Generalized Protein Structure Prediction Based on Combination of Fold-Recognition With De Novo Folding and Evaluation of Models

Andrzej Koliński[1]* and Janusz M. Bujnicki[2]†
[1]*Faculty of Chemistry, Warsaw University, Warsaw, Poland*
[2]*International Institute of Molecular and Cell Biology, Warsaw, Poland*

ABSTRACT **To predict the tertiary structure of full-length sequences of all targets in CASP6, regardless of their potential category (from easy comparative modeling to fold recognition to apparent new folds) we used a novel combination of two very different approaches developed independently in our laboratories, which ranked quite well in different categories in CASP5. First, the GeneSilico metaserver was used to identify domains, predict secondary structure, and generate fold recognition (FR) alignments, which were converted to full-atom models using the "FRankenstein's Monster" approach for comparative modeling (CM) by recombination of protein fragments. Additional models generated "de novo" by fully automated servers were obtained from the CASP website. All these models were evaluated by VERIFY3D, and residues with scores better than 0.2 were used as a source of spatial restraints. Second, a new implementation of the lattice-based protein modeling tool CABS was used to carry out folding guided by the above-mentioned restraints with the Replica Exchange Monte Carlo sampling technique. Decoys generated in the course of simulation were subject to the average linkage hierarchical clustering. For a representative decoy from each cluster, a full-atom model was rebuilt. Finally, five models were selected for submission based on combination of various criteria, including the size, density, and average energy of the corresponding cluster, and the visual evaluation of the full-atom structures and their relationship to the original templates. The combination of FRankenstein and CABS was one of the best-performing algorithms over all categories in CASP6 (it is important to note that our human intervention was very limited, and all steps in our method can be easily automated). We were able to generate a number of very good models, especially in the Comparative Modeling and New Folds categories. Frequently, the best models were closer to the native structure than any of the templates used. The main problem we encountered was in the ranking of the final models (the only step of significant human intervention), due to the insufficient computational power, which precluded the possibility of full-atom refinement and energy-based evaluation. Proteins 2005;Suppl 7:84–90.**

Key words: comparative modeling; folding simulation; GeneSilico; FRankenstein; CABS; model evaluation

INTRODUCTION

Methods for protein structure prediction have been divided into (1) template-based modeling, further subdivided into comparative (homology) modeling (CM) or fold recognition (FR), depending on the degree of similarity between the target and the template; and (2) de novo modeling, applicable to prediction of proteins with new folds (NF), for which no appropriate templates were available. Traditionally, these two types of methods relied on very different assumptions, typically, the principles of evolution (the "Darwinian" template-based modeling attempts to model the process of divergence of protein sequences and structures by searching for common ancestors and introducing mutations to simulate the evolutionary pathway) or the principles of physics (the "Boltzmannian" de novo modeling attempts to model the process of protein folding by simulating the conformational changes and searching for the free energy minimum).[1]

Assessments of protein structure prediction (initially Livebench2,[2] and later CAFASP3[3] and CASP5[4,5]) have demonstrated that the most successful approach for template-based modeling is that of the "metaservers" (i.e., to collect the results reported by many different FR servers and either generate a new overall ranking and select the potentially best model,[6] or to construct a hybrid from fragments of the original models).[7] A new generation of methods appeared that carry out the process of protein structure prediction by recombination of fragments of

Grant sponsor: Polish Ministry of Scientific Research and Information Technology; Grant number: PBZ-KBN-088/P04/2003. Grant sponsor: EMBO/HHMI Young Investigator Award and the Fellowship for Young Scientists from the Foundation for Polish Science (to J. M. Bujnicki).

†The authors wish that it be known that they should be regarded as joint first authors.

*Correspondence to: Andrzej Koliński, Faculty of Chemistry, Warsaw University, Pasteura 1, 02-093 Warsaw, Poland. E-mail: kolinski@chem.uw.edu.pl and Janusz M. Bujnicki, IIMCB, Trojdena 4, 02-695 Warsaw, Poland. E-mail: iamb@genesilico.pl

