BIOINFORMATICS

Vol. 18 Suppl. 1 2002
Pages S54–S61

Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA

Christopher Bystroff* and Yu Shao

Department of Biology, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

Received on January 24, 2002; revised and accepted on March 27, 2002

ABSTRACT

Motivation: The Monte Carlo fragment insertion method for protein tertiary structure prediction (ROSETTA) of Baker and others, has been merged with the I-SITES library of sequence structure motifs and the HMMSTR model for local structure in proteins, to form a new public server for the *ab initio* prediction of protein structure. The server performs several tasks in addition to tertiary structure prediction, including a database search, amino acid profile generation, fragment structure prediction, and backbone angle and secondary structure prediction. Meeting reasonable service goals required improvements in the efficiency, in particular for the ROSETTA algorithm.

Results: The new server was used for blind predictions of 40 protein sequences as part of the CASP4 blind structure prediction experiment. The results for 31 of those predictions are presented here. 61% of the residues overall were found in topologically correct predictions, which are defined as fragments of 30 residues or more with a root-mean-square deviation in superimposed alpha carbons of less than 6Å. HMMSTR 3-state secondary structure predictions were 73% correct overall. Tertiary structure predictions did not improve the accuracy of secondary structure prediction.

Availability: The server is accessible through the web at http://isites.bio.rpi.edu/hmmstr/index.html. Programs are available upon requests for academics. Licensing agreements are available for commercial interests.

Supplementary information: http://isites.bio.rpi.edu, http://predictioncenter.llnl.gov/casp4/

Contacts: bystrc@rpi.edu; shaoy@rpi.edu

Keywords: CASP; CAFASP; protein folding; motifs; hidden Markov model; knowledge-based.

INTRODUCTION

Recent progress has been made toward the *ab initio* prediction of protein structure from primary sequence alone. The results from the last two CASP experiments (Bonneau *et al.*, 2001; Pillardy *et al.*, 2001; Simons *et al.*, 1999a, 1997) have singled out the power of protein folding simulations in reduced representation, such as the Monte Carlo fragment insertion (MCFI) search method implemented in the Rosetta algorithm (Simons *et al.*, 1997), which correctly predicted protein fragments of up to 107 residues in length with an accuracy of 5Å root-mean-square deviation in superimposed alpha-carbon coordinates (RMSD). The strength of the MCFI approach is its simultaneous but independent predictions of local structure and tertiary structure. As such, the algorithm is a model for protein folding that views the local structure propensities as independent of the tertiary structure to a large extent.

The throughput of this algorithm is limited due to the computational burden and the necessity for human input. Here we investigated the possibility of streamlining and automating the process, making it possible to offer the service publicly via the web. The goal was to return a reasonable tertiary structure prediction in minutes, rather than hours or days.

Because of the necessary improvements in efficiency, the automated I-sites/HMMSTR/Rosetta server differs from the approach of Baker in several ways:

1. The moveset of predicted fragment comes from I-sites (Bystroff and Baker, 1998) or HMMSTR (Bystroff *et al.*, 2000), instead of using a nearest neighbor approach.
2. Shorter conformational searches are done, and fewer repetitions.
3. Simulations are done on overlapping short segments of the chain instead of the whole chain or manually derived segments
4. Partial predictions are recombined using a simple genetic algorithm, to produce a set of final tertiary structure predictions.
5. The human post-processing and cluster analysis are omitted.

The current server addresses the *ab initio* folding problem only. Users of the server should first determine whether

*To whom correspondence should be addressed.

a homolog of known structure exists, using a database search (Altschul *et al.*, 1997). Failing that, a remote homolog of known structure may sometimes be identified with a high confidence using fold recognition approaches (Fischer and Eisenberg, 1996; Jones and Thornton, 1996; Murzin and Bateman, 1997). Failing that, *ab initio* structure prediction may provide useful information about supersecondary structure.

Here we assess the abilities of the streamlined I-sites/Rosetta server, the first fully automatic *ab initio* protein tertiary structure prediction server, using *bonafide* results from the CAFASP2 (Fischer *et al.*, 2001) experiment of 2000. In assessing the method, we focus on structural characteristics that are correctly or incorrectly predicted. The local and secondary structure predictions are compared and contrasted with the tertiary structure predictions.

