

# Functional Inferences from Blind *ab Initio* Protein Structure Predictions

Richard Bonneau, Jerry Tsai, Ingo Ruczinski, and David Baker<sup>1</sup>

Department of Biochemistry, University of Washington, Box 357350, Seattle, Washington 98195

Received December 26, 2000, and in revised form May 1, 2001; published online July 5, 2001

***Ab initio* protein structure prediction methods have improved dramatically in the past several years. Because these methods require only the sequence of the protein of interest, they are potentially applicable to the open reading frames in the many organisms whose sequences have been and will be determined. *Ab initio* methods cannot currently produce models of high enough resolution for use in rational drug design, but there is an exciting potential for using the methods for functional annotation of protein sequences on a genomic scale. Here we illustrate how functional insights can be obtained from low-resolution predicted structures using examples from blind *ab initio* structure predictions from the third and fourth critical assessment of structure prediction (CASP3, CASP4) experiments.** © 2001 Academic Press

**Key Words:** genomics; Rosetta; CASP3; CASP4; new folds.

## INTRODUCTION

The prediction of protein structure from amino acid sequence is one of the longest standing problems in molecular biology. Despite considerable effort, methods for predicting protein structure in the absence of a related sequence with a known structure have had relatively little success until very recently. As late as 1996 *ab initio* structure prediction was at best able to produce reasonable structures for only very small alpha-helical proteins. The failure of these methods on the vast majority of protein structures was highlighted by the first (1994) and second (1996) CASP (critical assessment of structure prediction) protein structure prediction experiments (Sippl *et al.*, 1999; Venclovas *et al.*, 1999). The results from these first two blind structure prediction experiments led Arthur Lesk, the

assessor of the *ab initio* structure predictions for CASP2, to conclude:

It is probably true, as many believe, that as the database grows to the point where the known sequences and structures saturate the living repertoire, the problem of *ab initio* structure prediction will disappear as methods based on homology modeling become much more generally applicable. If so, not only are we facing a very difficult problem but we have a limited time in which to solve it, if the solution is to make a general and practical impact. This is a shame. (Lesk, 1997)

By the time of CASP3 in 1998, however, *ab initio* structure prediction methods had improved considerably. The Rosetta method, developed in our group, produced reasonable low-resolution structures for fragments of eight structures (Orengo *et al.*, 1999a), five of which were over 65 residues in length (Simons *et al.*, 1999a). Other groups also made multiple correct predictions using a variety of methods (Ortiz *et al.*, 1999; Samudrala *et al.*, 1999).

CASP4 showed additional progress in the prediction of new folds and the prediction of folds for which fold recognition methods generally failed to recognize the correct template. By incorporating into Rosetta insights gained from experimental studies of folding, our group produced good blind predictions (fragments greater than 50 residues predicted to less than 6-Å Root Mean Square Deviation (RMSD)) for 16 of the 22 domains under 300 residues attempted. These predictions contained higher percentages of strand, were longer, and were generally of better quality than those seen at CASP3.

The sustained level of reasonable predictions of large fragments of relatively complex domains in the CASP4 experiment and the likelihood that prediction methods can be improved still further in the immediate future suggest that *ab initio* structure prediction may be able to make useful contributions to biological research. This is particularly timely given the large amount of genomic sequence information currently being generated. As *ab initio* structure prediction requires only the sequence of a pro-

<sup>1</sup> To whom correspondence should be addressed. Fax: (206) 685-1792. E-mail: [dabaker@u.washington.edu](mailto:dabaker@u.washington.edu).

tein to generate a three-dimensional model, it is well suited to help interpret the function of the significant fraction of genes in sequenced genomes that do not have detectable sequence similarity to proteins of known structure or function.

Many of the most reliable techniques for functional genome annotation rely on query sequences being homologous to other sequences of known or suspected function. However, these methods frequently fail to detect very distant structural and functional relationships (Ponting and Russell, 1995; Russell and Ponting, 1998) and 30–50% of open reading frames (ORFs) in new genomes have no homology to previously classified genes (Fetrow *et al.*, 1998; Mewes *et al.*, 2000; Rychlewski *et al.*, 1998; Sanchez and Sali, 1998). Fortunately, structural similarity is retained over larger evolutionary distances than amino acid sequence similarity (Brenner *et al.*, 1998), and structural similarity in some but not all cases can be indicative of functional similarity (Martin *et al.*, 1998). This greater retention of structural similarity is the basis of fold recognition/threading approaches to remote homologue detection.