Received 4 April 2005; Accepted 24 May 2005

Published online 26 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20723

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

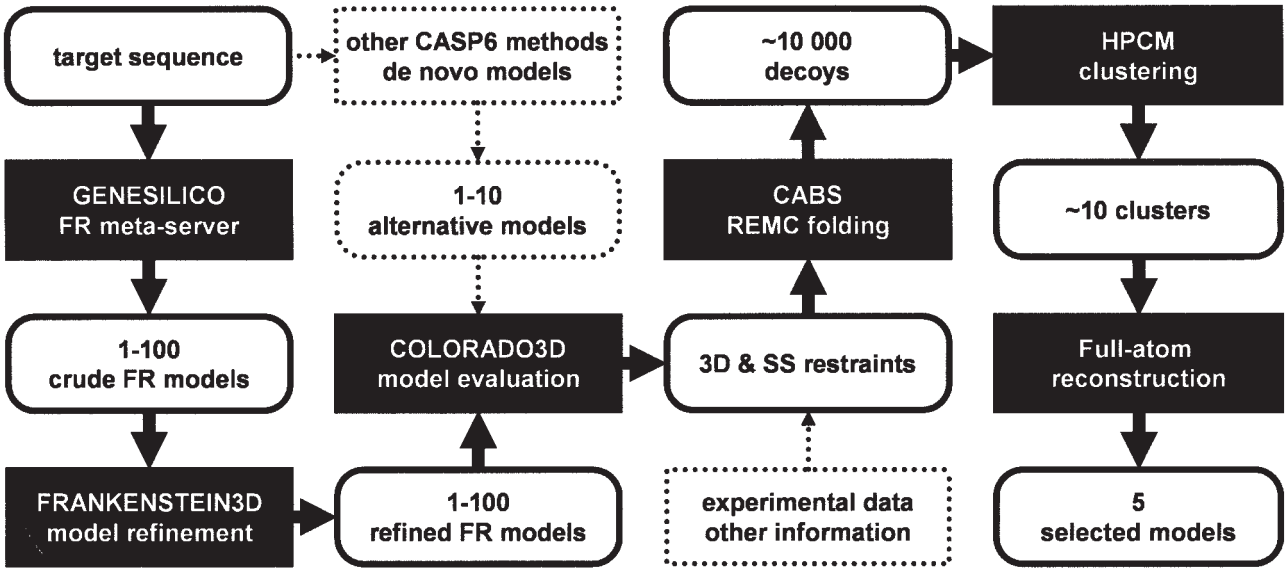

Fig. 1. A flowchart illustrating the key stages of our method. Black boxes indicate independent methods that make the pipeline. White boxes indicate sets of sequence–structure data. Thick arrows indicate obligatory stages. Thin, dotted arrows and dotted boxes indicate supplementary data and methods that can be optionally included (and have been indeed included for a few selected CASP6 targets).

other proteins at the level of the sequence alignments,[8–10] usually comprising iterations of model building, recombination, and assessment at the level of the resulting tertiary structure. As exemplified by the successful performance of our "FRankenstein's monster"[10] method in the CM category in CASP5, this approach can lead to very accurate target–template alignments, provided that a correct template is identified. This method usually outperforms the physics-based approaches in the CM category, but is generally unable to generate by reproduction models that are closer to the native structure than the combination of the best templates. Moreover, in the absence of the homologous templates (as in the NF category), even the best "Darwinian" methods are bound to fail and are usually outperformed by the "Boltzmanian" methods. According to the CASP5 evaluation, the two leading approaches for de novo modeling were the recombination of short fragments obtained from unrelated structures[11] and lattice-based folding,[12] both employing a Monte Carlo search of the conformational space. Thus, in CASP6, we decided to combine complementary strengths of our "Darwinian" (J. M. Bujnicki) and "Boltzmannian" (A. Koliński) methods.

METHODS

Our modeling strategy is illustrated by a flowchart in Figure 1.

FR Analysis, "Darwinian" modeling, and Derivation of Model-Based Restraints

Sequences of all CASP6 targets were processed by the GeneSilico structure prediction metaserver[13] at http://genesilico.pl/meta/, which is a gateway to a variety of third-party methods for prediction of protein primary and secondary structure, solvent accessibility, and protein fold recognition. The FR analysis was carried out using PDB-BLAST (a locally implemented version), FFAS03,[14] SAM-T02,[15] 3DPSSM,[16] FUGUE,[17] mGENTHREADER,[18] and SPARKS.[19] FR alignments reported by these methods were compared, evaluated, and ranked by the Pcons server[6] and structures corresponding for up to five most frequently reported folds were selected for further analysis.