SYSTEM AND METHODS

Hardware and interface

The server runs on a cluster of 24 Pentium 3 Linux machines, and typically returns the results (five sets of coordinates in PDB format, plus other results) in under 30 minutes. A queuing system controls the load, allowing at most two predictions to run concurrently. Over the last four months, the server has received an average of 300 jobs per month.

Figure 1 shows a flowchart describing the input and output of the programs included in the server. Most of the methods used by the server have been described elsewhere, including Psi-Blast (Altschul and Koonin, 1998), the I-sites Library (Bystroff and Baker, 1997, 1998), Rosetta (Bonneau *et al.*, 2001; Simons *et al.*, 1999a, 1997, 1999b), and HMMSTR (Bystroff *et al.*, 2000). ISLfrag, FragMaker and GArose are first described below. The server is written in *csh* script and *perl*, and runs in conjunction with a *csh* 'deamon' script, which handles job control. The results are returned as a web page and also via email.

The interface provides options for input formatting including single sequences, FASTA, ClustalW and SAF multiple sequence alignments and PDB files. Single sequences are optionally filtered for low-complexity using PSEG (Wootton and Federhen, 1996), and submitted to Psi-Blast, which returns a multiple sequence alignment. Earlier experiments show a greatly improved performance in local structure prediction when sequence profiles are used, as opposed to single sequences (Bystroff and Baker, 1997). No filtering was used for the CASP4 targets described here.

ALGORITHM AND IMPLEMENTATION

MakeVALL: calculates a sequence profile

The input multiple sequence alignment or the alignment produced by Psi-Blast is reformatted and converted to a sequence profile using sequence weights to correct for redundancy (Vingron and Argos, 1989).

FASTpred: predicts I-sites motifs given a sequence profile

The sequence profile is compared in a sliding-window fashion with each of the 261 I-sites Library scoring matrices (Bystroff *et al.*, 2000). This score is mapped to a confidence. The server returns a list of fragment predictions, expressed as backbone angles, sorted by confidence. The highest confidence fragments are referred to as 'I-sites predictions', the whole list as 'I-sites fragments'.

ISLfrag: generates a fragment moveset using I-sites predictions (Server 1)

For server 1, the I-sites fragment list was converted to a Rosetta move set. Rosetta uses fragment libraries of length 3 and 9 peptides as movesets for Monte Carlo fragment insertion. Each I-sites fragment with length $L \geq 9$ was divided into $L - 9 + 1$ subsegments of length 9 (*fragment moves*), each associated with a starting position in the target. Up to 25 of the highest confidence *fragment moves* are kept for each 9-residue window in the query. If not enough high confidence fragments were found, the list was augmented by extending 7 and 8 residue I-sites fragments. A similar procedure was done for the moveset of length 3.

HMMSTR: predicts secondary and local structure (Server 2)

For server 2, the profile was submitted to each of the three HMMSTR models, one for prediction of backbone angles (HMMSTR-r), one for the prediction of secondary structure (HMMSTR-d), and the third for the prediction of supersecondary structure 3D context (HMMSTR-c). In each of these models, the Markov states contain probability distributions over sets of symbols: 3-state secondary structure, backbone angle regions, or context symbols, for models d, r, and c, respectively.

Using the query sequence profile, an *a posteriori* probability may be calculated for each Markov state and each position, using the forward-backward algorithm (Rabiner, 1989), resulting in a matrix of conditional state probabilities, γ. The confidence for a predicted symbol, such as secondary structure symbol (ss) is the γ-weighted sum over all states, q, as follows:

