Pacific Symposium on Biocomputing 5:176-187 (2000)

Probing structure-function relationships of the DNA polymerase alpha-associated zinc-finger protein using computational approaches

RAM SAMUDRALA [a], YU XIA, MICHAEL LEVITT
Department of Structural Biology, Stanford University School of Medicine, Stanford CA 94305

NAOMI J COTTON
Department of Chemistry and Biochemistry, University of California at Santa Cruz, Santa Cruz CA 95064

ENOCH S HUANG
Cereon Genomics LLC, 45 Sidney Street, Cambridge MA 02139

RALPH DAVIS
Department of Pathology, Stanford University School of Medicine, Stanford CA 94305

We present the application of a method for protein structure prediction to aid the determination of structure-function relationships by experiment. The structure prediction method was rigourously tested by making blind predictions at the third meeting on the Critical Assessment of Protein Structure methods (CASP3). The method is a combined hierarchical approach involving exhaustive enumeration of all possible folds of a small protein sequence on a tetrahedral lattice. A set of filters, primarily in the form of discriminatory functions, are applied to these conformations. As the filters are applied, greater detail is added to the models resulting in a handful of all-atom "final" conformations. Encouraged by the results at CASP3, we used our approach to help solve a practical biological problem: the prediction of the structure and function of the 67-residue C-terminal zinc-finger region of the DNA polymerase alpha-associated zinc-finger (PAZ) protein. We discuss how the prediction points to a novel function relative to the sequence homologs, in conjunction with evidence from experiment, and how the predicted structure is guiding further experimental studies. This work represents a move from the theoretical realm to actual application of structure prediction methods for gaining unique insight to guide experimental biologists.

1 Introduction

The prediction of three dimensional protein structure from sequence with accuracy rivalling that of experiment is an unsolved problem. However, for certain classes of small globular proteins without homologs of known structure, it is possible in some cases to computationally build low resolution models

[a] Corresponding author; e-mail: ram@csb.stanford.edu

($\approx$ 6 Å C_α root mean square deviation of the coordinates (cRMSD) from the experimental structure)[1,2,3,4,5]. Given the large number of sequences being determined and the relatively slow progress of protein structure determination methods, low resolution models generated by current approaches can be used to elucidate details and yield valuable insight about the structure and function for proteins whose atomic structure has not been determined experimentally.

We have used a combination of approaches described in the literature, and primarily developed in-house, to construct tertiary models of protein sequences that have the correct topological arrangement of secondary structure elements. The hierarchical approach was tested rigourously by making blind predictions for thirteen proteins at the third meeting on the critical assessment of protein structure prediction methods (CASP3), with encouraging results [4].

The focus of this work is to move forward to the next step of using predicted structure for predicting function. We describe how we applied the combined approach to predict the structure of the 67-residue C-terminal zinc-finger region of the DNA polymerase-alpha associated zinc-finger (PAZ) protein, and how we used the predicted model to explore its function, simultaneously guiding and guided by experiment. The combined theoretical and experimental evidence points to a novel function for this protein compared to its sequence homologs. We discuss the implications of this type of approach for exploring structure-function relationships in a large-scale automated manner.

2 Methods

2.1 Summary of the combined hierarchical approach

For a given target protein, all possible self-avoiding compact conformations were exhaustively enumerated using a tetrahedral lattice model [6,7]. The computation is made tractable by reducing the chain length to no more than 50 lattice vertices (with two to three residues per vertex, depending on the size of the protein) and the degrees of freedom (three). This procedure yielded 10 million to 10 billion lattice conformations, and of these up to 40,000 best scoring conformations were selected using a simple lattice-based pairwise scoring function [7].

All-atom models were constructed by "fitting" the predicted secondary structure to the best-scoring lattice models. The secondary structure prediction was accomplished by generating twenty multiple sequence alignments of a homologous set of sequences to the target protein (using a bootstrapping procedure) and using them as input for three previously published secondary structure prediction methods: PHD [8], DSC [9], and Predator [10]. The consen-

sus of the twenty predictions for each method was used to assign helical and sheet residues where all three methods agreed. A greedy off-lattice build up procedure with a 4-state (ϕ/ψ) representation (one state helix, one sheet, two other) [11] was used minimise the cRMSD between the lattice model and the all-atom model taking into account predicted helix and sheet assignments. The most frequently observed rotamer values in protein structures were used for constructing side chains. The all-atom models were refined by applying 200 steps of steepest descent minimisation using ENCAD [12,13,14,15].

