PROTEINS: Structure, Function, and Genetics 53:369–379 (2003)

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

Jan Kosinski, Iwona A. Cymerman, Marcin Feder, Michal A. Kurowski, Joanna M. Sasin, and Janusz M. Bujnicki*
Bioinformatics Laboratory, International Institute of Molecular and Cell Biology, Trojdena 4, 02-109 Warsaw, Poland

ABSTRACT **We applied a new multi-step protocol to predict the structures of all targets during CASP5, regardless of their potential category. 1) We used diverse fold-recognition (FR) methods to generate initial target-template alignments, which were converted into preliminary full-atom models by comparative modeling. All preliminary models were evaluated (scored) by VERIFY3D to identify well- and poorly-folded fragments. 2) Preliminary models with similar 3D folds were superimposed, poorly-scoring regions were deleted and the "average model" structure was created by merging the remaining segments. All template structures reported by FR were superimposed and a composite multiple-structure template was created from the most conserved fragments. 3). The average model was superimposed onto the composite template and the structure-based target-template alignment was inferred. This alignment was used to build a new (intermediate) comparative model of the target, again scored with VERIFY3D. 4) For all poorly scoring regions series of alternative alignments were generated by progressively shifting the "unfit" sequence fragment in either direction. Here, we considered additional information, such as secondary structure, placement of insertions and deletions in loops, conservation of putative catalytic residues, and the necessity to obtain a compact, well-folded structure. For all alternative alignments, new models were built and evaluated. 5) All models were superimposed and the "FRankenstein's monster" (FR, fold recognition) model was built from best-scoring segments. The final model was obtained after limited energy minimization to remove steric clashes between sidechains from different fragments. The novelty of this approach is in the focus on "vertical" recombination of structure fragments, typical for the *ab initio* field, rather than "horizontal" sequence alignment typical for comparative modeling. We tested the usefulness of the "FRankenstein" approach for non-expert predictors: only the leader of our team had considerable experience in protein modeling - he registered as a separate group (020) and submitted models built only by himself. At the onset of CASP5, the other five members of the team (students) had very little or no experience with modeling. They followed the same protocol in a deliberately naïve way. In the fourth step they used solely the VERIFY3D criterion to compare their models and the leader's model (the latter regarded only as one of the many alternatives) and generated the hybrid or selected only one model for submission (group 517). In order to compare our protocol with the traditional "one target-one template-one alignment" approach, we submitted (as a separate group 242) models selected from those automatically generated by all CAFASP servers (i.e. obtained without any human intervention). Here, we compare the results obtained by the three "groups", describe successes and failures of the "FRankenstein" approach and discuss future developments of comparative modeling. The automatic version of our multi-step protocol is being developed as a meta-server; the prototype is freely available at http://genesilico.pl/meta/. Proteins 2003;53:369–379.**

Key words: homology modeling; bioinformatics; GeneSilico; consensus generation; model evaluation

INTRODUCTION

Assessments of protein structure prediction (CASP,[1] CAFASP,[2] Livebench[3]) have demonstrated that fold-recognition (FR) methods can identify remote similarities when standard sequence search methods fail, but the reported target-template alignments are often only partially correct, leading to models with misfolded parts. The use of additional information, such as secondary structure (SS), and/or localization of ligand-binding residues can help to improve the target-template alignments. Moreover, models constructed from multiple parents were often

Grant sponsor: Polish State Committee for Scientific Research and EMBO & HHMI Young Investigator Programme; Grant numbers: 6P04A01124 and 3P05A02024.

*Correspondence to: Janusz M. Bujnicki, Bioinformatics Laboratory, International Institute of Molecular and Cell Biology, Trojdena 4, 02-109 Warsaw, Poland. E-mail: iamb@genesilico.pl

Received 14 February 2003; Accepted 8 April 2003

found to be more accurate than models constructed from single parents only. The final prediction accuracy can be therefore improved if the best fragments obtained from various FR alignments can be judiciously combined to generate a consensus model. Interestingly, in CASP4, several FR methods were shown to perform quite well also in the comparative modeling (CM) category.[4] The performance difference between the human experts and computer predictors continues to narrow, which suggests that most of the refinement procedures used by humans can be fully automated. These observations led us to consider that the FR-based consensus approach may be applicable not only to modeling based on remotely related templates, where the critical issue is to identify the correct template, but to _classical_ comparative modeling as well, where the possible templates can be identified relatively easily, but the real challenge is to obtain a perfect sequence alignment.

