Prediction of novel and analogous folds using fragment assembly and fold recognition

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

Impact Factor: 2.63 · DOI: 10.1002/prot.20731 · Source: PubMed

7 AUTHORS, INCLUDING:

Kevin Bryson
University College London
39 PUBLICATIONS **5,666** CITATIONS
SEE PROFILE

Liam J McGuffin
University of Reading
58 PUBLICATIONS **6,662** CITATIONS
SEE PROFILE

Michael I Sadowski
MRC National Institute for Medical Research
34 PUBLICATIONS **963** CITATIONS
SEE PROFILE

Jonathan J Ward
European Molecular Biology Laboratory
18 PUBLICATIONS **2,599** CITATIONS
SEE PROFILE

Prediction of Novel and Analogous Folds Using Fragment Assembly and Fold Recognition

D. T. Jones,[1,2*] K. Bryson,[1] A. Coleman,[1] L. J. McGuffin,[1] M. I. Sadowski,[1] J. S. Sodhi,[1] and J. J. Ward[1]
[1]*Bioinformatics Unit, Department of Computer Science, University College London, London, United Kingdom*
[2]*Department of Biochemistry and Molecular Biology, University College London, London, United Kingdom*

ABSTRACT **A number of new and newly improved methods for predicting protein structure developed by the Jones–University College London group were used to make predictions for the CASP6 experiment. Structures were predicted with a combination of fold recognition methods (mGenTHREADER, nFOLD, and THREADER) and a substantially enhanced version of FRAGFOLD, our fragment assembly method. Attempts at automatic domain parsing were made using DomPred and DomSSEA, which are based on a secondary structure parsing algorithm and additionally for DomPred, a simple local sequence alignment scoring function. Disorder prediction was carried out using a new SVM-based version of DISOPRED. Attempts were also made at domain docking and "microdomain" folding in order to build complete chain models for some targets. Proteins 2005;Suppl 7:143–151.** © **2005 Wiley-Liss, Inc.**

Key words: protein structure prediction; CASP; protein sequence analysis; bioinformatics

INTRODUCTION

By looking at the results from previous CASP experiments, it is quite clear that methods for automatically inferring three-dimensional (3D) models for newly characterized proteins based on distant sequence similarity to known structures are reaching a high level of success. However, progress in modeling structures with no detectable evolutionary relationship to known structures has been somewhat slower. These problems can be divided into two categories. First, there are cases where the target protein does indeed share a great deal of similarity with a previously determined structure, but where there is no evidence for an evolutionary relationship. Second, there are cases where the target protein turns out to have a novel fold, in which case there are no useful complete templates in the structure databases. Oddly enough, there has perhaps been the most recent progress in the second category (i.e., in the prediction of new folds). Most of the recent successful methods for predicting novel folds have exploited the fact that when a new fold is discovered, it is generally found that the fold is still composed of common structural motifs at the supersecondary structural level. We originally exploited this observation in a method called FRAGFOLD,[1] the latest version of which is briefly described in this article. Methods such as FRAGFOLD attempt to greatly narrow the search of conformational space by preselecting structural fragments from a library of known protein structures. The original FRAGFOLD was used with some success in the CASP2 experiment in 1996, and was the first demonstration of the power of fragment assembly in building novel folds in a blind prediction study. Of course, there now exists a number of fragment assembly methods such as the highly successful Rosetta method of David Baker and colleagues.[2]

The most obvious approach to predicting the structures of the second category of target (where there exists a structural analog of the target protein) is to evaluate the fit of the target sequence on likely template folds, and to select the model that achieves the highest sequence–structure compatibility. This is the basis of the threading approach to protein structure prediction.[3] Since the initial excitement in the field surrounding threading methods (which were the most successful methods in the first two CASP experiments), progress has been relatively slow, and these methods are now being squeezed from both the distant homology methods on the one hand, and the fragment assembly methods on the other. Despite this, the idea of finding a useful template for a target protein from a limited set of alternatives remains a useful one, even though the purists in the threading field have long since become pragmatists, who have now adapted and extended the basic threading concept to suit a wide range of different prediction problems.

METHODS

An outline of the basic prediction strategy we used is shown in Figure 1. The first stage is what we refer to as "preprocessing." This involves the basic steps of predicting secondary structure, domain parsing, and identifying any

Grant sponsor: Joint Research Councils (D.T. Jones). Grant sponsor: Biotechnology and Biological Sciences Research Council (L. J. McGuffin, A. Coleman). Grant sponsor: Department of Trade and Industry (L. J. McGuffin). Grant sponsor: Wellcome Trust (K. Bryson). Grant sponsor: European Union Framework 6 BioSapiens Network of Excellence (M. I. Sadowski). Grant sponsor: Medical Research Council (Studentships to J. J. Ward and J. S. Sodhi).

*Correspondence to: D. T. Jones, University College London, Department of Computer Science, Bioinformatics Unit. Gower Street, London WC1E 6BT, UK. E-mail: dtj@cs.ucl.ac.uk

Received 15 April 2005; Accepted 1 June 2005

Published online 26 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20731

This article was originally published online as an accepted preprint. The "Published Online" date corresponds to the preprint version.