*Ab initio* structure prediction methods can also potentially contribute to genome annotation. A significant fraction of proteins of unknown function are within the size range accessible to *ab initio* protein structure prediction (upper limit 100–150 residues) or can be parsed into domains in this size range using multiple sequence alignment-derived breakpoints. Once models are generated for unannotated ORFs, functional information may be obtained by global structural similarity searches (Simons, 2001) or by searching for conserved sequence/structure motifs characteristic of protein active sites or other functional regions (Fetrow *et al.*, 1999; Fetrow and Skolnick, 1998; Jonassen *et al.*, 2000; Moodie *et al.*, 1996; Wallace *et al.*, 1996).

In this report we illustrate how *ab initio* protein structure prediction can potentially contribute to genome annotation using as examples several of our blind protein structure predictions from CASP3 and CASP4. As many of the structures of the CASP4 prediction targets are not currently available publicly, we focus on the small number of proteins whose structures have already been published. A more complete description of the CASP4 *ab initio* structure predictions will be published in an upcoming supplemental issue of *Proteins: Structure Function and Genetics*.

#### BLIND STRUCTURE PREDICTIONS USING ROSETTA

The Rosetta method is based on a view of folding in which each short segment of the chain samples a

subset of the possible local conformations (dependent on its amino acid sequence), and folding to the native state occurs when the local segments simultaneously adopt conformations and relative orientations in which the hydrophobic residues are buried, the beta-strands are paired, and other nonlocal interactions are favorable (Bonneau *et al.*, 2000; Simons *et al.*, 1997, 1999b). The fundamental assumption underlying the method is that the distribution of conformations sampled by a particular sequence segment in isolation is reasonably well approximated by the distribution of conformations adopted by that sequence segment in known protein structures. Twenty-five to 200 fragments of known protein structures are selected based on sequence similarity for each 3- and 9-residue window of the query sequence. Tertiary structures are then generated using a Monte Carlo search of the possible combinations of these local structures, minimizing a scoring function that accounts for nonlocal interactions such as hydrophobic burial, compactness, strand pairing, and specific pair interactions. One thousand to 100 000 conformations are generated for each sequence, and a simple clustering procedure is used to identify the most frequently occurring families of structures; the centers of these clusters are the predicted models for the protein structure (Bonneau and Strauss, *in press*; Shortle *et al.*, 1998).

For the CASP3 and CASP4 structure prediction experiments, we generated five models for each of the sequences that lacked detectable sequence homology to proteins of known structure. The structure–structure comparison method Dali (Holm and Sander, 1995) was used to compare each of the models to the proteins in the PDB and to the true structure after it was released following submission of the predictions. The Dali *Z* score is a convenient measure of structure similarity; *Z* scores greater than 3–4 indicate significant structural similarity.

This procedure has some similarity to threading methods that use multiple targets. Rosetta builds up structures from large numbers of small fragments, rather than a small number of large fragments, through a large-scale search of conformational space. The differences between Rosetta and a multiple template threading approach are illustrated by the need for the Dali structure-based search of the protein structure database to determine what protein structure family, if any, a newly generated structure belongs to—in threading approaches, the protein structure family is identified before the model is generated, whereas in Rosetta, since the fragment libraries are derived from a wide range of completely unrelated proteins, a match to a known structure family is not evident until after the model is generated. Because of these differences, Rosetta,