Three subsets consisting of the best 50, best 100, and best 500 all-atom conformations from the set of all-atom models were selected by a combined scoring function. The combined function consisted of an all-atom distance-dependent conditional probability discriminatory function (RAPDF) [16], a simple residue-level pairwise contact function (Shell) [17], and a hydrophobic compactness function (HCF) [3]. The most frequently observed C_α-C_α distances in each of the three subsets were used as constraints to a distance geometry procedure (by the TINKER software suite) [18] to generate up to 36 models. Predicted secondary structures were once again fitted to the consensus distance geometry models, and the models refined and scored by the all-atom (RAPDF) function. Detailed descriptions of the individual components of the combined hierarchical approach are given elsewhere [3,4,5].

For the initial test set of twelve proteins, only the final conformation was used to evaluate the results. For CASP3 predictions, four lowest scoring conformations after the consensus distance geometry procedure, and the lowest scoring conformation from the set of $\approx$ 40,000 as evaluated by RAPDF, were submitted as final models. The best model (out of five that were submitted) was used to evaluate the results.

2.2 Predicting the structure and function of the PAZ protein

Figure 1 shows the sequence for the C-terminal zinc-finger region of the PAZ protein, along with the predicted secondary structure using the PSIPRED secondary structure prediction server[19], and a multiple-sequence alignment to a family of homologous zinc-finger proteins. The homologous family is the ARFGAP sequence family, which has been found to play a role as a coatomer in GTP hydrolysis involved in vesicle formation during transport of proteins between intra-cellular compartments within an eukaryotic cell. The PAZ sequence is particularly interesting given the presence of the human homologs, all of which are classified as "hypothetical proteins" in SWISS-PROT[20]. The PSIPRED secondary structure prediction method was chosen because of its performance at CASP3.

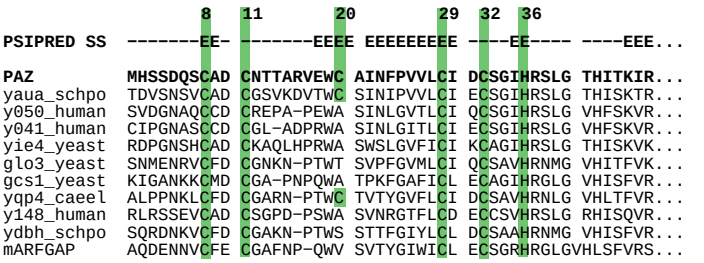

Figure 1: PAZ multiple sequence alignment with predicted secondary structure from the PSIPRED server[19]. The related sequences are labelled with their SWISS-PROT[20] identifiers. Residues thought to be important in coordinating zinc in our predicted model are highlighted.

Encouraged by the results of our approach both in our initial test and at CASP3, 67 residues from the C-terminal region of PAZ, chosen based on the consensus of residues observed in the multiple sequence alignment, were predicted in an *ab initio* manner using the approach described above. We visually examined the five lowest scoring all-atom conformations before and after the consensus distance geometry step (total of ten). Based on this visual analysis, we selected the second lowest scoring conformation from the set of 40,000 models prior to the consensus distance geometry step for detailed functional studies (see "Analysis of lowest scoring conformation" in the Results section for why this conformation was chosen). All further functional analyses were performed on this model using interactive graphics.

2.3 Summary of experimental studies

The prototypical DNA polymerase alpha-primase complex is composed of four different gene products; the DNA polymerase catalytic subunit, two polypeptides involved in a primase activity, and a fourth subunit with no proven catalytic activity referred to as the B subunit. We purified the DNA polymerase alpha-primase complex from the fission yeast *S. pombe*. The polymerase alpha-primase complex fractionated into two complexes: One was the DNA polymerase catalytic subunit complexed with the two primase subunits. The other complex comprised of a truncated form of the DNA polymerase catalytic subunit, an immunologically distinct 100 kDa polypeptide (ergo, PAZ,

for polymerase alpha-associated zinc-finger protein), the B subunit, and the two primase subunits. Direct comparisons of a yeast strain without the PAZ protein vs. the wild-type strain shows several biochemical and cell cycle differences: The PAZ deletion strain has an S phase perturbation. Purification of the DNA polymerase alpha complex from the PAZ deletion strain yields virtually none of the truncated catalytic subunit. Also, in the PAZ deletion strain, a large fraction of the primase subunits are not tightly associated with the DNA polymerase alpha complex, in contrast to the wild-type strain.