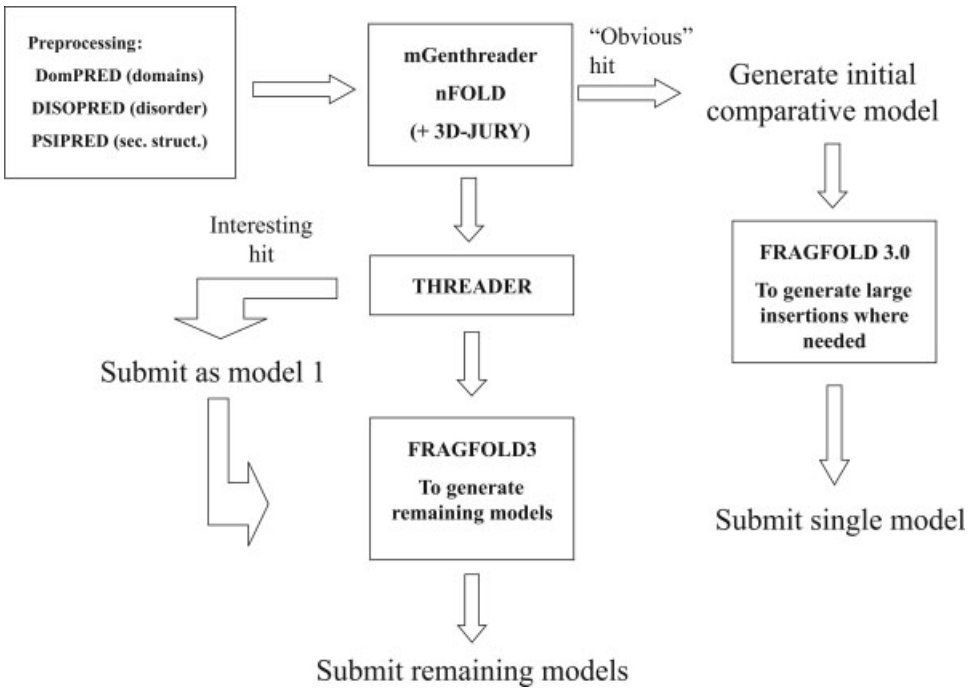

Fig. 1. Diagram showing the outline steps employed in predicting the complete set of CASP.

useful sequence features (including predicted regions of disorder). The next stage involves the application of automated fold recognition methods to try to determine the category of the target (or target domain), that is, whether there is an obvious template structure on which to base a model. We describe below a few of the newer programs referenced in Figure 1, but the main focus of this section is on the new developments incorporated into FRAGFOLD3.

Domain Prediction (DomPred and DomSSEA)

The DomPred Server[4] implements our previously published method for domain prediction, DomSSEA,[5] combined with a newly developed method called Domains Predicted from Sequence (DPS). DomSSEA uses a fold recognition approach, based on aligning the PSIPRED[6] predicted secondary structure for the query sequence against the DSSP[7] assigned secondary structures of a fold library. It then transfers the SCOP[8] assigned domain structure from the best fold match to the query sequence. DPS, however, starts out with a PSI-BLAST[9] search of the query sequence against a database consisting of NRDB90[10] augmented with sequences from Pfam-A.[11] Significant local alignment fragments are examined, and the total numbers of C- and N-terminals for the fragments are recorded for each residue position in the query sequence. These distributions are smoothed and are then combined giving additional weight to positions that have high values for both the C- and N-termini, since this provides more evidence for a domain boundary in which one conserved sequence region ends and another starts. The combined values are then turned into *Z*-scores by dividing throughout by the standard deviation over the entire query protein. A threshold is then applied to these *Z*-score values in order to predict domain boundaries.

The DomSSEA method is most effective when the fold library contains a complete structural match to the query. Hence, this approach bears some resemblance to remote homology detection or fold recognition. The DPS method makes no such use of complete structural matches, since, in this case, the alignment would just have its N- and C-terminal positions lying close to the N- and C-termini of the query sequence. Large scores close to the N- and C-termini of the query sequence are simply excluded as end-effects when predicting domain boundaries, for obvious reasons. DPS relies more on the database containing sequences having fragments that in combination reveal the domain structure of the protein, rather than a complete sequence with a structural match. Thus, it can be seen that these two methods are based on largely orthogonal information; hence, they are particularly effective in combination.

Currently the results from the two methods need to be combined by the user, but there are plans to form a consensus method, combining both of these approaches. Also, we wish to have the server carry out an initial screening stage in which it detects obvious homologs to Protein Data Bank (PDB) structures and just reports back their inferred domain structure. This will lead to a robust server that can deal with both easy and difficult cases.

Fold Recognition (THREADER, mGenTHREADER, and nFOLD)

THREADER 3.5 is the latest incarnation of our original program to implement threading,[3] and although it now incorporates a number of new features (in particular the use of sequence profiles) and a set of alignment parameters optimized with a genetic algorithm, the overall components of the current implementation remain more or less

unchanged since CASP2. THREADER 3.5 was used to predict targets that were not predicted with high confidence by our server-based prediction methods (which were submitted separately to the automatic server sections of CASP).

There have been a number of improvements in our fully automated fold recognition methods since CASP5. The widely used (m)GenTHREADER[12,13] method has been improved through the inclusion of profile–profile alignments. We have also developed a new method called nFOLD (manuscript in preparation) that is based on the new mGenTHREADER protocol, but which also incorporates a number of extra inputs into the underlying neural network.