$$P(ss \mid i) = \sum_{q} \gamma(q \mid i) P(ss \mid q). \qquad (1)$$

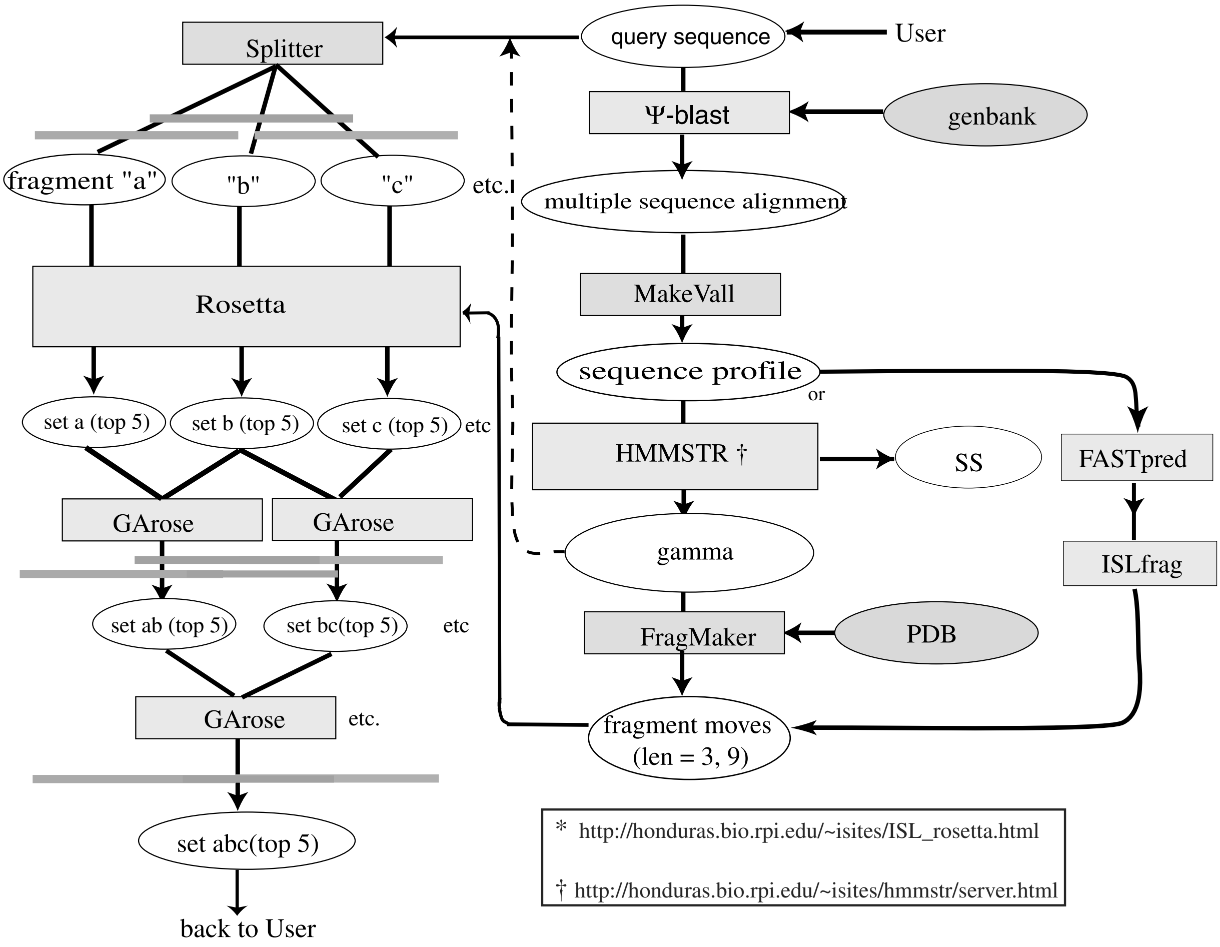

Fig. 1. See text explanations.

Secondary structure is reported as a string of ss symbols and a reliability index, which is $P(ss \mid i)$ multiplied by 10 and truncated to an integer.

FragMaker: generates a fragment moveset using *HMMSTR (Server 2)*

γ-matrices were calculated and stored for each of the 794 proteins in the PDBselect database (Hobohm and Sander, 1994) using dual input data: sequence profiles and backbone angle symbols. Dual input data assured that the Markov states at each position are consistent with both the sequence and structure. The 794 γ-matrices were compared using a vector product to each 3 and 9-residue window of the target. The top-scoring 25 PDBselect segments were saved for each target position. Fragment moves are expressed as backbone angles.