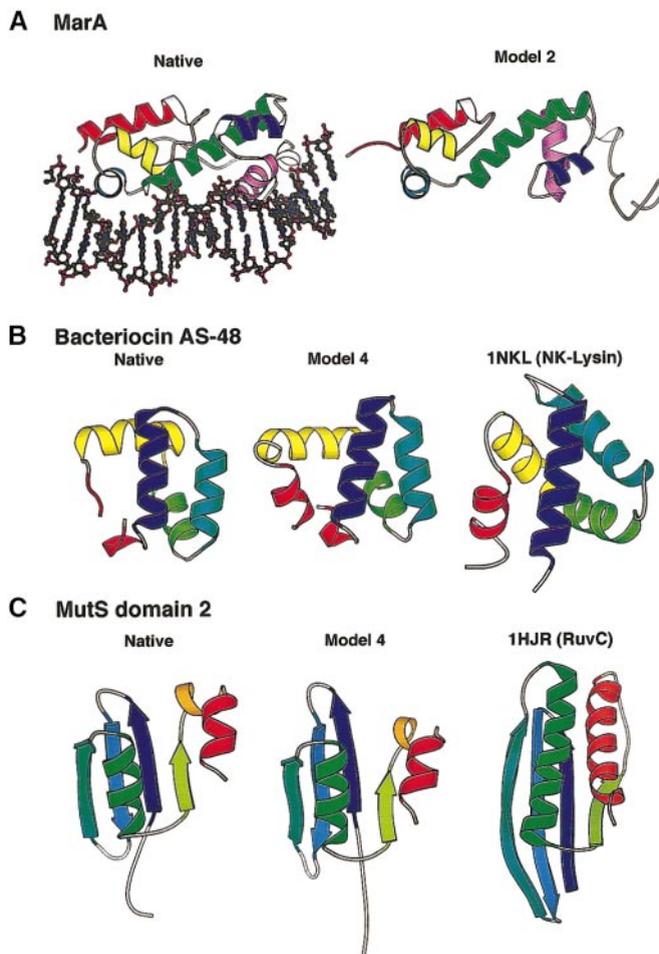
unlike traditional fold recognition methods, can generate structures for proteins with novel folds (Targets 54 in CASP3 and 91, 106, and 115 in CASP4, for example).

### MarA

One of the most interesting predictions by Rosetta in CASP3 was for MarA, a transcriptional activator responsible for multiple drug resistance in *Escherichia coli* and a member of the AraC family of transcriptional regulators (Rhee *et al.*, 1998). Our second model for this target had an RMSD of 6.4 Å over 100 residues and a Dali *Z* score of 6.0 to the native structure (Fig. 1A). The native structure and the model have two subdomains each, the overall structure being a dumbbell shape. The first subdomain of MarA is a helix-turn-helix DNA-binding motif. Our model has significant sequence independent structural matches to several proteins (1a04-A, 1qbj-A, 1bl0-A, and 1bia), all of which bind DNA with binding modes similar to that of MarA. These structure matches result in 30- to 45-residue stretches with sequence identities of 8–18% to the sequence of MarA; thus the local sequence and structure matches mutually reinforce the prediction that the first domain of MarA binds DNA. The second subdomain of our model also shows a structure match to a DNA-binding protein; thus the DNA-binding function of both subdomains could have been predicted based solely on our predicted models for MarA. The model produced by Rosetta is far more similar to the native structure than any other known protein structure is to the native structure and thus is considerably more accurate than any model that could be produced using a fold recognition method. Most interestingly, the relative orientation of the two subdomains in our model positions the DNA-binding helices in the two domains so as to fit well into the DNA major groove. MarA was known to be a transcription factor at the time of our prediction, and inspection of the model could well have suggested the mode of DNA binding, despite the errors in the model.

### Bacteriocin AS-48

One of the first CASP4 structures to be published was that of Bacteriocin AS-48 from *E. faecalis*, a cyclic bacterial lysin 70 residues in length (Gonzalez *et al.*, 2000). Our models 1 and 4 were quite good; model 4 had an RMSD of 3.5 Å over all 70 residues and a Dali *Z* score of 5.3 to the native structure. A search of the protein structure database with this model yielded 1nkl (NK-lysin) as the first structural match of comparable length. As is evident in Fig. 1B, the native structure of the bacteriocin is quite sim-



**FIG. 1.** Blind protein structure predictions from CASP3 and CASP4. (A) Left, crystal structure of MarA bound to double-stranded DNA (1bl0); right, our best submitted model. (B) Left, the crystal structure of Bacteriocin AS-48 (the peptide bond between the N- and the C-terminal residues is not shown); middle, our best submitted model; and right, the structural homologue (1nkl) identified using this model in a Dali search. (C) Left, crystal structure of the second domain of MutS (1ewq); middle, our best submitted model for this domain; and right, a structural homologue (RuvC) with a related function recognized using the model in a Dali search.

ilar to our model and to that of 1nkl, but the sequence identity in the structure-based alignment of the two proteins is only 4%. Importantly, despite the very low sequence identity, the two proteins have very similar functions (both are lysins). Thus Rosetta structure prediction, followed by a search of the structure database, identifies a protein of similar function with no detectable sequence similarity. Given the similar structure and function of the two proteins, it is likely that they have similar mechanisms of action and thus insights into the function of one are likely to hold for the other.