The major change to the original mGenTHREADER algorithm is the implementation of a simple profile–profile alignment algorithm. The comparison method used was designed to directly compare PSI-BLAST position-specific scoring matrix (PSSM) scores, and makes use of an optimized heuristic comparison metric. The heuristic we use is essentially based on the dot product of the two PSSM vectors X (from the target) and Y (from the template), though only considering positive values in the target PSSM:

$$\sum_{i=1}^{20} \max \left\{ \begin{array}{c} 0 \\ X_i \end{array} \right\} Y_i$$

To find the correct alignment parameters (e.g., gap penalties) for this scoring function, the parameters were optimized using a genetic algorithm to maximize a sum of model quality over a benchmark set of 50 difficult fold recognition targets.

The nFOLD method is an extension of the new mGenTHREADER protocol. Three additional inputs are fed into the neural network, which include the secondary structure element alignment (SSEA) score,[13] a new functional site detection score (MetSite),[14] and a simple model quality checking algorithm, MODCHECK.[15] The nFOLD neural network is also trained directly on a model quality score that allows for a greater correlation between neural network output and model quality.

The functional site predictions were calculated using a set of classifiers based on the MetSite method,[14] which was initially developed in order to predict the location of residues forming commonly occurring metal binding sites in low-resolution structural models. The top-ranking MetSite predictions were extracted for the top models generated from the mGenTHREADER profile–profile alignments. Analysis of the MetSite scores showed a significant improvement in distinguishing native and near-nativelike models from decoy hits and was therefore implemented as an extra input in the nFOLD method.

The MODCHECK score was also used to directly assess the quality of the models from the profile–profile alignments. An early version of the MODCHECK program was used previously for our CASP5 predictions[15]; however, this is the first time it has been implemented in a fully automated method. A full description and evaluation of MODCHECK has also been carried out.[16]

Last, a relatively minor but important improvement to the fold recognition servers since CASP5 has been the implementation of fully automated weekly updates of both the fold recognition library and sequence databases, which reduces the chance that obvious homologs or fold templates will be missed.

In order to facilitate later function analysis (not discussed in this article), all-atom models were generated for as many models as possible. For obvious homologs, MODELLER[17] was used to generate an all-atom model. For the more difficult fold recognition targets, no attempt was made to model loops, and the new side-chain generation module in FRAGFOLD3 was used to generate side-chains for the incomplete backbone structures. For new fold predictions, side-chains were already added by FRAGFOLD3 (see below).

New Fold Prediction (FRAGFOLD3)

For CASP6 targets that we believed could not be reliably predicted using fold recognition methods, FRAGFOLD3[1,15] was used to generate up to five structures. This approach to protein tertiary structure prediction is based on the assembly of recognized supersecondary structural fragments taken from highly resolved protein structures using a simulated annealing algorithm. FRAGFOLD3 differs from previous versions by making use of both fixed-length and supersecondary structural fragments, explicitly modeling side-chains using a fast rotamer generation method, and an improved treatment of main-chain hydrogen bonding using a simple Morse potential. Up to 1000 structures were generated for each target domain using in-house Grid software running on a 100 CPU Linux cluster, and a simple rigid-body structural clustering algorithm was used to select the models representing the largest clusters of conformations. Submitted predictions were made using little or no human intervention apart from initial domain assignment and preparation of input secondary structure and sequence alignment files.

The method used in CASP6 is based on an earlier method[1] that was first used in the CASP2 experiment (and improved versions in CASP4 and CASP5), but here we will describe only the most recent changes to the method.

As with the original method, at the heart of the objective function is a set of pairwise potentials of mean force, determined by a statistical analysis of highly resolved protein X-ray crystal structures, and the application of the inverse Boltzmann equation to convert observed frequencies of residue pair interactions to free energy changes. In addition to the pair potentials, a solvation potential is also employed. These potentials are identical to those currently in use by the latest versions of our threading programs: THREADER[3] and GenTHREADER.[12]

For threading applications, pairwise and solvation potentials are often sufficient on their own to discriminate correct from incorrect protein folds. However, for ab initio prediction, it is also necessary to include extra terms to ensure that low energy folds are compact, have optimal

hydrogen bond networks, and have no steric clashes. In threading, these additional terms are unnecessary, because real protein folds are almost always compact, have no steric clashes, and have well-defined hydrogen bonding networks.

A major difference between FRAGFOLD3 and FRAGFOLD2 (used in CASP5) is that now an all-atom representation is used, rather than main-chain plus Cβ. A very simple rotamer search method is employed, with a maximum of just two rotamers considered per side-chain. Although on the face of it, this limits the possible accuracy of side-chain orientation, in testing on backbone structures taken from crystal structures we find that very good results are obtained, which are remarkably comparable to the results obtained from more rigorous side-chain placement methods.

As we now work with a full atom model, a simplified steric potential is no longer needed and steric clashes are now penalized using a soft repulsive term of the form:

$$E_{clash} = (R_a + R_b)^2 - (d_{ab})^2 \quad \text{when } R_a + R_b > d_{ab}, \text{or}$$

$$E_{clash} = 0 \quad \text{otherwise},$$

where R_a and R_b are the van der Waals radii of the atoms, and d_{ab} is the separation between the atoms.

Long-range main-chain hydrogen bonding in FRAGFOLD3 is now handled by generating an explicit amide hydrogen position and calculating a simple Morse potential as follows:

$$E_{hbond} = 5.5(\chi^2 - 2\chi)\cos^2\theta$$

where

$$\chi = e^{-1.327(d_{\text{H-O}}-2.1)},$$

θ is the N-H-O angle and $d_{\text{H-O}}$ is the distance between the O and H atoms.