In CASP4, one of us (J.M.B.) participated as a member of the *BioInfo.PL* duumvirate, as well as one of four experts of the *CAFASP-consensus* group. Within *BioInfo.PL*, J.M.B. was responsible for building and refinement of models for most of the targets in the HM and FR categories, while in *CAFASP-consensus*, he participated in the identification of the best models generated automatically by FR servers and in inference of a rational consensus between them. While the unrefined predictions gave *CAFASP-consensus* the overall rankings of 7th and 26th in the FR and CM categories, respectively, the refined predictions gave *BioInfo.PL* even better score (5th in FR, 4th in CM).[4,5] However, it was not always clear if the improvement stemmed from application of different criteria for selection of the best automated models as the starting points for modeling by the two groups or from different degree of model refinement.

In CASP5, we attempted to assess the applicability of the FR-based consensus approach in targets beyond the "core" FR category, i.e. also in the CM (easy modeling) and FR/NF (extremely hard modeling) categories. Since it is impossible to distinguish between NF (novel fold) and FR/NF targets a priori, we applied the same modeling protocol to all targets, regardless of their apparent difficulty. Because our modeling protocol was quite complex (see below), we were also interested in testing if it can be useful in the hands of non-experts, compared to simple selection of one of the crude models generated by the individual FR servers. Previously, comparisons of human-refined models and crude automatic models have been made.[2,6] Fully automated protein structure prediction is usually applied in large-scale analyses for instance on a genome scale, where human intervention is not feasible for practical reasons (refined modeling of thousands of proteins would take thousands of person-hours). However, many researchers are interested in protein structure prediction only for one particular protein at a moment and they usually have some knowledge about the prediction target (for instance knowledge of the catalytic residues), which can be applied to selection of the potentially best model from several alternatives. Hence, in our opinion it would be informative to compare the quality of predictions made by non-experts, who either select crude models from the CAFASP set based on their agreement with data from the literature or generate refined models according to an elaborate, multi-step "expert" modeling protocol.

METHODS

At the outset of CASP5, our group comprised one protein modeling expert (J.M.B.) and five students with background in experimental biology, only fundamental knowledge of protein structure and function, and little or no experience with protein structure prediction and modeling. We attempted to mimic three possible real-life scenarios:

i) a group of biologists with at most basic knowledge of protein modeling attempts to obtain the possibly most useful model from the set of alternatives provided by publicly available automatic servers,
ii) a protein modeling expert uses a refined protocol and his intuition to build and refine the model,
iii) the same biologists as in scenario i) are provided with the model built by the expert in scenario ii), but they do not fully trust his prediction and attempt to use the expert's refinement protocol and select the "best" model for submission based not on intuition, but on an acclaimed objective method for evaluation of protein structures.

Hence, we submitted predictions as three independent groups, and applied the same sets of protocols for submission of all targets, irrespective of their potential classification to the CM, FR and NF categories:

Group 242 *GeneSilico-servers-only* (unrefined FR models selected from the CAFASP results)

Group 020, *Bujnicki-Janusz* (models refined by a single experienced predictor),

Group 517 *GeneSilico* (consensus obtained after objective evaluation and comparison of models generated independently by six members of our team).

Selection of the best automated model by *GeneSilico-servers-only* in CASP5 was carried out in a similar manner to that of *CAFASP-consensus* in CASP4, but in a more disciplined way. For instance, in *CAFASP-consensus*, preparation of the model involved limited human intervention when homology modeling programs failed to generate any reasonable structures from the FR alignments because of large deletions or insertions in the protein core (in these cases gaps in the alignment were shifted to the surface-exposed regions). On the other hand, human intervention of the *GeneSilico-servers-only* group in CASP5 involved only selection of one of the FR alignments or one of the atomic models generated by HM or *ab initio* servers. The criteria for selection included: the formation of a compact globular structure (assessed visually using RASMOL),[7] the functional similarity between the target and the template according to the literature and the SCOP database,[8] the alignment of the functionally important

and/or conserved residues, and the agreement between the secondary structure in the target and in the template. Similar criteria could be easily applied in the real-life scenario, by real researchers interested in obtaining a crude model of their protein. Automated FR models submitted by *GeneSilico-servers-only* were all based on single templates. Most of them were submitted in the AL format without explicit modeling of sidechains and insertions or deletions (indels) to avoid the inevitable distortion of the raw data by automatic homology modeling in cases, such as disruption of the protein core. Occasionally, full-atom models built by *ab initio* and homology modeling servers were selected for submission.