For each candidate fold, the alignments between the target sequence and the structures of selected templates were used as a starting point for modeling using the "FRankenstein's monster" approach.[10] Initially, preliminary models based on unrefined FR alignments were built with MODELLER.[20] Termini and large insertions in the target without a counterpart in any template were always trimmed. Evaluation of the sequence–structure fit was carried out by VERIFY3D[21] via the COLORADO3D[22] server (http://asia.genesilico.pl/colorado3d/). A hybrid model was built from fragments conserved in $> 40\%$ of intermediate models, and the nonconsensus regions were built from fragments with highest VERIFY3D score. The hybrid model (i.e., the "FRankenstein's monster") typically exhibited numerous stereochemical problems such as steric clashes or breaks in the polypeptide chain at the junctions of fragments. Therefore, it was not directly refined, but was instead superimposed onto the template structure, yielding a new target–template sequence alignment, which was only then used to generate a new model that satisfied criteria of stereochemical "protein-likeness" implemented in MODELLER. The sequence–structure fit in the new model was evaluated again with VERIFY3D, and regions of low local score were selected for further refinement. For each poorly scored region, a set of new alignments was generated by progressively shifting the target sequence with a step of 1 aa (amino acid) in the direction of either

terminus, within the region of overlap between the secondary structure elements found in the template structure and those predicted for the target. All resulting alignments were used to generate a new family of intermediate models, which were again evaluated and recombined to produce a hybrid model. Model building, evaluation, realignment in poorly scored regions, and merging of best scoring fragments was reiterated until all regions in the protein core obtained acceptable VERIFY3D score (≥ 0.2) or their sequence–structure fit could not be improved by any manipulations. For a few targets, which exhibited very high similarity to the template structure(s) of only one fold, full-length, high-scoring models were obtained already at this stage and were submitted without any further refinement. For the great majority of the targets, however, best models obtained at this stage (1–15 for each fold) were used to derive spatial restraints from those amino acids that exhibited VERIFY3D score ≥ 0.2 to be used at the final stage of modeling (see below). Restraints were collected only from residues ≥ 7 aa apart in the sequence.

Additional Sources of Restraints

In the case of exceptionally difficult targets, where no consensus fold could be selected, additional tertiary restraints were derived from models submitted to CASP by third-party, fully automated servers for de novo structure prediction, such as ROBETTA,[23] PROSPECT,[24] or SIMFOLD[25]; we also used only fragments with VERIFY3D score > 0.2. Secondary structure restraints were derived from the consensus of results returned by PSIPRED,[26] PROFsec,[27] PROF,[28] SABLE,[29] JNET,[30] JUFO,[31] and SAM-T02.[15]

"Boltzmannian" Modeling of Full-Length Structures Based on Restraints

Tertiary restraints derived from the FR and de novo models, as well as secondary restraints derived from the consensus prediction, guided the replica exchange Monte Carlo (REMC) folding simulation using a new high-resolution reduced lattice model.[32,33] The CABS model (Cα, Cβ, and the remaining portion of the side group) employs lattice-confined Cα representation of the main-chain backbone, with 800 possible orientations of the Cα–Cα virtual bonds. The lattice spacing of the underlying simple cubic lattice is equal to 0.61 Å. As a result, the Cα trace of a Protein Data Bank (PDB) protein structure could be projected onto this lattice with the average accuracy range of 0.35 Å, without any measurable effects of the lattice anisotropy. The Cα trace provides a geometrical reference frame for the other types of united atoms: the Cβ, the center of the side group, and the center of the virtual Cα–Cα bonds. The last were employed in definition of the model main-chain hydrogen bonds. Besides the Cα's, the positions of the remaining united atoms are not restricted to the lattice. Overall accuracy of the model is limited by the assumed resolution of the potential functions describing interactions between the all types of united atoms. The force field of the CABS model contains several components that mimic averaged interactions derived from statistical analysis of the structural regularities seen in globular proteins. The effect of the solvent is treated in an implicit manner as an averaged contribution to the interaction of the side-chains. The details of the force field, including the numerical data for all potentials used could be found (and downloaded) at www.biocomp.chem.uw.edu.pl. A detailed description of the force field design can be found in recent publications.[32,33] Here, we just outline the structure of the interaction scheme. The short-range interactions contain two types of components. The generic (sequence-independent) components provide energetic biases that simulate proteinlike conformational stiffness. Sequence-dependent short-range interactions control the distances between the Cα atoms of the second, third and fourth neighbors along the chain and take into account the chirality of polypeptides. The model of the main-chain hydrogen bonds is designed in a way that mimics geometric regularity of protein secondary structure. Long-range potentials (between side groups) take into account mutual orientation of the interacting units as well as the local conformations of the main-chain fragments involved. This approach provides significantly higher specificity than simple, context-independent, statistical pairwise contact potentials.[34]