3 Results and discussion

3.1 Summary of model construction using the combined hierarchical approach

For 14/25 proteins, we were able to identify the correct topology of the protein or a significant fraction of the protein ($\approx$ 60 residues) and produce conformations that are $\approx$ 6.0 Å to the experimental structure. For 18/25 proteins, we sampled the conformational space adequately to ensure that a conformation representing the correct topology was available in the sample space. The correct topologies were sampled and identified even in cases where the secondary structure assignments were not very accurate. There is no clear dependence of success on protein size, but the method works better on α-helical proteins compared to β-sheet proteins. Detailed discussion of these results is given elsewhere [3,4].

3.2 Analysis of the lowest scoring conformations for the PAZ sequence

All five lowest scoring models after the consensus distance geometry procedure yielded similar structures, containing an α-helix at the N- and C-terminal ends, and a zinc-finger motif (Figure 2a). Among the five lowest scoring conformations from the set of 40,000 (before the consensus distance geometry step), the second lowest scoring conformation as evaluated by the RAPDF had the lowest average cRMSD to the consensus distance geometry models. This model was used for further structural and functional analyses, since the consensus distance geometry models do not have regular secondary structures and contain only C_α atoms.

The zinc-finger motif region spans residues 7-37 in the PAZ model, and involves cysteines 8, 11, 29, and 32, as would be expected from the multiple sequence alignment (Figure 1). However, the predicted structure reveals two additional residues, the non-conserved cysteine 20 (C20) and the conserved histidine 36 (H36), interacting with these four cysteines (Figures 2b and 2c).

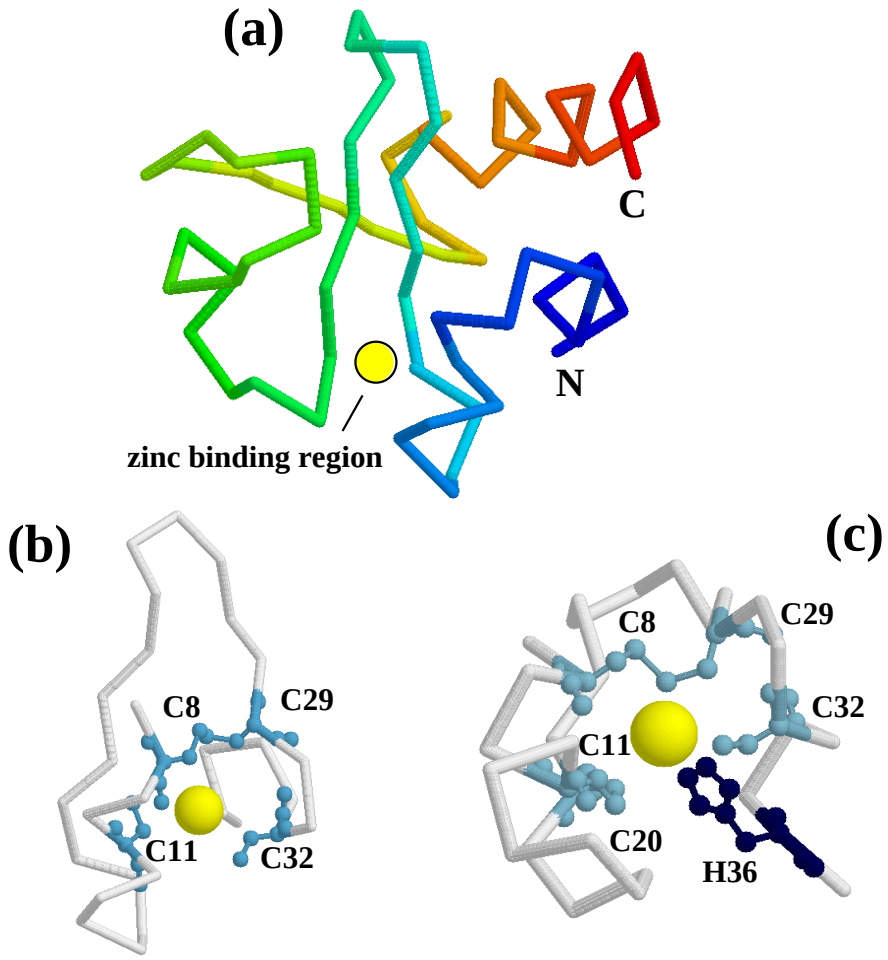

Figure 2: Illustrations of the PAZ model. Shown are (a) the entire 67-residue region *ab initio* prediction coloured by the direction of the chain, (b) the zinc-finger region (residues 7-37) with only the conserved cysteines coordinating a zinc atom, (c) the zinc-binding site (looking up the finger) with C20 and H36 interacting with the conserved cysteines and the zinc atom.