The energy of hydrogen bonds between residues i and j is set to zero, where $j < i + 6$ and a maximum of two hydrogen bonding interactions are allowed per residue (apart from proline).

The above energy terms are applied to an all-atom representation of the polypeptide chain, and summed thus:

$$E_{total} = W_1E_{short\text{-}range} + W_2E_{long\text{-}range} + W_3E_{solv} + W_4E_{steric} + W_5E_{hbond}$$

where $W_{1\ldots5}$ are adjustable weights.

The above potential terms are summed across multiply aligned sequences, rather than just a single sequence. For the steric terms, however, the clash energies are calculated only for the target sequence.

The first stage of the folding simulation involves the selection of favorable supersecondary structural fragments at each residue position along the target sequence. Supersecondary structures are defined by taking two or three sequential secondary structures from a library of protein structures. Currently, the following supersecondary structures are defined:

- α-hairpin consecutive α-helices in a compact arrangement
- α-corner consecutive α-helices in a noncompact arrangement
- β-hairpin hydrogen-bonded consecutive β-strands
- β-corner non-hydrogen-bonded consecutive β-strands
- β-α-β unit parallel hydrogen-bonded β-strands with intervening α-helix
- Split β-α-β unit parallel non-hydrogen-bonded β-strands with intervening α-helix

The fragment selection stage of the folding procedure involves the summation of pair potential terms and solvation terms for the target sequence (and aligned homologs) threaded onto each supersecondary motif, at each position in the sequence. So for a target sequence of length L, and a motif of length M, $L - M + 1$, threadings are considered. A new feature of fragment selection used for the CASP6 predictions was to also consider fixed-length fragments (of length 9) in addition to supersecondary fragments, though still only considering fragments that agree with the reliable regions of predicted secondary structure. Secondary structure was predicted with PSIPRED[6] (version 2.4), and these predictions, along with the associated PSI-BLAST multiple sequence alignments, were used as inputs to FRAGFOLD. Note that apart from biasing the selection of fragments, secondary structure prediction information is still not used elsewhere in the FRAGFOLD method.

In addition to the sequence-specific fragment list, a general fragment list is also constructed from all tripeptide, tetrapeptide, and pentapeptide fragments from the library of highly resolved structures. These smaller fragments are not preselected.

Having selected the starting fragment lists, a single folding simulation progresses in the following way. First a random conformation for the target sequence is generated by selecting fragments entirely randomly. Fragments are spliced by superposing the α-carbon, the main-chain nitrogen, and carbonyl-carbon atoms of the C-terminus of one fragment on the equivalent atoms of the N-terminus of the other fragment. Each randomly selected fragment is spliced onto the end of the growing chain until all N residues have been covered. Having generated a random conformation for all N residues, a simple steric clash check is carried out, and the conformation is rejected if any pair of atoms (main-chain and β-carbon) is found to be closer than a predetermined minimum distance. A residue-specific table of minimum distances was used, which was compiled from a set of highly resolved protein structures (resolution better than 1.5 Å). If parts of the randomly generated chain overlap according to the table of minimum distances, then the conformation is rejected and another randomly generated conformation is selected using the same procedure. This continues until the starting conformation has no overlapping atoms.

Before the simulation starts, it is necessary to calculate the relative weighting of the components of the potential function ($W_1 \ldots W_5$). To find these weights, a number of random chain conformations (typically 1000) are gener-

ated using the above procedure. For each of these conformations, the component energy terms are calculated, and the standard deviations of each component are calculated. The ratios of these standard deviations to that of the short-range pairwise potential component are used as weights.

Given a random starting conformation and an appropriate set of weights on the component energy terms, a simulated annealing algorithm is used to search for chain conformations that produce a minimum of the energy function. A random move is made by either selecting a locally optimum fragment from the lists of preselected fragments at each position in the target sequence, or a completely free choice is made from the additional list of small fragments. Half of the moves made involve a locally optimum fragment, and the other half of the moves involve a free selection from the small fragment list. In this way, approximately half of the random moves will result in forming a supersecondary structural motif at the selected position in the polypeptide chain. Computationally, these large fragment moves allow a great deal of conformational space searching to be bypassed.

In addition to the random fragment moves, side-chain rotamers are also randomly changed after a fragment move has been allowed with one rotamer being changed per accepted fragment move.

Although FRAGFOLD offers the option of using a genetic algorithm to search conformational space, for all of the CASP6 predictions, simulated annealing was used as follows. Random moves are made as detailed above, but are accepted with a probability $e^{-\Delta E/kT}$, where ΔE is the energy change caused by the move. The starting temperature (T_0) for the simulation is selected by making 500 random moves to the starting conformation and calculating the largest absolute energy change between any two moves. The simulation is started at a temperature corresponding to 10 times this ΔE (i.e., from $E = kT$, $T_0 = 10\ \Delta E/k$). FRAGFOLD3 now uses an adaptive cooling schedule based on the standard deviation of energies sampled at the current temperature [$\sigma(T')$], where the temperature is reduced according to the expression

$$\frac{T'}{T} = e^{\frac{-0.5T}{\sigma(T')}}$$

after M random moves have been allowed, or a total of N moves have been tried. The values of M and N are chosen according to the length of the target protein as follows:

$l < 80\ M = 1000, N = 10{,}000$
$80 \leq l < 120\ M = 3000, N = 30{,}000$
$l \geq 120\ M = 5000, N = 50{,}000$

When every random move is rejected at a given temperature, then it is assumed that the current structure is "frozen." At this point, the temperature is set to zero, and a further 50 000 random moves made to allow the system to "quench."