Rosetta: a conformational search algorithm for proteins

Rosetta searches protein conformational space using fragment insertion moves and a Monte Carlo acceptance critereon. An insertion point in the target is selected at random, then a fragment (either length 3 or 9) is selected at random from the move set. The backbone angles are changed to those of the fragment, new coordinates are computed from the backbone angles, and the move is accepted or rejected, using Monte Carlo. The energy function (Simons *et al.*, 1999b) is composed of structure-based Bayesian conditional probability expressions, drawn from the same PDBselect database.

Simulated annealing. The acceptance critereon for Monte Carlo fragment insertion (MCFI) depends on the energy and the ‘temperature’ setting (T). T is set initially set to a high value so that most physically-possible moves are accepted, then decreased linearly over 12 000 moves (simulated annealing). The optimal temperature schedule for simulated annealing depends on the length of the chain being simulated, or more specifically, the number of degrees of freedom. For practical reasons, a fixed temperature schedule is used by the server, and the length of the input sequence in restricted to a narrow range.

Those practical reasons included the anticipated availability of hardware, and the perceived need for speed. Our server cluster, having 24 nodes, can support about ten simultaneous Rosetta jobs per machine before memory is exhausted. If the jobs are very long or there is a surge of submissions to the server, it would crash one or more of the processors, creating a nuisance for the cluster administrators. A queuing system now prevents this from happening. However, users are likely not to want to wait more than a few hours for results. Therefore, the run length was fixed at 12 000 MCFI cycles, long sequences were split into manageable pieces, and the number of replicates was fixed at 15. An appropriate temperature schedule was chosen by trial-and-error. These settings were tested on a small sample of known structures.

Restraining selected backbone angles. High confidence I-sites predictions were restrained to their predicted backbone angles to increase efficiency. Fragment insertion was allowed in the restrained regions, but moves were rejected if any angle deviated by more then 60Å from the I-sites prediction. A maximum of one-third of the residues may be restrained, because over-restraining was found to impair the search efficiency in test cases.

Splitter: Dividing long sequences into overlapping segments. If the target sequence had more than 36 un-restrained residues, then it was divided into overlapping segments having 36 un-restrained residues each. Adjacent segments overlap by at least 18 un-restrained residues, plus any number of restrained residues.

GArose: a simple algorithm for assembling segment predictions A total of 15 fragment predictions were produced by Rosetta for each segment. The 5 best predictions for adjacent segments were re-combined by exhaustive splicing. Starting with two sets of overlapping segment predictions, all possible crossover hybrid models were made and the five with the lowest energy were saved for the next round, or for final output to the user (See Figure 1).

RESULTS AND DISCUSSION

Two servers were used in the CASP4 experiment, differing only in the algorithm for generating the move sets. Server 1, (prediction group ISITES, group 216 in the CASP4 records) used ISLfrag movesets. Server 2 (prediction group BYSTROFF, group 55) used a moveset generated by FragMaker. Without further elaboration, it was found that the two movesets gave essentially equivalent results, and that the moveset, if it contained a reasonably accurate set of fragment predictions, was not the limiting reagent for success.

Tertiary structure predictions by the ISITES group, and secondary structure predictions by the BYSTROFF group were sent to CAFASP2 (Fischer *et al.*, 2001). Thus, only these are truly, *bonafide,* fully automated. The statistics and conclusions presented here refer to Server 1 and the best (model 1) of its five submissions for each target. Fragments from the other four models are shown for illustrative purposes. A full and detailed analysis of the predictions made by the I-sites server and others can be obtained from the web site for the CASP4 experiment (http://predictioncenter.llnl.gov).

Of the 40 target proteins predicted by the server, solved structures were reported for 32 at the time of the CASP4 meeting (December, 2000). One of these (T0092) contained only alpha-carbons, and therefore the backbone angles could not be evaluated. The remaining 31 structures are discussed here, with an emphasis on overall statistical results that might be helpful in understanding the strengths and shortcomings of the prediction protocol.