C20 is seen only in the similar proteins from *S. pombe* (the same organism with the PAZ sequence) and *C. elegans*.

3.3 Comparison to the similar ARFGAP sequence with known structure

During the initial sequence analysis, and structure prediction, we did not discover any homolog to the PAZ sequence with known structure. After the *ab initio* prediction was made, a structure of the mouse ARFGAP (mARFGAP) sequence was published [21]. The sequence of the mARFGAP is 28% identical to the PAZ sequence for the 67 residues and 16% identical for the region around the zinc-finger domain (Figure 1). We obtained the coordinates from the author and compared it visually to our predicted model.

The related experimental structure of mARFGAP and the PAZ model superpose to a cRMSD of 7.9 Å overall (5.9 Å for the fifth best scoring structure), 4.1 Å for the zinc-finger motif region, and 1.5 Å for the four cysteine residues. The excellent superposition of the cysteines coordinating the zinc atom in known structure corroborates our structural prediction (Figure 3). However, if a model is constructed based on the sequence similarity between PAZ and mARFGAP (i.e., by comparitive modelling methods), it would indicate a non-functional role for C20 and H36. Given (*i*) the low level of overall sequence identity (28%); (*ii*) the low level of sequence identity around the zinc-finger region (16%); (*iii*) the divergence of conformations at these levels (as high as 6.0 Å [22,23]), (*iv*) the presence of the cysteine in the related protein from the same species (Figure 1), and (*v*) the experimental evidence that indicates a novel function, we feel this is a case where the comparative model is not complete and that there is merit to performing mutagenesis experiments involving these two residues.

3.4 Functional role for the PAZ protein

It would appear, on the surface, based on the sequence relationships alone, that the PAZ sequence is a zinc-finger protein which is involved in vesicle formation and protein transport. However, the predicted structural and experimental data indicate otherwise: The association of PAZ with DNA polymerase alpha complex during purification, the perturbation of the S phase in the cell cycle when the PAZ protein is deleted, the lack of truncation of the catalytic subunit of the DNA polymerase alpha complex without the PAZ protein suggests a role that is different from protein transport and involvement in DNA replication and/or S-phase progression. The predicted structure of PAZ suggests a functional role for residues C20 and H36 because of their interaction with the conserved cysteines forming the zinc cluster (Figure 2). Experiments to test

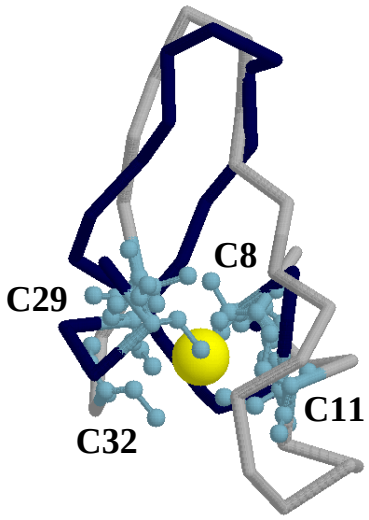

Figure 3: Comparison of the zinc-finger motif region for the PAZ and mARFGAP structures. The mARFGAP protein (black) shares a 16% sequence identity with the PAZ sequence for 30-residue region displayed and the cRMSD for the regions is 4.1 Å between the model and the homolog experimental structures. The four cysteine residues in each of the two structures are located at similar locations and superpose to an RMSD of 1.5 Å. The prediction was made without any knowledge of the homolog's structure.

whether these residues are important for the structure and function of PAZ are currently ongoing.

3.5 Does PAZ form a bi-nuclear zinc cluster?

An intriguing hypothesis is whether or not C20 and H36 help enable the coordination of an additional zinc atom, forming a binuclear zinc cluster as observed in the DNA binding domain of the yeast transcription factor, GAL4. There are two primary reasons for even considering this hypothesis: (*i*) from visual inspection of how C8, C11, C20, C29, C32, and H36 interact in the predicted structure, and comparing it to the GAL4 experimental structure, the putative coordination of the two zinc atoms is remarkably similar (see Figure 4), and (*ii*) the experimental evidence suggests a role for PAZ in DNA replication. If this hypothesis is true, the PAZ protein and homologous sequences would represent a novel family of binuclear zinc-containing motifs. Experiments to test this hypothesis and to see if the PAZ sequence binds DNA are also underway.