For each prediction target, 1000 separate simulations (using different random number seed values) were run on a Linux-based compute farm. The final conformations were clustered by rigid body superposition, where two conformations were placed in the same cluster if at least 33 residues could be superposed with a distance between equivalent atoms of < 6 Å. The representatives of the five largest clusters were submitted to the CASP6 assessment. In some cases, however, less than five clusters were produced, and in some of these cases, additional models were added "by eye" considering features such as compactness and quality of sheet formation (where applicable).

Domain Docking and Microdomain Folding (FRAGFOLD-MODEL)

A variant of FRAGFOLD was used to assist in the generation of complete chain models for some targets. This version of FRAGFOLD is designed to fold only specified parts of a chain, and in CASP6 this was used in two main ways. First, for comparative models or fold recognition models where there was a large [$>$ 15 amino acids (aa)] overhang at either the N- or C-terminus, but not long enough ($<$ 40 aa) to warrant consideration as a separate domain, FRAGFOLD-MODEL was used to generate plausible conformations for these segments. Second, in cases where proteins were modeled as separate domains, in some cases, FRAGFOLD-MODEL was used to "dock" the domains together by searching possible linker peptide conformations. In both cases, certain regions of chain were preformed from previous prediction work, and although the residues in these regions were considered in the calculation of energy, the conformations of these regions were held fixed throughout the FRAGFOLD-MODEL run.

Disorder Prediction (DISOPRED2)

For all targets (including CM and FR targets), regions of native disorder were predicted using DISOPRED2.[18,19] DISOPRED2 was trained on a set of around 750 nonredundant sequences with high-resolution X-ray structures. Disorder was identified with those residues that appear in the sequence records but with coordinates missing from the electron density map. This is an imperfect means for identifying disordered residues, as missing coordinates can also arise as an artifact of the crystallization process. False assignment of order can also occur as a result of stabilizing interactions by ligands or other macromolecules complexed with the protein. However, this is the simplest means for defining disorder in the absence of further experimental investigation of the protein.

A sequence profile was generated for each protein using a PSI-BLAST search against a filtered sequence database. The input vector for each residue was constructed from the profiles of a symmetric window of 15 positions. The data were used to train linear support vector machines (SVMs). The SVM controls overfitting by ensuring that the decision surface separates the two classes with a large margin. Apart from the use of SVMs, DISOPRED2 differs from DISOPRED in that it does not use secondary structure predictions from PSIPRED to postfilter predictions, and it uses separate classifiers for the N- and C- termini of protein chains.

TABLE I. Global Distance Test Total Score (GDT_TS) for Models Submitted to the CASP6 Server

Targets	No. of residues in domain	Best model no.	GDT_TS of best model	GDT_TS of Model 1
(a) FR/A targets				
T0198	225	1	33.22	33.22
T0199_3	82	1	20.43	20.43
T0209_1	108	4	22.91	14.12
T0212	124	1	45.37	45.37
T0215	53	1	54.25	54.25
T0230	102	3	41.17	31.86
T0235_2	43	1	27.33	27.33
T0239	98	4	22.2	26.53
T0248_1	79	5	47.78	32.59
T0248_3	87	3	32.19	26.15
T0262_1	72	2	51.04	32.64
T0272_1	85	2	38.53	25.0
T0272_2	99	2	27.52	24.49
T0273	186	5	24.87	18.55
T0280_2	51	1	25.0	25.0
T0281	70	1	71.79	71.79
(b) NF targets				
T0201	94	1	49.47	49.47
T0209_2	57	4	39.48	36.84
T0216_1	209	5	15.07	10.65
T0216_2	213	4	12.56	9.27
T0241_1	117	3	24.15	20.94
T0241_2	119	4	22.06	18.49
T0242	115	5	30.65	28.26
T0248_2	87	1	45.98	45.98

Scores were calculated using the method of Zemla et al.[20] and were obtained from the CASP6 summary tables, which can be found at http://predictioncenter.org/casp6. (a) Models submitted in the fold recognition/analogous category (FR/A). (b) Models submitted in the new fold category (NF).

RESULTS

The methods described above were applied to all of the CASP6 targets according to the strategy diagram shown in Figure 1. FRAGFOLD was only applied to targets that seemed most likely to be novel folds on the basis of fold recognition results from THREADER[3] and mGenTHREADER,[12,13] but compared to CASP5, we made more use of fragment assembly methods this time around. As described, up to five predictions were submitted for each target based on the results of the structure based clustering. Table I summarizes the submitted predictions for the eight target domains that turned out to have novel folds or folds with very limited similarity to existing folds, along with the 16 target domains that had similarities to known folds but where there was no evidence of an evolutionary relationship (i.e., analogous folds). Figure 2 shows the best of our submitted predictions.

FR/A (Analogous) Targets

We submitted predictions for 16 target domains for which there was a significant similarity to a previously determined structure, but where there was no evidence for an evolutionary relationship. If we just look at the Model 1 predictions for these targets, then we can say that seven of the targets were predicted with reasonable success [Global Distance Test Total Score (GDT-TS) > 30]. If we expand the analysis to all submitted models, then nine of the 16 targets can be said to have been modeled at a similar level of prediction accuracy. Target T0281 [Fig. 2(b)] was clearly the highlight with a GDT-TS of 71.79 and a Cα root-mean-square deviation (RMSD) from the experimental structure of only 2.44 Å. Although this target had a similarity to known folds, the prediction in this case was generated by FRAGFOLD, and the accuracy of the model is likely due to the fact that we are now modeling side-chains, albeit with a restricted set of rotamer choices.