Depending on the size and difficulty of particular targets, the REMC folding simulations employed 10–50 replicas and required from a couple of hours to 10 days on 2–5 fast LINUX boxes (3.06 GHz Intel Xeon and AMD Opteron 246). In the cases of easy CM/FR targets, the initial models required relatively minor refinement simulations, which could be considered as a conformational averaging of ambiguous fragments, controlled by the force field of the CABS model. The sets of initial models (replicated if necessary) were used as the initial pools of replicas. In more difficult cases, where the initial models exhibited significant conformational divergence, or in the cases where large (20–100 aa) insertions had to be modeled, longer simulations were carried out. In all cases the sets of distance restraints were derived only from well-scored fragments of the initial models. For the putative CM/FR targets, the strength of restraints was very high, keeping the well-scoring fragment near their initial geometry during the simulations. For more difficult FR/NF targets, the strength of the restraints was very low, allowing large-scale rearrangements of the initial models. For the most difficult targets "ab initio" simulations were also carried out, using large sets of different starting conformations and relatively weak restraints superimposed on the predicted secondary structure elements. If the target was known or predicted to exhibit an oligomeric structure and for at least one of the templates the quaternary structure was known, the final monomer structure was simulated in the environment of neighboring subunits positioned as in the template, to reflect the nativelike protein–protein interaction.

After initial relaxation of the simulated lattice structures, a large number (1000–5000) of conformations were collected from the low-temperature replicas. These were

subject to the average linkage hierarchical clustering algorithm with the distance root-mean-square separation as a measure of structures similarity.[35] Depending on the level of diversity of the obtained structures, the clustering algorithm produced 1–10 clusters. These were ordered according to cluster size, average energy, and the cluster's density. For each cluster, its centroid was calculated and a full atom model rebuilt.[36] Selection of the final set of five models for submission to CASP6 was based on the combination of objective criteria, such as the average energy and the size of the respective clusters, and subjective visual analysis, to reject models that exhibited unlikely features, such as atypical angles of strands in β-sheets or rare handedness of connections between elements of secondary structure. Due to the limited computer power, in most cases we did not attempt to refine the conformations of the side-chains or to employ a full-atom energy function to score the models.

RESULTS AND DISCUSSION

Target T0198

This putative phosphate transport system regulator from *Thermotoga maritima* (now 1sum in PDB) was predicted by the GeneSilico server to comprise a tandem of PhoU domains with a short C-terminal tail. All secondary and tertiary structure prediction methods predicted that each of the PhoU domains comprises three helices, potentially forming a coiled-coil structure; nonetheless, the FR results exhibited large differences in both the type of the templates (left- and right-handed, double and triple coiled-coils) and the alignments reported. Thus, a large number of starting structures was built by the FRankenstein method and used to derive restraints for the CABS folding simulation. Comparison of the resulting models with the native structure (Fig. 2, T0198) revealed that we correctly modeled the ensembles of two N-terminal and four C-terminal helices (aa 3–71, 1.8 Å from native and 80–200, 3.9 Å from native, respectively), but we failed to model their mutual orientation. We have also mispredicted a helix in the C-terminus, which in the native structures forms a β-hairpin. This case illustrates the ability of our method to obtain a reasonably good model form a set of starting structures with different folds, but also reveals a problem with the assembly of subdomains.

Target T0201

TM1457, a putative protein from *T. maritima* (now 1s12 in PDB) has no obvious homologs in the sequence database. The fold recognition analysis suggested that it may be related to the ferredoxin fold, which comprises two layers: a β-sheet with four strands and two α-helices, but secondary structure prediction suggested that this target should contain an additional fifth β-strand in the N-terminus. The fold-recognition models were at best moderately scored by VERIFY3D, had mutually inconsistent alignments, and exhibited large differences from each other. Hence, we decided to derive additional restraints from de novo models submitted to CASP6 by ROBETTA, which had five strands, grouped together in one β-sheet with three strands and a separate β-hairpin formed by the N-terminus. All our models had a similar topology, either a ferredoxin fold or its variant with an additional N-terminal β-strand added at the edge of the β-sheet and hydrogen-bonded to the C-terminal strand in an antiparallel manner. Analysis of the native structure revealed that TM1457 has a new architecture, indeed, similar to the ferredoxin fold, but with the N-terminal strand inserted into the β-sheet between the second strand and the C-terminal strand. Thus, the N- and C-terminal strands in our model are flipped compared to the native structure. However, amazingly, the CABS refolding based on fuzzy restraints has led to a rearrangement of the two layers (helices and strands), leading to a very good superposition of the model with the native structure (Fig. 2, T0201). According to the assessment, our model_4 seems to have the best Global Distance Total Test Score (GDT_TS; 61.17) among all models submitted for this target in the course of CASP6 (our model_1 is slightly worse, but is second best according to the GDT_TS, 51.06). It is noteworthy that besides the geometry, the alignment is also very good (total Cα root-mean-square deviation (RMSD): 3.54 Å, without the swapped strands: 2.81 Å), which shows that our procedure can identify a reasonable sequence–structure fit for the correctly modeled core even if the initial alignment is poorly defined, and even if the peripheral elements have incorrect structure. We speculate that the reason for our inability to obtain a correct topology of the terminal strands was due to disregarding close-range restraints (from residues < 7 aa apart), which has led to erasure of half of nativelike restraints between the two N-terminal strands (residues 1–16) that formed a hairpin in some models and instead, promotion of restraints between the whole N-terminus and the C-terminal β-strand. As a result, the two N-terminal strands surround the C-terminal strands from both sides instead of forming a hairpin inserted into the protein core. This suggest that it may be advantageous to implement a different scheme for weighting restraints, possibly based on predicted boundaries of secondary structure elements that allow us to keep strong restraints for well-folded, sharp β-turns. Among models submitted by other predictors, we found only one (Group 450, Model 5) with the correct overall topology, which, however, exhibits incorrect alignment, wrong local and global geometry, and low GDT score (42.02).