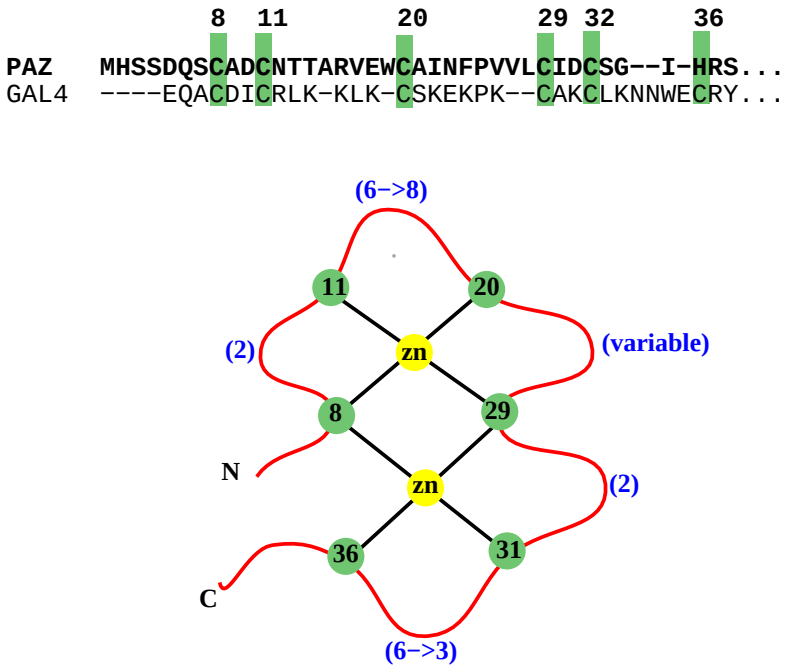

Figure 4: Model for how PAZ could form a binuclear zinc cluster similar to the one observed in the DNA-binding domain of the GAL4 transcription factor. Shown is an alignment between the GAL4 and PAZ sequences for the region of interest and a schematic diagram illustrating how the C8, C11, C20, C29, C32 and H36 could coordinate two zinc atoms. The change, if any, in the number of residues involved in the loops between C11 and C20, and C31 and C36, in PAZ relative to GAL4, are indicated by arrows.

3.6 Predictive power of this approach

As would be expected from the results based on the initial test set, the method should predict conformations to about 6.0 Å roughly capturing the topology for proteins/fragments of length 60 for slightly more than half the proteins modelled. This is borne out by the results from the blind prediction experiments. Besides evaluating the general accuracy of the method as a measure of the quality of a given prediction, we have two primary reasons to believe that the PAZ model is fairly accurate, especially in the functional region: (*i*) the consensus distance geometry models are similar to each other, suggesting that the lowest scoring structures have similar inter-atomic distances (*ii*) the zinc-finger motif with the four conserved cysteines coordinating the zinc atom

(Figure 2) is modelled extremely well (considering this was done in a purely *ab initio* manner), and (*iii*) there is a structural similarity, particularly in the zinc-finger region (cRMSD of 4.1 Å) and the position of the cysteines, between our prediction and the mARGAP experimental structure (a distant homolog).

The question then becomes, how useful is this rough model for predicting function? While it is clear that rough models cannot be used directly for rational drug design and other functional studies that require high-resolution models[24], the model we have built for the PAZ sequence has been useful in guiding mutagenesis studies and corroborating experimental data.

3.7 Applicability of this approach to other (large-scale) problems

While the focus of this paper is on one protein, we have applied this approach using a combination of theoretical and experimental data, guided by intuition, to attempt to predict structure/function relationships of three other proteins. These have produced similar results which are being used to guide experiments. This indicates that we have developed tools that, when used carefully in the hands of a structural biologist, can help elucidate function in a rational manner. In general, our work represents an important step of moving from pure prediction of structure to actually suggesting experiments to wet-lab biologists. As a result, there can be iterative improvement of our methodologies: as we codify the intuitions and heuristics we use, it may be possible to automate the function-prediction step further.

3.8 Availability of test sets and software

The ensembles of structures that were generated and much of the software used to generate them are available at <`http://dd.stanford.edu`> and <`http://www.ram.org/computing/ramp/ramp.html`>, respectively. The TINKER software suite is available at <`http://dasher.wustl.edu/tinker/`>.