### MutS

The largest protein target in the CASP4 experiment was the 811-residue mismatch repair protein from *E. coli*, MutS (Obmolova *et al.*, 2000). Based on the multiple sequence alignment, we parsed the sequence into five domains, one of which had some similarity to a protein of known structure and was modeled using our comparative modeling methods, and four that had little detectable sequence similarity to proteins of known structure and hence were modeled using Rosetta. The recently published structure (1ewq) showed that our decisions on how to divide the protein into tractable domains were reasonably accurate; this is encouraging given the importance of such domain parsing in generating models for proteins of greater than 150 amino acids. One of our predictions for domain 2 was particularly good, having an RMSD of 2.5 Å over 70 residues and a Dali *Z* score of 6.0 to the corresponding region of the native structure (Fig. 1C). A search of the protein structure database with this model revealed strong structural similarities to proteins with Ribonuclease H-like folds (Lo Conte *et al.*, 2000; Murzin *et al.*, 1995) involved in large DNA-binding multidomain assemblies, including RuvC resolvase, a Holliday junction resolvase (Fig. 1C, right). Other matches to the Ribonuclease H-like fold class include the retroviral integrases (responsible for the integration of viral DNA into the host genome), polymerase domains, and Ribonuclease H (responsible for cleaving RNA duplexed with unwound double-stranded DNA). Thus our model, at the most basic level, would correctly suggest DNA-binding functionality involving three or more strands of DNA/RNA and on a more detailed level would identify a set of four possible functional families including one, RuvC, responsible for an analogous function (resolving improperly paired double-stranded DNA). The sequence divergence between domain 2 of MutS and the RNaseH-like domains is great enough that traditional sequence comparison and fold recognition methods produced models less accurate than our prediction in CASP4.

### CONCLUSION

The examples cited above illustrate a new approach for functional annotation of sequences that lack detectable sequence homology to proteins of known structure: generate three-dimensional models using structure prediction methods such as Rosetta and search the protein database with these models for proteins of similar structure. As in the case of MutS and Bacteriocin AS-48, the functions of the proteins whose structures are similar to that of the model may provide clues about the function of

the query sequence (in these cases functional information was already available, but this will not generally be true). Alternatively, in cases where the function of the protein is known, but the mechanism of action is not well understood, the structure may provide mechanistic insights, as the predicted structure of MarA could have provided insight into the mode of DNA binding. With the human genome sequence nearly complete, it will be exciting to see what *ab initio* structure prediction methods can contribute to the functional interpretation of the genome.

This strategy is complementary to traditional fold recognition methods that attempt to match a sequence with a previously determined fold. Such methods are likely to outperform *ab initio* structure prediction methods for larger and more complex proteins, which are beyond the range of current structure prediction methods. The *ab initio* structure prediction-based methods may be more powerful when the structures have diverged to the point that the query sequence no longer fits well onto the template structure, either because of changes in the lengths of the secondary structure elements, as in the MutS example in Fig. 1C, or because of significant changes in the solvent accessibility and residue-residue interaction patterns accompanying changes in the orientations of secondary structural elements. Consistent with this expectation, for many of the difficult fold recognition targets in CASP4, the best models were produced by *ab initio* methods rather than threading methods. There is, however, still considerable work to be done prior to applying Rosetta on a genome scale: in particular the method must be completely automated, the false-positive rate reduced, and a method for producing accurate confidence values developed.

### REFERENCES

- Bonneau, R., and Strauss, C. E. M. (2001) Improving the performance of ROSETTA using multiple sequence alignment information and global measures of hydrophobic core formation. *Proteins* **43**, 1–11.
- Brenner, S. E., Chothia, C., and Hubbard, T. J. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA* **95**(11), 6073–6078.
- Fetrow, J. S., Godzik, A., and Skolnick, J. (1998) Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: Identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* **282**, 703–711.
- Fetrow, J. S., Siew, N., and Skolnick, J. (1999) Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. *FASEB J.* **13**(13), 1866–1874.
- Fetrow, J. S., and Skolnick, J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure

- ture-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* **281**(5), 949–968.
- Gonzalez, C., Langdon, G. M., Bruix, M., Galvez, A., Valdivia, E., Maqueda, M., and Rico, M. (2000) Bacteriocin AS-48, a microbial cyclic polypeptide structurally and functionally related to mammalian NK-lysin. *Proc. Natl. Acad. Sci. USA* **97**(21), 11221–11226.
- Holm, L., and Sander, C. (1995) Dali: A network tool for protein structure comparison. *Trends Biochem. Sci.* **20**(11), 478–480.
- Jonassen, I., Eidhammer, I., Grindhaug, S. H., and Taylor, W. R. (2000) Searching the protein structure databank with weak sequence patterns and structural constraints. *J. Mol. Biol.* **304**, 599–619.
- Lesk, A. M. (1997) CASP2: Report on ab initio predictions. *Proteins Suppl.* **1**, 151–166.
- Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G., and Chothia, C. (2000) SCOP: A structural classification of protein database. *Nucleic Acids Res.* **28**(1), 257–259.
- Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B., Taroni, C., and Thornton, J. M. (1998) Protein folds and functions. *Structure* **6**(7), 875–884.
- Mewes, H. W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuller, C., Stocker, S., and Weil, B. (2000) MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.* **28**(1), 37–40.
- Moodie, S. L., Mitchell, J. B., and Thornton, J. M. (1996) Protein recognition of adenylate: An example of a fuzzy recognition template. *J. Mol. Biol.* **263**, 486–500.
- Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K., and Pedersen, J. T. (1997) Critical assessment of methods of protein structure prediction (CASP): Round II. *Proteins Suppl.* **1**, 2–6.
- Moult, J., Hubbard, T., Fidelis, K., and Pedersen, J. T. (1999) Critical assessment of methods of protein structure prediction (CASP): Round III. *Proteins Suppl.* **3**, 2–6.
- Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
- Obmolova, G., Ban, C., Hsieh, P., and Yang, W. (2000) Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. *Nature* **407**(6805), 703–710.
- Orengo, C. A., Bray, J. E., Hubbard, T., LoConte, L., and Sillitoe, I. (1999a) Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins Suppl.* **3**, 149–170.
- Orengo, C. A., Todd, A. E., and Thornton, J. M. (1999b) From protein structure to function. *Curr. Opin. Struct. Biol.* **9**(3), 374–382.
- Ortiz, A. R., Kolinski, A., Rotkiewicz, P., Ilkowski, B., and Skolnick, J. (1999) Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins Suppl.* **3**, 177–185.
- Ponting, C. P., and Russell, R. B. (1995) Swaposins: Circular permutations within genes encoding saposin homologues [Letter] [See comments]. *Trends Biochem. Sci.* **20**(5), 179–180.
- Rhee, S., Martin, R. G., Rosner, J. L., and Davies, D. R. (1998) A novel DNA-binding motif in MarA: The first structure for an AraC family transcriptional activator. *Proc. Natl. Acad. Sci. USA* **95**(18), 10413–10418.
- Russell, R. B., and Ponting, C. P. (1998) Protein fold irregularities that hinder sequence analysis. *Curr. Opin. Struct. Biol.* **8**(3), 364–371.
- Rychlewski, L., Zhang, B., and Godzik, A. (1998) Fold and function predictions for *Mycoplasma genitalium* proteins. *Fold. Des.* **3**(4), 229–238.
- Samudrala, R., Xia, Y., Levitt, M., and Huang, E. S. (1999) A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. *Pac. Symp. Biocomput.*, 505–516.
- Sanchez, R., and Sali, A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. USA* **95**(23), 13597–13602.
- Shortle, D., Simons, K. T., and Baker, D. (1998) Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl. Acad. Sci. USA* **95**(19), 11158–11162.
- Simons, K. T., Bonneau, R., Ruczinski, I., and Baker, D. (1999a) Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins Suppl.* **3**, 171–176.
- Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225.
- Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C., and Baker, D. (1999b) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* **34**(1), 82–95.
- Simons, Kim T., Strauss, Charlie E. M., Baker, David. (2001) Prospects for ab initio protein structural genomics. *J. Mol. Biol.* **306**, 1191–1199.
- Sippl, M. J., Lackner, P., Domingues, F. S., and Koppensteiner, W. A. (1999) An attempt to analyse progress in fold recognition from CASP1 to CASP3. *Proteins* **37**(S3), 226–230.
- Venclovas, C., Zemla, A., Fidelis, K., and Moult, J. (1999) Some measures of comparative performance in the three CASPs. *Proteins Suppl.* **3**, 231–237.
- Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1996) Derivation of 3D coordinate templates for searching structural databases: Application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* **5**(6), 1001–1013.