OVERALL RESULTS

Over the 31 target proteins, 61% of the residues were found in ‘topologically correct’ large fragments, defined as fragments of 30 residues or more with RMSD <6Å. The locations of these fragments and longer fragments with the same RMSD cutoff are shown in the lower band of Figure 2, shaded by the length range of the fragment. At 6A RMSD, the correct overall chain trace has been reproduced, but not the finer details of structure. Occasionally beta strand may be out of order in a sheet, and strands may be substituted for helices.

A smaller percentage of large fragments, 44%, were predicted with a 5Å accuracy. At 5Å RMSD, secondary structure is occasionally mispredicted, loop structures may be wrong in detail, and axial rotations of secondary structure units are possible. However, much or most of the non-local packing interactions are faithfully though roughly reproduced at this level of accuracy, and strand mispairing is not observed.

In practice, the details of the local structure are often correctly predicted when a fragment was globally correct, but the RMSD measure is insensitive to this. Therefore, another measure is used to evaluate the local accuracy of the predictions . The maximum deviation in backbone angles (*mda*) over a window of 8 residues is usually ~180Å or small, and serves as a strictly local measure of correctness. 8-residue peptides that have *mda* <90Å and obey all of the stereochemical constraints of a polypeptide, have an RMSD of 1.4Å at most (Bystroff and Baker, 1998). The top band of Figure 2 shows the locations of fragments with *mda* <90Å. It is immediately obvious that the good local structure predictions do not always superimpose on the good, large fragment predictions.