Target T0223

TM1586, a putative nitroreductase, *T. maritima* (now 1vkw in PDB) has a very interesting quaternary structure that made it a challenging prediction target for the automatic methods, despite being in the comparative modeling class. The homology of T0223 to numerous members of the nicotinamide adenine dinucleotide (NADH) oxidase/flavin reductase superfamily was easily detected by most FR methods, with the confident scores and alignments spanning the N-terminal and central part of the sequence. Only SAM-T02 departed from the consensus and reported alignments, in which the sequence match spanned the C-

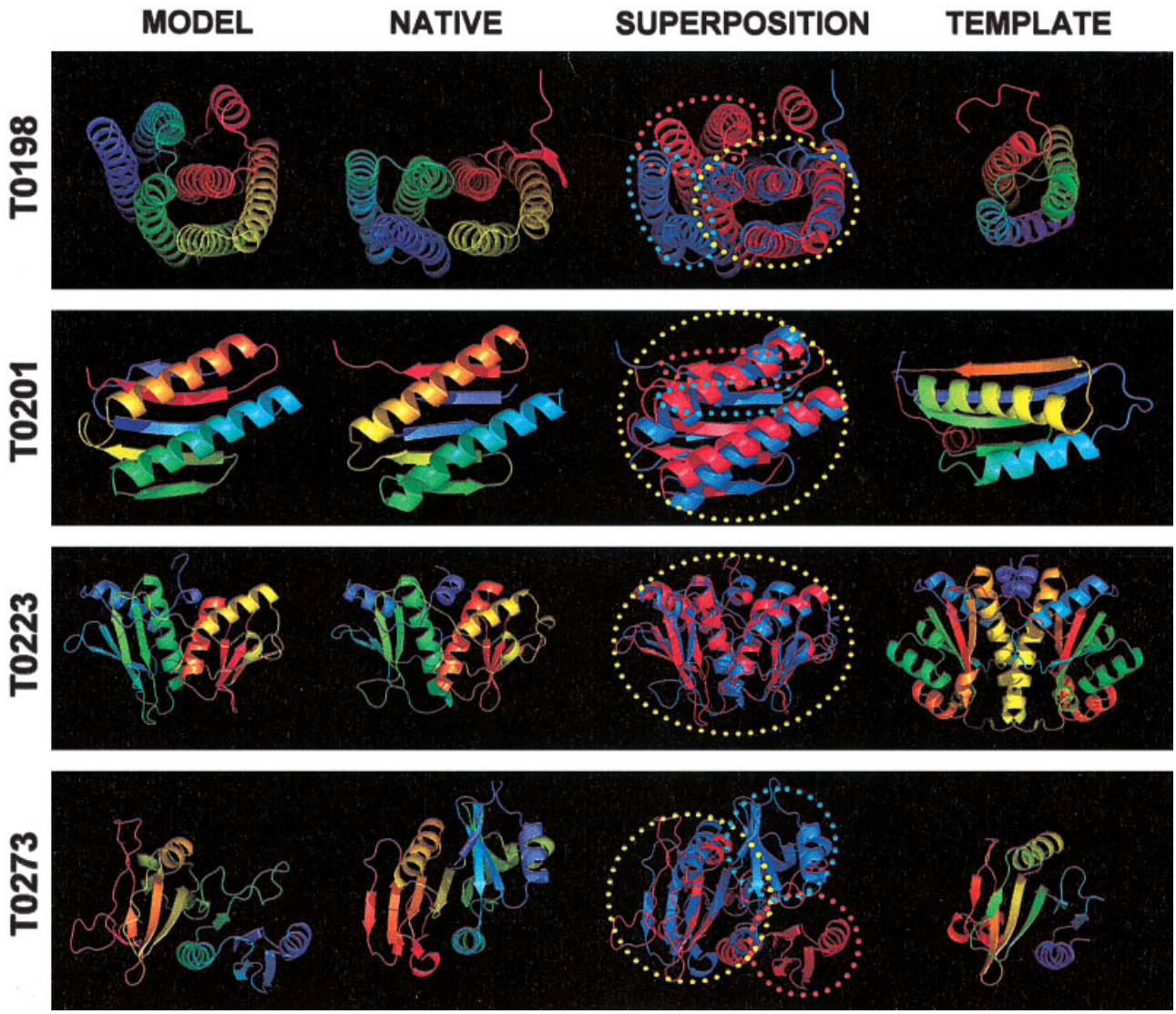

Fig. 2. Cartoon diagrams of four targets analyzed in this article (our best model, the native structure, their superposition, and the possibly best template we could identify in the PDB: T0198, 1uur; T0201, 1lxj; T0223, 2bkj; and T0273, 1gef). Proteins are colored according to the sequence index (N-terminus, blue; C-terminus, red), only in the superposition, where the model is in red and the native structure is in blue. Dotted elipses indicate correctly predicted (sub)structures according to global superposition (yellow) and regions with correctly predicted local structures, but incorrectly placed with respect to the remaining part of the protein (model, red; native, blue). The RMSD/GDT values of the model versus the native structure are as follows: T0198, 9.85/35.11; T0201, 3.54/61.17; T0223, first domain 2.98/72.37, second domain 3.29/75.27; T0273, 11.02/29.97.

terminal region. Interestingly, in both versions of the alignment, the local secondary structure predicted for the target agreed with the structure observed in the templates. It is noteworthy that the template structures were found to be homodimers. This suggested that TM1586 may be a monomeric pseudodimer (i.e., a product of intragenic duplication or fusion of two related reductase domains). Thus, we carried out the FRankenstein-type optimization of the target–template alignment using a dimeric template versus a monomeric target. Evaluation of the optimized models revealed that the N-terminal repeat obtained higher scores than the C-terminal repeat, resulting in corresponding difference in the density of restraints for the CABS folding simulation. Comparison of our best model with the native structure (Fig. 2, T0223) reveals that we correctly predicted the fold and the mutual arrangement of both domains. The values of Cα RMSD between the native structure and the three starting models were 3.2 Å, 3.9 Å, and 4.2 Å, respectively. The final model, after simulations, was better than any of the templates used with the Cα RMSD of 3.0 Å for the entire structure, and 3.1 Å and 2.8 Å for the N- and C-terminal domains, respectively. The largest inaccuracies were observed in the loop connecting the two domains (in spite of the correct mutual orientation of the domains) and in a few longer loops in the N-terminal domain. This case illustrates the advantage of combining restraints from different FR alignments that are mutually incompatible and even largely nonoverlapping. It also underscores that it may be beneficial to carry out template-based modeling and model evaluation based on the struc-

