (T0100 and T0101) from the names alone. This sort of intuition must be rigorously checked to ensure it does not lead us astray, but it also gives enormous advantage when trying to decide if a hypothesis (model) is likely to be correct. In addition, the visual nature of JOY annotation works well with human skills of pattern recognition and is a key element in our handling of hypotheses. Healthy skepticism seems to be a human attribute that is particularly difficult to automate.

ACKNOWLEDGMENTS

The authors thank the following for funding: the Wellcome Trust, the BBSRC, Pfizer, the Overseas Research Students Award Scheme, the Cambridge Overseas Trust, the Cambridge European Trust and Tanabe Seiyaku Co. Ltd. R.N.M. thanks the "Ministerio Español de Educación y Cultura" for a postdoctoral grant (EX 99 09274532). H.G.N. is a BOYSCAST fellow sponsored by the Department of Science and Technology, Government of India.

REFERENCES

1. Bernstein FC, Keotzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
3. Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM. Knowledge-based prediction of protein structures and the design of novel molecules. Nature 1987;326:347–352.
4. Burke DF, Deane CM, Nagarajaram HA, Campillo N, Martin-Martinez M, Mendes J, Molina F, Perry J, Reddy BV, Soares CM, Steward RE, Williams M, Carrondo MA, Blundell TL, Mizuguchi K. An iterative structure-assisted approach to sequence alignment and comparative modeling. Proteins 1999;Suppl 3:55–60.
5. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J Mol Biol 1990;216:167–180.
6. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
7. Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. Protein Sci 1996;5:947–955.
8. Hargbo J, Elofsson A. Hidden Markov models that use predicted secondary structures for fold recognition. Proteins 1999;36:68–76.
9. Johnson MS, Overington JP, Blundell TL. Alignment and searching for common protein folds using a data bank of structural templates. J Mol Biol 1993;231:735–752.
10. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
11. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
12. Rice DW, Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 1997;267:1026–1038.
13. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
14. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14: 755–763.
15. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detec-

tion of distantly related proteins. Proc Natl Acad Sci USA 1987;84:4355–4358.
16. Karplus K, Barrett C, Hughley R. Hidden Markov models for detecting remote protein homologues. Bioinformatics 1998;14:846–856.
17. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles: strategies for structural predictions using sequence information. Protein Sci 2000;9:232–241.
18. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci 1998;7:2469–2471.
19. de Bakker PIW, Bateman A, Burke DF, Miguel RN, Mizuguchi K, Shi J, Shirai H, Blundell TL. HOMSTRAD: adding sequence information structure-based alignments of homologous protein families. Bioinformatics 2001;17:748–749.
20. Sowdhamini R, Burke DF, Huang JF, Mizuguchi K, Nagarajaram HA, Srinivasan N, Steward RE, Blundell TL. CAMPASS: a database of structurally aligned protein superfamilies. Structure 1998;6:1087–1094.
21. Overington J, Johnson MS, Sali A, Blundell TL. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proc R Soc Lond B Biol Sci 1990;241:132–145.
22. Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci 1992;1:216–226.
23. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001; 310:243–257.
24. Dayhoff MO, Schwartz RM, Orcutt BC. Atlas of protein sequence and structure. National Biomedical Research Foundation; 1978. p 345–358.
25. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein sequence database. Science 1992;256:1443–1445.
26. Henikoff S, Henikoff JG. Amino-acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
27. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 1992;8:275–282.
28. Fischer D, Elofsson A, Rychlewski L, Pazos F, Valencia A, Rost B, Ortiz AR, Dunbrack RL. CAFASP2: the second critical assessment of fully automated structure prediction methods. Proteins 2001; Suppl 5:171–183.
29. Mizuguchi K, Deane CM, Blundell TL, Johnson MS, Overingon JP. JOY: protein sequence-structure representation and analysis. Bioinformatics 1998;14:617–623.
30. Altschul SF, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database programs. Nucleic Acid Res 1997; 25:3389–3402.
31. Higgins DG, Thompson JD, Gibson TJ. Using CLUSTAL for multiple sequence alignments. Methods Enzymol 1996;266:383–402.
32. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. JPred: a consensus secondary structure prediction server. Bioinformatics 1998;14:892–893.
33. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
34. Deane CM, Kaas Q, Blundell TL. SCORE: predicting the core of protein models. Bioinformatics 2001;17:541–550.
35. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
36. Deane CM, Blundell TL. CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 2001;10:599–612.
37. Donate LE, Rufino SD, Canard LHJ, Blundell TL. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. Protein Sci 1996;5:2600–2616.
38. Rufino SD, Donate LE, Canard LHJ, Blundell TL. Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modelling. J Mol Biol 1997;267:352–367.
39. Burke DF, Deane CM, Blundell TL. Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure. Bioinformatics 2000;16:513–519.
40. Bower MJ, Cohen FE, Dunbrack RL Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 1997;267:1268–1282.
41. Laskowski RA, Macarthur MW, Moss DS, Thornton JM. PROCHECK—a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283–291.
42. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with 3-dimensional profiles. Nature 1992;356:83–85.
43. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins 1993;17:355–362.
44. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 1999;285:1735–1747.
45. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. Visualizing and quantifying molecular goodness of fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol 1999;285:1711–1733.
46. Sali A, Blundell TL. Definition of general topological equivalence in protein structures: a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. J Mol Biol 1990;212:403–428.
47. Domingues FS, Koppensteiner WA, Sippl MJ. The role of protein structure in genomics. FEBS Lett 2000;476:98–102.
48. Reddy BVB, Nagarajaram HA, Blundell TL. Analysis of interactive packing of secondary structural elements in alpha/beta units in proteins. Protein Sci 1999;8:573–586.