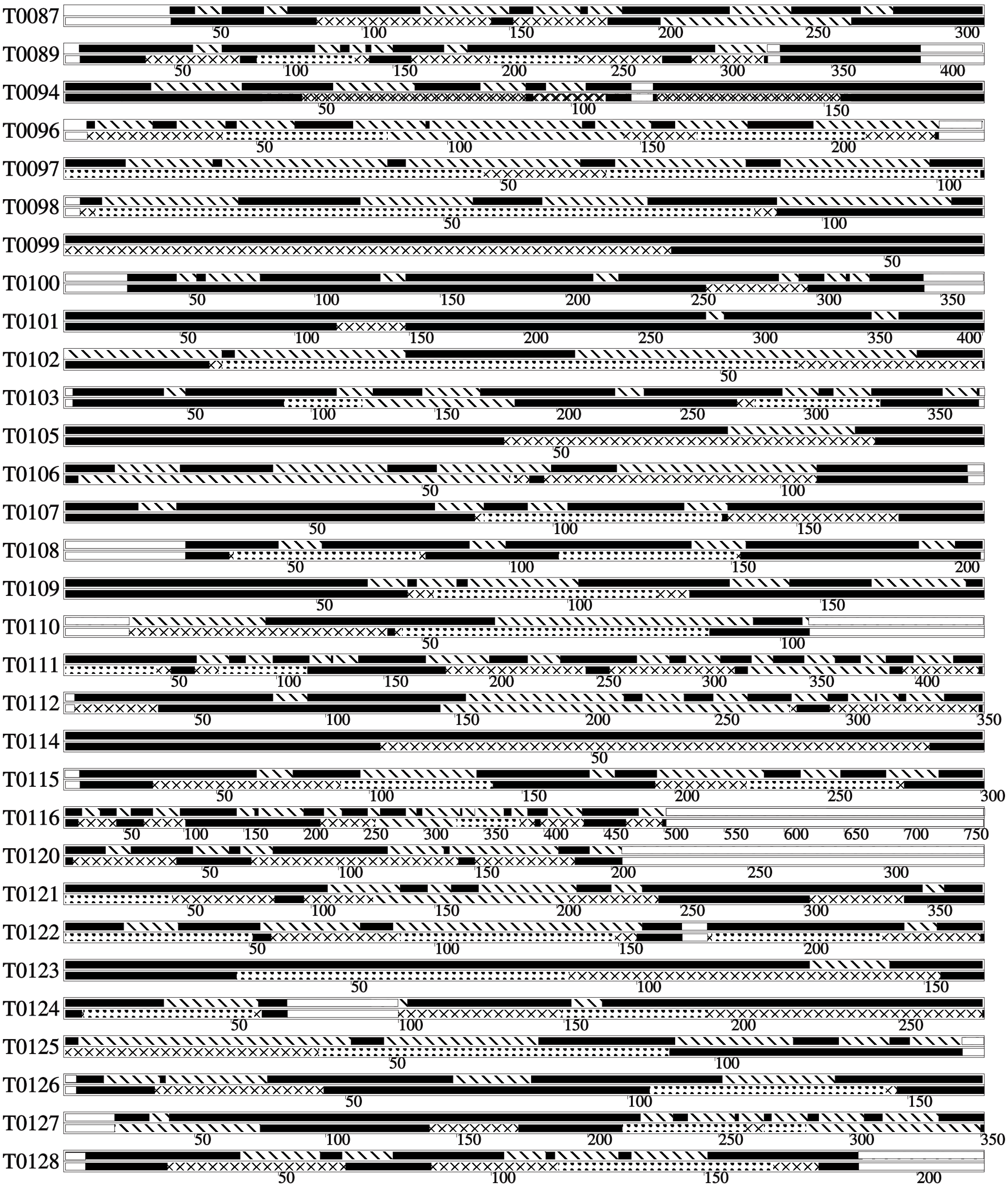

Fig. 2. Locations of correctly predicted local structure (upper band) marked in slashed lines, and topologically correct fragment predictions (lower band) of length 30–39 (crossed lines), 40–49 (dots), and 50 or greater (slashed lines), for each target (left margin). White bands are regions that were either not predicted or missing in the target coordinates.

Topologically correct large fragment predictions: potential improvements

Figures 3–5 show examples of fragments longer than 30 residues with 6Å RMSD. In general, the residues found in the core were correct, and their 3D arrangement was roughly correct. In fragments that contained helices, the N and C capping residues were usually but not always correctly located, and the direction of the chain coming off of the helix was generally correct. The orientation of parallel sheets to helices was reproduced to within about 60Å, and the axial orientation of the helices with respect to strands was almost always correct, even though rolling the helix would not greatly effect the RMSD value.

Some characteristics of even the 'correct' fragment predictions suggested ways in which the algorithm could be improved. The most obvious of these is the distortion of alpha helices. True native helices retain very straight

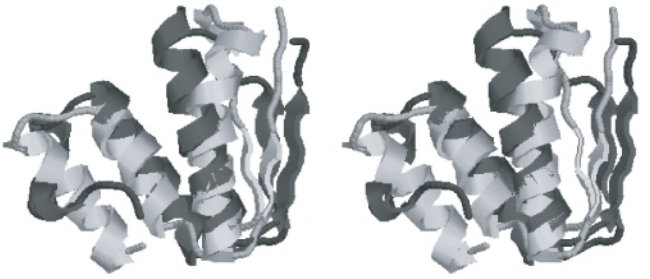

Fig. 3. T0121 residues 126–199, RMSD = 5.9Å. Both the overall structural orientation and the secondary structures are mostly correct. Light: true structure, dark: predicted. Here we see a bent helix in the foreground.

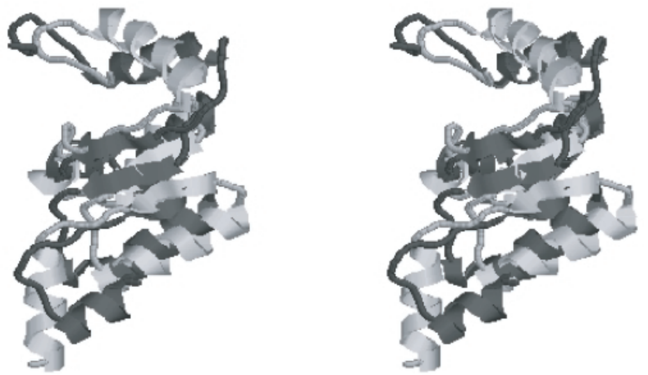

Fig. 4. T0122, residue 57–153, RMSD = 5.9Å. This is a successful prediction of a repeating beta-alpha-beta motif. All the helices and strands are correctly predicted. Light: true structure, dark: predicted.