tures of multimeric complexes rather than using isolated subunits/domains. The results for T0223, as well as for some other targets, demonstrate that the proposed combination of "Darwinian" and "Boltzmannian" methods allows us to build models that are closer to the target structure than any of the templates.

Target T0273

The hypothetical cytosolic protein Tt1808 from *Thermus thermophilus* (now 1wdj in PDB) belongs to a large family of proteins with unknown function (COG4636), which are abundant in cyanobacteria.[37] Previously we have found that this family belongs to the PD-(D/E)xK superfamily of nucleases, whose best known representatives are restriction enzymes and Holliday junction resolvases,[38,39] and we modeled the structure of another member of this family, all3650. Interestingly, in most members of COG4636, one of the catalytic residues has migrated to another region in the sequence, which makes the detection of its relationship to the PD-(D/E)xK superfamily very challenging.[37] In the case of Tt1808 (as well as for all3650), there was no consensus among FR methods, and the alignments to PD-(D/E)xK proteins were poorly represented. Folding with CABS based on restraints derived from all templates failed to generate models that would be similar to the known nuclease structures. Thus, in the case of T0273, we decided to depart from the general protocol and derive the restraints only from the templates we were confident of, namely, the Holliday junction resolvases Hjc and Hje (1gef and 1ob8 in PDB), as well as from the predicted secondary structure. Still, the prediction was far from trivial, as the starting models spanned only about 50% of the target length. Comparison of the resulting models with the native structure (Fig. 2, T0273) revealed that we correctly modeled the catalytic core (aa 38–73, 83–91, and 110–145, Cα RMSD 3.7 Å), but we failed to model the mutual arrangement of peripheral elements, for which no template structure was available. As a result, the Cα RMSD for the entire structure was large, 11.9 Å. It seems that the tertiary restraints in one of the peripheral regions were too strong and resulted in contraction of one of the loops that in the native structure serves as a docking platform for the two other elements. This suggests that it may be beneficial to weight the restraints differently depending on whether they are derived from regions modeled based on secondary structures in template or inserted into the loops. However, we have also incorrectly predicted that the C-terminus of T0273 would form an α-helix by aligning it to a secondary structure element commonly present in most members of the PD-(D/E)xK superfamily. Instead, in the native structure of T0273, it forms an unusual additional β-hairpin, which has never been observed at this position in related proteins. Thus, it seems that a method for confident division of the modeled protein into the invariable core and the variable shell (that can also include peripheral secondary structures) would be very helpful. Remarkably, T0273 turned out to be so difficult as a prediction target that even the CASP6 assessors initially classified it as a NF, even though, in our opinion, it is a clear case of homologous FR. Also most groups failed to identify the fold of T0273; thus, our poor model turned out to be one of the best among those submitted to CASP6. For a more detailed description of sequence analyses, modeling of a relative of T0273 and comparison of its model with the native structure, see Ref. 37.