In fact, for most of the FR/A targets, we were not able to find convincing fold recognition results either using mGenTHREADER or THREADER, so we opted to submit fragment assembly predictions. The two exceptions to this were targets T0212 and T0262. In the case of T0212 [Fig. 2(a)], we were able to identify a strong match to PDB entry 1IU1 (human γ1-adaptin ear domain) using "classical" threading techniques (THREADER 3.5), and our template-based prediction for this target was in fact the third best prediction out of all submissions. Although, were one to be pessimistic, this looks like a disappointing result for threading techniques, it must be borne in mind that only a small number of targets were really amenable to prediction using this kind of approach. The majority of fold recognition cases were in the FR/H category, where we were able to identify good templates using mGenTHREADER, and the majority of FR/A targets either showed only weak similarity to a known fold or strong but local similarity. It is also worth pointing out that only five of our predictions were made using classical threading (targets T0212, T0216, T0218, T0250, and T0257) and of these, three were cancelled targets, and of the remaining two, one turned out to be an excellent prediction (T0212). Nevertheless, the boundaries between classical potential-based threading and fragment assembly grow ever more indistinct, and it is clear that the best way of tackling targets in the FR/A category will likely be a combination of template-based and fragment-based modeling.

NF (New Fold) Targets

Of the submitted predictions that eventually were determined to be new folds, four out of eight domains were predicted with reasonable accuracy.

Target T0201

Target T0201 [Fig. 2(e)] is a small protein with a simple $\alpha + \beta$ topology comprising two helices and a five-stranded sheet. Our best submitted prediction for this target had a respectable GDT-TS (49.47) and the correct number of strands and helices, but clearly did not accurately model the topology of the sheet. Looking closely at the model, it is clear that the loops leading into and out of the second β-hairpin are not long enough to allow this hairpin to fold back and complete the sheet. Looking back at all of the models generated for this target, it seems that our model