helix axes despite variability in the backbone angles. Helices in the predictions, however, were often distorted (see Figure 3), sometimes bending the axis by 90Å over its length. A combination of factors produce these errors. Rosetta has no energy penalty for helix distortion, while it gives a large energetic bonus for packing hydrophobic residues in the core and for maintaining a low radius of gyration. Bent helices are found to replace helix kinks and alpha-alpha corners (Efimov, 1996). Adding a penalty for helix distortion might fix the problem.

Good local structure prediction correlates weakly with good tertiary structure prediction

A popular view of protein folding pathways could be described roughly as ‘local structure first’ followed by tertiary structure (Nolting and Andert, 2000). If the Rosetta simulations are following such a pathway, then we would expect to see good supersecondary structure predictions coinciding with good local structure predictions. However, this is not always the case. Figure 1 shows (in top strip,

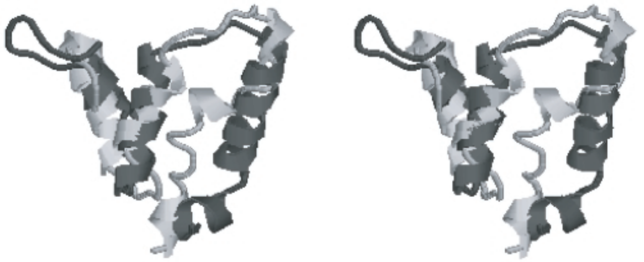

Fig. 5. T0116, residues 262–322, RMS = 5.9Å. The prediction of this 3-helix bundle is topologically correct but one of the helices is mostly unfolded. The core residues are correctly predicted. Light: true structure, dark: predicted.

yellow shading) the regions of the target where the local structure predictions were good. Specifically, they show 8-residue or longer segments with no backbone angle deviations greater than 120Å. Frequently, the topologically correct large fragments have the wrong local structure, even though at least one fragment of correct local structure existed in the moveset for approximately 90% of the sequences.

Tertiary structure prediction does not improve secondary structure prediction

3-state secondary structure (SS) predictions were made using a version of HMMSTR that was trained on a large dataset of proteins of known structure with SS states assigned using DSSP (Kabsch and Sander, 1983). The accuracy of these predictions over the 31 targets was 73.3%. This is only slightly lower than the state of the art in SS prediction (Jones, 1998). SS predictions based on tertiary structure (TS) predictions from Rosetta had the potential of benefiting from the added TS information, however this proved not to be the case.

SS assignments for the TS predictions made using DSSP or STRIDE (Frishman and Argos, 1995), performed poorly (50–60% Q3), because these programs depend on precise positioning of the hydrogen-bonding residues in assigning the strand state (E). Instead, we derived the SS predictions from the fragments used to assemble the TS predictions. The fragment SS assignments are derived from their native proteins. Using this method, the overall Q3 score improved to 72.4%, but this is no better than the SS predictions that use sequence alone.

If the simulation were reproducing the folding process, one might expect that the correctly-predicted tertiary interactions would add information to the secondary structure prediction. One explanation for the lack of improvement in secondary structure, despite some success in tertiary packing, is that topologically correct tertiary structures are possible even when the wrong local structure is used to build it.

Predictions have lower average contact order than true structures.

Relative contact order (Plaxco *et al.*, 1998) is calculated from the coordinates as follows:

$$CO = \frac{1}{L \bullet N} \sum^{N} \Delta S_{ij} \qquad (2)$$

where ΔS_{ij} is the sequence separation $|i - j| \geqslant 5$, for residues, ij, that are in contact ($C\alpha - C\alpha$ distance <8Å). N is the number of contacts, and L is the length of the sequence. The overall average CO in the targets was 0.252, while the CO for the 32 predictions was 0.119. The lower CO is mostly the result of an increased number of beta hairpins. Contacts that are local, such as those in beta hairpins, are easier to find in a search, and thus may represent kinetic intermediates, trapped at the end of the simulation. Kinetic trapping may be exacerbated by the more computationally efficient server protocol. A possible solution is to do more replicates and rely on cluster analysis to identify the global energy minimum. Practical limitations currently stand in the way of implementing this.