Acknowledgments

We are extremely grateful to Patrice Koehl for providing us with efficient FORTRAN source code to construct protein models given a set of $\phi/\psi/\chi$ angles and to calculate the best-fit cRMSD between conformations, and to Jay Ponder for TINKER and helpful advice on its application. This work was supported in part by a Burroughs Wellcome Fund Postdoctoral Fellowship awarded by the NSF Program in Mathematics and Molecular Biology to Ram Samudrala, a Howard Hughes Medical Institute Predoctoral Fellowship to Yu Xia, a Jane Coffin Childs Memorial Fund Fellowship to Enoch Huang, NIH

Grant CA 14835 and CA 54415 to Teresa Wang for the support of Ralph Davis, and NIH Grant GM 41455 to Michael Levitt.

References

1. K.T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.*, 268:209–225, 1997.
2. A. Ortiz, A. Kolinski, and J. Skolnick. Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments. *J. Mol. Biol.*, 277:419–448, 1998.
3. R. Samudrala, Y. Xia, M. Levitt, and E.S. Huang. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. In R.B. Altman, A.K. Dunker, L. Hunter, T.E. Klein, and K. Lauderdale, editors, *Proceedings of the Pacific Symposium on Biocomputing*, pages 505–516, 1999.
4. R. Samudrala, Y. Xia, E.S. Huang, and M. Levitt. *Ab initio* protein structure prediction using a combined hierarchical approach. *Proteins: Struct. Funct. Genet. (in press)*, 1999.
5. Y. Xia, E.S. Huang, M. Levitt, and R. Samudrala. *Ab initio* construction of protein tertiary structures with a hierarchical approach. *In preparation*, 1999.
6. D.A. Hinds and M. Levitt. A lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci. USA*, 89:2536–2540, 1992.
7. D.A. Hinds and M. Levitt. Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.*, 243:668–682, 1994.
8. B. Rost and C. Sander. Prediction of protein structure at better than 70% accuracy. *J. Mol. Biol.*, 232:584–599, 1993.
9. D. Ross and M.J.E. Sternberg. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.*, 5:2298–2310, 1996.
10. D. Frishman and P. Argos. Knowledge-based secondary structure assignment. *Proteins: Struct., Funct., Genet.*, 23:566–579, 1995.
11. B. Park and M. Levitt. The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.*, 249:493–507, 1995.
12. M. Levitt and S. Lifson. Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.*, 46:269–279, 1969.

13. M. Levitt. Energy refinement of hen egg-white lysozyme. *J. Mol. Biol.*, 82:393–420, 1974.
14. M. Levitt. Molecular dynamics of native protein. *J. Mol. Biol.*, 168:595–620, 1983.
15. M. Levitt, M. Hirshberg, R. Sharon, and V. Daggett. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comp. Phys. Comm.*, 91:215–231, 1995.
16. R. Samudrala and J. Moult. An all-atom distance dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.*, 275:895–916, 1998.
17. B. Park, E.S. Huang, and M. Levitt. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.*, 266:831–846, 1997.
18. E.S. Huang, R. Samudrala, and J. Ponder. Distance geometry generates native-like folds for small helical proteins using the consensus distances of predicted protein structures. *Protein Sci.*, 7:1998–2003, 1998.
19. L.J. McGuffin, K. Bryson, and D.T. Jones. Psipred: a protein structure prediction server. <`http://globin.bio.warwick.ac.uk/psipred/`>, 1999.
20. A. Bairoch and R. Apweiler. The swiss-prot protein sequence data bank and its supplement trembl. *Nucleic Acids Res.*, 25:31–36, 1997.
21. J. Goldberg. Structural and functional analysis of the ARF1-ARFGAP complex reveals a role for coatomer in GTP hydrolysis. *Cell*, 1999:893–902, 1999.
22. S. Mosimann, R. Meleshko, and M.N.G. James. A critical assessment of comparative molecular modeling of tertiary structures in proteins. *Proteins: Struct., Funct., Genet.*, 23:301–317, 1995.
23. A.C.R. Martin, M.W. MacArthur, and J.M. Thornton. Assessment of comparative modelling in CASP2. *Proteins: Struct., Funct., Genet.*, To be published:0–0, 1997.
24. L. Wei, E.S. Huang, and R.B. Altman. Are predicted structures good enough to preserve functional sites? *Structure*, 7:643–650, 1999.