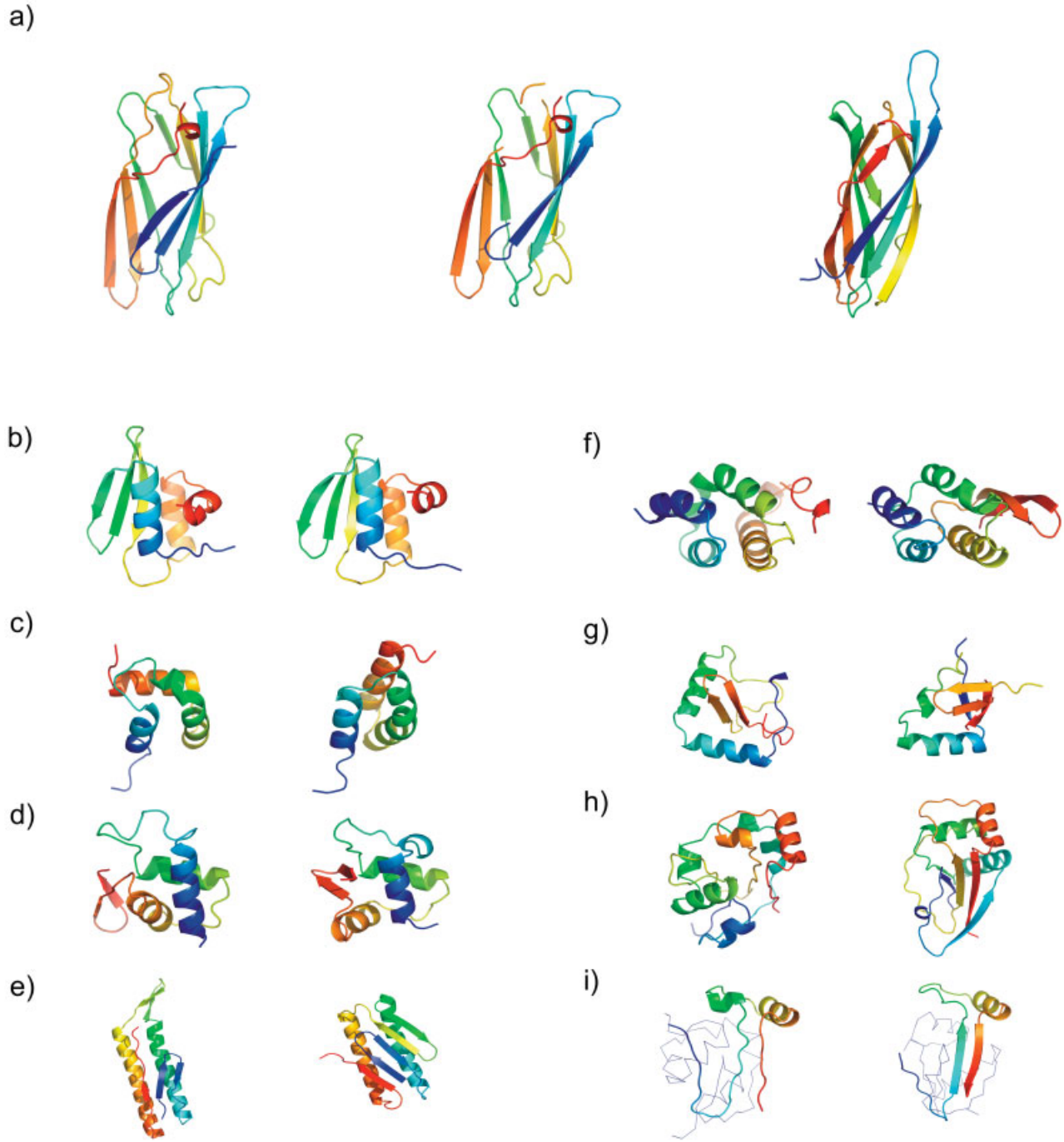

Fig. 2. Ribbon diagrams highlighting the best prediction results in the new fold (NF) and analogous fold recognition (FR/A) categories. Images were generated using PyMol.[21] In (a), the template structure is shown on the left, the submitted model, in the middle, and the observed structure on the right. In other cases, the model is on the left and the experimental structure is shown on the right. (**a**) Target T0212, predicted by classical threading (THREADER 3.5). (**b**) Target T0281 (GDT-TS, 71.79) predicted by fragment assembly. (**c**) Target T0215 (GDT-TS, 54.25) predicted by fragment assembly. (**d**) Target T0262 Domain 1 (GDT-TS, 51.04) predicted by fragment assembly. (**e**) Target T0201 (GDT-TS, 49.47) predicted by fragment assembly. (**f**) Target T0248 Domain 2 (GDT-TS, 45.98) predicted by fold recognition (nFOLD). (**g**) Target T0209 Domain 2 (GDT-TS, 39.48) predicted by fragment assembly. (**h**) Target T0242 (GDT-TS, 30.65) predicted by fragment assembly. (**i**) Best predicted fragment of Target T0242 is shown highlighted.

selection strategy was effective, as there were no significantly better models in the ensemble of 1000 models generated for this target. The reason for this failure was clearly down to the overprediction of the two helical regions using PSIPRED as a result of the lack of any sequence homologs for this target. A more accurate secondary structure prediction would almost certainly have led to better models being generated in the ensemble, and indeed when we recently reran the FRAGFOLD simulations (i.e., a "postdiction" experiment) on this target with a better

secondary structure prediction, a number of nativelike conformations were generated in the ensemble.

Target T0248 (Domain 2)

The second domain of target T0248 [Fig. 2(f)] is a small, mostly helical domain. The domain is essentially a four-helix bundle with a C-terminal β-hairpin, though the four-helix bundle has a relatively unusual topology. Perhaps the most interesting aspect of this prediction is that it was a novel fold prediction based on a structural template. The template was the B chain of 1qbk (karyopherin β2), which is a chain of 890 residues forming a helical armadillo repeat structure. In this case, an mGenTHREADER model was selected by the nFOLD algorithm with a convincing score, so this prediction was entered as our first model. This model turned out to include an accurate prediction for this domain (Cα RMSD of 6.3 Å over all residues), although the other domains were not modeled to any reasonable degree of accuracy. Given that the modeled domain (77 residues) matched only 8.7% of the template structure, it is remarkable that this local similarity was recognized so strongly. The MODCHECK signal in particular seems to have been the main reason for this model scoring so well. This does of course raise the issue of whether this is really a novel fold at all, and even if there might be an extremely distant evolutionary relationship between the proteins, though this would seem to be unlikely.

Target T0209 (Domain 2)

Domain 2 of Target T0209 [Fig. 2(g)] was predicted by FRAGFOLD with some reasonable success. The domain comprises two helices and a small half-barrel-like arrangement of four strands. In this case, the two helices are modeled quite well, along with the central β-hairpin. Unfortunately, the first and last strands do not occupy the correct locations in the sheet.

Target T0242

Despite the relatively low GDT-TS of 30.65, our best prediction [Model 5, shown in Figure 2(h and i)] for target T0242 was the best submitted prediction from all groups. Our Model 1 prediction was also the best Model 1 prediction submitted. The protein in this case has a rather unusual α + β fold, with a mixed sheet of four strands and three helices. What makes this structure particularly difficult to predict is that a large fraction of the structure is in irregular coil regions. This is compounded by the fact that few homologs are available to allow an accurate secondary structure prediction. In the predicted structure, two layers of helices are predicted, which pack against both faces of a fairly distorted sheet. In the experimental structure, the sheet is very well defined, but the second layer of helices in fact appears to be a relatively unstructured region with no regular secondary structure at all. Figure 2(i) shows the highlight of this prediction, where we accurately modeled a section with a helix and two strands, along with a long section with no regular secondary structure.

Other Targets

Table I(b) summarizes the submitted predictions for the five further domains that were classified as novel fold targets. All but one of these target domains are part of much larger target structures, and the low success of these target domains is therefore of little surprise. In almost all cases, inaccurate domain boundary assignment led to only parts of the domain structures being modeled as complete polypeptide segments, so the FRAGFOLD simulations were very unlikely to produce sensible results.

What Went Right?

In the analogous fold category (FR/A), reasonable models were submitted for up to nine out of 16 of the target domains. Most of the success came from the use of fragment assembly using FRAGFOLD, but in one case a good first model was submitted using "classical threading." Although it is hard to ignore the fact that fragment assembly is proving to be very effective at modeling small- to medium-size domains, there still seems to be some mileage left in threading methods, particularly if used in conjunction with fragment assembly for larger domain folds. The prediction for Target T0281 was of remarkable accuracy, including reasonable side-chain placement, so we are pleased that our relatively simple method for handling side-chains seems to work fairly well.