Previously, in J.M.B.'s hands the carefully refined multiple sequence alignments gave much better results as FR queries compared to single sequences, even in those FR servers which used their own BLAST utility to construct alignments (our unpublished data and CASP4 results: the performance of *BioInfo.PL* vs *CAFASP-consensus* group). In many cases we observed significant improvement of the prediction quality if sequences from unfinished genomes were included in the alignment, when the position of indels was manually refined and when highly diverged parts of the alignment or long insertions in the target were deleted. However, none of the available FR metaservers[9–11] offered satisfactory options to build and process user-defined alignments. In order to reduce the workload and simplify submission of different variants of prediction jobs for the same target to multiple servers, we developed a novel metaserver (http://genesilico.pl/meta).[12] It serves as a gateway to many of the FR servers available via the CAFASP metaserver (at the time of the writing: PDB-BLAST, 3DPSSM,[13] BIOINBGU,[14] FFAS-03,[15] FUGUE 2.0,[16] MGENTHREADER,[17] RAPTOR,[18] and SAM-T02[6]) and offers a few options not available elsewhere, including submission of user-defined sequence alignments and generation of many variants of the consensus sequence.[12]

Figure 1 shows the flow-chart of our sequence analysis and modeling strategy. As a prerequisite (step 0) for the refined modeling analysis carried out by *Bujnicki-Janusz* and *GeneSilico* in CASP5, as many homologs of the target sequence as possible were identified and included in the alignment. For this purpose, we created a database of putative translation products (length>20aa) of all unfinished genomes, whose sequences were publicly available. Combined with the non-redundant database (NCBI), this allowed a roughly two-fold increase of the size of the database used in local PSI-BLAST[19] searches. In a few cases, it allowed to increase the number of homologs of the target from ca. 5 to over 20 and hence, much better delineation of conserved and variable regions. The PSI-BLAST output (aligned sequence fragments) was saved as a multiple sequence alignment following the removal of all columns with > 30% gaps. Full-length sequences were retrieved from the ENTREZ database, and realigned by CLUSTALX,[20] using the "align sequences to the profile" option (the PSI-BLAST output served as the "profile"). Alignments were refined manually and used to divide the query sequence into domain-size fragments, which were submitted to our new metaserver as independent prediction queries (http://genesilico.pl/meta). For the full-length sequences, the FR results (target-template alignments in the AL format) as well as comparative and *ab initio* models (full-atom structures in the TS format) were obtained from the CAFASP metaserver (http://bioinfo.pl/cafasp/). Our server was also used to carry out FR analysis for the full-length alignments and for the alignment sections corresponding to the individual domains. Three options were used: i) columns with >30% of gaps were deleted (i.e. only the core regions were analyzed), ii) gaps were treated as unknown characters (X) (i.e. the variable regions of the target sequence were "extended" to the size of the entire alignment, using the longest insertions present in homologous sequences as the reference), iii) the consensus sequences were generated for submission as additional single-sequence jobs, using the majority-rule criterion for selection of the most frequently observed amino acids in each position or the BLOSUM matrix criterion. All results for each CASP5 target (regardless of its category), its parts, corresponding multiple sequence alignments and consensus sequences were collected. For consensus and "core only" models, the original length and the amino acid sequence of the prediction target was restored by introducing insertions and deletions (indels) into the corresponding FR alignments. All alignments were converted to a common format, including the sequence of the target (or one of its domains) and the template.

In the first step of the modeling protocol, all FR alignments were converted into preliminary full-atom models by comparative modeling using MODELLER 6v1[21] with default parameters. For each CASP5 target, a database of models was created using the FR-based homology models and full-atom *ab initio* and homology models obtained from the CAFASP website. All these models were evaluated by VERIFY3D,[22] using the atypically small window size of 5 in order to identify well- and poorly-folded fragments of the size comparable to the smallest secondary structure elements.

In the second step, preliminary models were divided into clusters based on the relationship of the template folds, according to SCOP.[8] Within each major cluster (> 5 FR alignments), all models were superimposed with SWISSP-DBVIEWER and the superposition was used to generate a multiple sequence alignment. Structurally superimposable regions were identified and analyzed for the consistency of the alignment and the quality of the sequence-structure fit, according to VERIFY3D. A hybrid "consensus" model was created from the well-scored fragments (>10 aa) of models corresponding to the most frequently reported alignments. If there were several alternative clusters (alternative folds), the consensus model composed of best scoring fragments over the entire length of the target sequence was selected for further analysis (step 3). If no recurring fold could be identified among the FR results (no major clusters with > 5 superimposable models), the best-scoring preliminary model was selected for further analysis (step 4, see below). Alternatively, the analysis was halted with the conclusion that the target most likely