Alternatively, the predominance of beta hairpins may reflect an error in the energy function with regard to the backbone angles. Positive ϕ angles, favored only in glycine residues and usually required for turns, are found in the same proportion in the targets (8%) and in the predictions (7%), but in the targets, 44% of these turn residues are glycines, while in the prediction only 16% are glycines. This suggests that a larger energetic penalty for positive ϕ angles in non-glycine residues, might correct the overabundance of hairpin turns.

CONCLUSIONS

Our results suggest that a combination of improvements in efficiency may increase the potential of the Rosetta algorithm as a high-throughput engine for tertiary structure prediction at the 30–100 residues length scale. We suggest that a combination of structure comparison metrics be used for the evaluation of correctness; a low RMSD in the context of low backbone angle deviations is shown to identify predictions that were ‘correct for the right reasons’.

Secondary structure assignments were not improved by the use of tertiary structure predictions, partly because it was possible to obtain a globally correct tertiary structure prediction by inserting fragments of the wrong local structure.

An overall low contact order was observed in the predictions relative to the true structures. This is at least partly due to the absence of an energetic penalty for unfavorable backbone torsion angles. These may also represent kinetically-trapped intermediate structures from a simulation that was too short.

REFERENCES

Altschul,S.F. and Koonin,E.V. (1998) Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci.*, **23**, 444–447.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bonneau,R., Strauss,C.E. and Baker,D. (2001) Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. *Proteins*, **43**, 1–11.

Bystroff,C. and Baker,D. (1997) Blind predictions of local protein structure in CASP2 targets using the I-sites library. *Proteins*, **Suppl 1**, 167–171.

Bystroff,C. and Baker,D. (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.*, **281**, 565–577.

Bystroff,C., Thorsson,V. and Baker,D. (2000) HMMSTR: A hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.*, **301**, 173–190.

Efimov,A.V. (1996) A structural tree for alpha-helical proteins containing alpha-alpha-corners and its application to protein classification. *FEBS Lett.*, **391**, 167–170.

Fischer,D., Elofsson,A., Rychlewski,L., Pazos,F., Valencia,A., Rost,B., Ortiz,A.R. and Dunbrack,Jr,R.L. (2001) CAFASP2: The second critical assessment of fully automated structure prediction methods. *Proteins*, **45 Suppl 5**, 171–183.

Fischer,D. and Eisenberg,D. (1996) Protein fold recognition using sequence-derived predictions. *Protein Sci.*, **5**, 947–955.

Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.

Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.

Jones,D.T. (1998) *Critical Assessment of Protein Structure Prediction 3*. Asilomar, CA.

Jones,D.T. and Thornton,J.M. (1996) Potential energy functions for threading. *Curr. Opin. Struct. Biol.*, **6**, 210–216.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Murzin,A.G. and Bateman,A. (1997) Distant homology recognition using structural classification of proteins. *Proteins*, **Suppl 1**, 105–112.

Nolting,B. and Andert,K. (2000) Mechanism of protein folding. *Proteins*, **41**, 288–298.

Pillardy,J., Czaplewski,C., Liwo,A., Lee,J., Ripoll,D.R., Kazmierkiewicz,R., Oldziej,S., Wedemeyer,W.J., Gibson,K.D., Arnautova,Y.A., Saunders,J., Ye,Y.J. and Scheraga,H.A. (2001) Recent improvements in prediction of protein structure by global optimization of a potential energy function. *Proc. Natl Acad. Sci. USA*, **98**, 2329–2333.

Plaxco,K.W., Simons,K.T. and Baker,D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, **277**, 985–994.

Rabiner,L.R. (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

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**, 82–95.

Vingron,M. and Argos,P. (1989) A fast and sensitive multiple sequence alignment algorithm. *Comput. Appl. Biosci.*, **5**, 115–121.

Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, **266**, 554–571.