CONCLUSIONS: WHAT WENT RIGHT AND WHAT WENT WRONG

We developed a novel method for protein structure prediction that combines the FRankenstein algorithm for "Darwinian" comparative modeling with the CABS algorithm for "Boltzmannian" protein folding. Nearly all steps of our method were fully automated. Considering targets from all classes, according to the CASP6 rankings, we were significantly outperformed only by one modeler (Ginalski), who extensively used intervention in the process of model building based on his intuition (i.e., something we also successfully applied in CASP5,[10] but now have purposely limited to the minimum in order to make our method objective and scalable). We find it especially advantageous that our method is able to systematically improve the global conformation and the sequence–structure fit, even if initiated with partially incorrect structures with poor alignments. Thus, our approach seems to be among the best methods for protein structure prediction that rely on computation rather than expert judgment (although it obviously allows expert intervention at nearly all stages). Among the individual categories, we did not score very high in FR (within the top 10, but not at the very top of the rankings), which we attribute to the limited number of FR methods implemented in our metaserver during CASP6 and the deficiency of the procedure for consensus calculation, which is an obvious area for the future improvement.

The major drawback was our inability to correctly rank the top models: This is most clearly seen in the NF category, where we are outperformed by the Baker group if only the first model is considered, but where we excel in the ranking if the best model of five is considered. Clearly, our subjective ranking of models for submission was suboptimal. We speculate that an objective ranking based on full-atom reconstruction and refinement of the decoys using more refined energy functions (possibly physics-based rather than statistics-based) would allow us to improve the generation and identification of more native-like models. This, however, requires much larger computing power than was available for us during the CASP6 experiment. Another drawback of our methodology for CASP6, also related to the limited computing resources, was to make predictions (especially CABS simulations) only for the target sequence. As demonstrated by several other leading groups (e.g., that of Baker), considering results obtained not only for the target sequence but also for other homologous proteins seems to consistently improve the coverage of the conformational space. Such procedure might be actually helpful not only in generation of better decoys but also (what is probably more important) it might facilitate better selection of the final models. Our goal for the CASP7 experiment will be to improve the

method in the aforementioned areas, as well as to fully automate the FRankenstein/CABS pipeline and, we hope, with the availability of more computing power than presently, to provide it to the community as a server.

ACKNOWLEDGMENTS

The FRankenstein method could not exist without the availability of third-party methods and servers included in the GeneSilico metaserver. J. M. Bujnicki would like to thank their developers, in particular: David Baker, Geoff Barton, Roland Dunbrack, David Eisenberg, Arne Elofsson, Lukasz Jaroszewski, David Jones, Adam Godzik, Kevin Karplus, Lawrence Kelley, Jarek Meller, Jens Meiler, Kenji Mizuguchi, and Burkhard Rost. We would like to thank members of our groups for their technical assistance during CASP6, for the participation in the development of our methods, and for useful comments on the manuscript. Special thanks to Janek Kosiński for generating images for Figure 2.