In the new fold category (NF), four of the eight submitted predictions turned out to be reasonably accurate. This is perhaps a slightly lower success rate than we have achieved in previous CASP experiments, but there were relatively few independent new fold domain targets in this round, and the majority of the targets in this category were small parts of large multidomain structures.

Another encouraging aspect was that domain parsing was less of a problem in CASP6, showing an improvement in our domain prediction strategy. Partly this was due to a new domain assignment method being employed (DomPred) but also the additional value of looking at disorder predictions, in that domain linkers appear often to be long disordered regions (LDRs). We hope to look into this as a possible new feature of future domain prediction algorithms.

Finally, we were pleased to see that our ability to rank our models had increased. More often than not, our Model 1 prediction was not only the best of the five submitted models, but also among the very best structures generated in the ensemble of 1000 structures generated. There are still some problems to be resolved, however. We generated a number of extremely good predictions for Target T0198, for example, but our existing model selection strategy did not rank these predictions in the top five clusters.

What Went Wrong and Why?

The biggest disappointment for us is that there is still slow progress in the correct folding of proteins with complex β-sheet topologies. As we have pointed out in previous CASP articles, the formation of β-sheets is a highly cooperative process requiring many regions of the polypeptide chain to converge in just one configuration,

and it is thus easy to see why fragment-based methods have such difficulty in handling this type of structure. Our improved hydrogen bonding potentials (i.e., a simple Morse potential) have resulted in better sheet conformations being generated with a greater proportion of satisfied hydrogen bonds; nevertheless, the accurate modeling of the twists of sheets, particularly those forming β-barrels, is still a major challenge. We have made some progress in combining template-based predictions with fragment-based modeling (e.g., ModelFRAGFOLD), but again, much more work needs to be done in this area.

Although there has been a lot of progress in assigning domain boundaries from sequence since CASP5, this still remains problematic in a number of cases. In cases where an incorrect domain boundary assignment is made, there is very little chance that the final predictions generated will be successful.

Finally, model selection, like the domain problem, has improved substantially since CASP5, but still a lot more progress is needed. In particular, the ability to rank very similar models is something that is still not optimal. Our current model selection strategy is able to pick out clusters of structures that have a nativelike fold, but it is not often able to make a fine distinction between, say, a model that has an RMSD of 1.5 Å and one that has an RMSD of 4.0 Å. We hope, now that our models include side-chain atoms, that a set of more discriminating model quality assessment criteria can be applied.

CONCLUSIONS

It is clear from the results presented here that, as with earlier versions of the method, FRAGFOLD3 is capable of generating compact structures with significant similarity to the experimentally determined structures even for proteins with entirely novel folds. This has been demonstrated for both α-helical proteins and proteins that include β-sheets, though the performance on proteins with sheets is markedly worse than that on mostly helical proteins. It is, however, apparent that in most cases, there still remains a large gap between the quality of structures produced by FRAGFOLD and fold recognition or comparative modeling techniques. Nevertheless, in several cases, including one startlingly clear case (Target T0281), the prediction by FRAGFOLD was superior to most, if not all, of the predictions made by fold recognition methods even though a suitable template structure did exist in the structural databases. It is now becoming clear that fragment-based prediction methods are becoming standard tools for modeling small- to medium-size protein domains where no obvious template fold can be found. There still remains the problem of how to handle larger domains, and multidomain proteins, however. Perhaps with access to high-throughput computing resources (e.g., using Grid technology) even these targets will begin to be tractable.

ACKNOWLEDGMENTS

Our thanks to the organizers and assessors of the CASP6 experiment for their hard work, and in particular we would like to thank the experimentalists for making their structures available.

REFERENCES

1. Jones DT. Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. Proteins 1997;Suppl 1:185–191.
2. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
3. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
4. Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT. Protein structure prediction servers at University College London. Nucleic Acids Res 2005;33:w36–w38.
5. Marsden RL, McGuffin LJ, Jones DT. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. Protein Sci 2002;11:2814–2824.
6. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
7. KabschW, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 1983;22:2577–2637.
8. Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32:D226–D229.
9. Altschul SF, Madden TL, Schaffer AA, Zhang J, 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.
10. Holm L, Sander C. Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 1998;14:423–429.
11. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res 2004;32:D138–D141.
12. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287:797–815.
13. McGuffin LJ, Jones DT. Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics 2003;19:874–881.
14. Sodhi JS, Bryson K, McGuffin LJ, Ward JJ, Wernisch L, Jones DT. Predicting metal binding sites in low resolution structural models. J Mol Biol 2004;342:307–320.
15. Jones DT, McGuffin LJ. Assembling novel protein folds from super-secondary structural fragments. Proteins 2003;53(Suppl 6):480–485.
16. Pettitt CS, McGuffin LJ, Jones DT. Improving sequence-based fold recognition by using 3D model quality assessment. Bioinformatics 2005;21:3509–3515.
17. Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA, Pieper U, Stuart AC, Marti-Renom MA, Madhusudhan MS, Yerkovich B, Sali A. Tools for comparative protein structure modeling and analysis. Nucleic Acids Res 2003;31:3375–3380.
18. Jones DT, Ward JJ. Prediction of disordered regions in proteins from position specific score matrices. Proteins 2003;Suppl 6:573–578.
19. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 2004;337:635–645.
20. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;Suppl 3:22–29.
21. DeLano WL. The case for open-source software in drug discovery. Drug Discov Today 2005;10:213–217.