REFERENCES

1. Bystroff C, Shao Y. Modeling protein folding pathways. In: Bujnicki JM, editor. Practical bioinformatics: Vol. 15. Nucleic acids and molecular biology. Berlin: Springer-Verlag; 2004. p 97–122.
2. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. LiveBench-1: continuous benchmarking of protein structure prediction servers. Protein Sci 2001;10:352–361.
3. Fischer D, Rychlewski L, Dunbrack RL, Jr., Ortiz AR, Elofsson A. CAFASP3: the third critical assessment of fully automated structure prediction methods. Proteins 2003;53(Suppl 6):503–516.
4. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins 2003;53(Suppl 6):352–368.
5. Kinch LN, Wrabl JO, Krishna SS, Majumdar I, Sadreyev RI, Qi Y, Pei J, Cheng H, Grishin NV. CASP5 assessment of fold recognition target predictions. Proteins 2003;53(Suppl 6):395–409.
6. Lundstrom J, Rychlewski L, Bujnicki JM, Elofsson A. Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci 2001;10:2354–2362.
7. Fischer D. 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins 2003;51:434–441.
8. John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 2003;31:3982–3992.
9. Contreras-Moreira B, Fitzjohn PW, Bates PA. In silico protein recombination: enhancing template and sequence alignment selection for comparative protein modelling. J Mol Biol 2003;328:593–608.
10. Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM. A "FRankenstein's monster" approach to comparative modeling: merging the finest fragments of fold-recognition models and iterative model refinement aided by 3D structure evaluation. Proteins 2003;53(Suppl 6):369–379.
11. Bradley P, Chivian D, Meiler J, Misura KM, Rohl CA, Schief WR, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, Strauss CE, Baker D. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. Proteins 2003; 53(Suppl 6):457–468.
12. Skolnick J, Zhang Y, Arakaki AK, Kolinski A, Boniecki M, Szilagyi A, Kihara D. TOUCHSTONE: a unified approach to protein structure prediction. Proteins 2003;53(Suppl 6):469–479.
13. Kurowski MA, Bujnicki JM. GeneSilico protein structure prediction meta-server. Nucleic Acids Res 2003;31:3305–3307.
14. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles: strategies for structural predictions using sequence information. Protein Sci 2000;9:232–241.
15. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins 2003;53(Suppl 6):491–496.
16. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
17. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001;310: 243–257.
18. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
19. Zhou H, Zhou Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 2004;55:1005–1013.
20. Fiser A, Sali A. Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol 2003;374:461–491.
21. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature 1992;356:83–85.
22. Sasin JM, Bujnicki JM. COLORADO3D, a web server for the visual analysis of protein structures. Nucleic Acids Res 2004; 32(Web Server issue):W586–W589.
23. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 2004;32(Web Server issue):W526–W531.
24. Guo JT, Ellrott K, Chung WJ, Xu D, Passovets S, Xu Y. PROSPECT-PSPP: an automatic computational pipeline for protein structure prediction. Nucleic Acids Res 2004;32(Web Server issue): W522–W525.
25. Chikenji G, Fujitsuka Y, Takada S. A reversible fragment assembly method for de novo protein structure prediction. J Chem Phys 2003;119:6895–6903.
26. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000;16:404–405.
27. Rost B, Yachdav G, Liu J. The PredictProtein server. Nucleic Acids Res 2004;32(Web Server issue):W321–W326.
28. Ouali M, King RD. Cascaded multiple classifiers for secondary structure prediction. Protein Sci 2000;9:1162–1176.
29. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. Proteins 2004;56:753–767.
30. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502–511.
31. Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. Proc Natl Acad Sci USA 2003;100:12105–12110.
32. Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A. Protein fragment reconstruction using various modeling techniques. J Comput Aided Mol Des 2003;17:725–738.
33. Kolinski A. Protein modeling and structure prediction with a reduced representation. Acta Biochim Pol 2004;51:349–371.
34. Pokarowski P, Kloczkowski A, Jernigan RL, Kothari NS, Pokarowska M, Kolinski A. Inferring ideal amino acid interaction forms from statistical protein contact potentials. Proteins 2005;59: 49–57.
35. Gront D, Kolinski A. HCPM—program for hierarchical clustering of protein models. Bioinformatics 2005;21:3179–3180.
36. Feig M, Rotkiewicz P, Kolinski A, Skolnick J, Brooks CL III. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins 2000;41:86–97.
37. Feder M, Bujnicki JM. Identification of a new family of putative PD-(D/E)XK nucleases with unusual phylogenomic distribution and a new type of the active site. BMC Genomics 2005;6:21.
38. Bujnicki JM. Phylogeny of the restriction endonuclease-like superfamily inferred from comparison of protein structures. J Mol Evol 2000;50:39–44.
39. Bujnicki JM. Crystallographic and bioinformatic studies on restriction endonucleases: inference of evolutionary relationships in the "midnight zone" of homology. Curr Protein Pept Sci 2003;4:327–337.