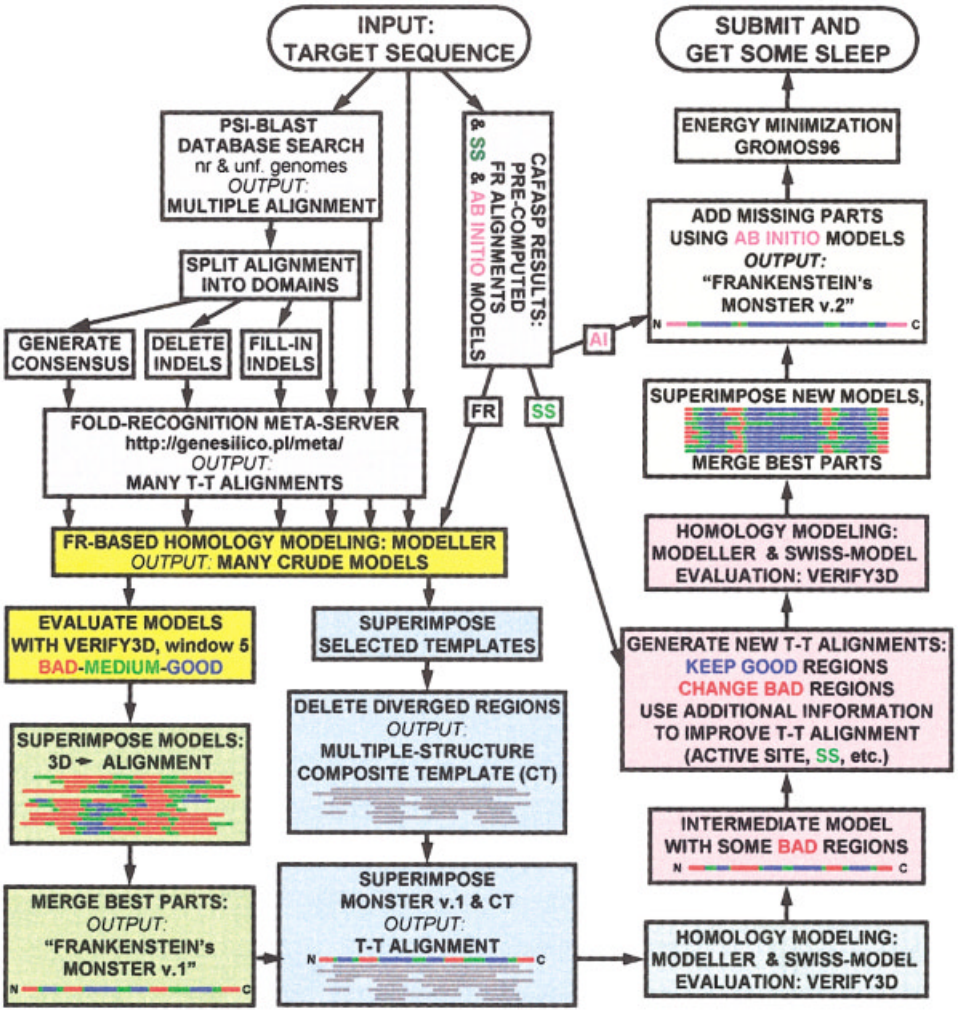

Fig. 1. A flowchart illustrating the major stages of construction of the "FRankenstein's monster"–from the target sequence to successful model submission. The individual steps are color coded: step 0 (pre-modeling, sequence analysis)–gray; step 1 (generation of crude models)–yellow; step 2 (generation of the first hybrid model)–green; step 3 (generation of the composite template and construction of the first "real" comparative model based on the structural alignment between the composite template and the first hybrid model)–cyan; step 4 (sampling of the alignment space)–pink; step 5 (creation of the final hybrid model and unleashing the "FRankenstein's monster")–cornsilk.

represents a novel fold and neither a suitable template exists in PDB nor any attractive models were generated by *ab initio* servers for CAFASP.

In the third step, a multiple structure alignment was created by pairwise superposition of all template structures and the consensus model of the target. The template most frequently reported by FR was used as a reference structure. The most diverged elements of the template structures (for instance large insertions not present in other templates and in the consensus model) were removed and the remaining parts of the template structures were regarded as a composite multiple-structure template. Based on this superposition, a sequence alignment was inferred. The alignment of the target sequence to the composite multiple-structure template was used to build a new (intermediate) comparative model of the target, using MODELLER[21] and SWISSMODEL.[23] We used both programs, because we observed that SWISSMODEL introduces less distortion into the protein core (compared to the structure of the template), while it sometimes has problems with insertions or ligation of loose ends generated by deletions. On the other hand, MODELLER is more "sloppy" in modeling the protein core, but has an extraordinary ability to accommodate indels in virtually any region of the protein structure, including "prohibited" regions of the protein core (our unpublished observations). This was also the reason for using only Modeller for generation of the preliminary models based on FR alignments (step 1), which often have indels in "prohibited" regions of the template structure.

In the fourth step, the quality of the local structure of the intermediate model (both the MODELLER and SWISS-MODEL versions) was again evaluated using VERIFY3D. For all regions with unsatisfactory VERIFY3D scores in

either version of the intermediate model, series of alternative alignments were generated by progressively shifting the "unfit" sequence fragment in either direction. Here, additional information, such as secondary structure, placement of insertions and deletions in loops, conservation of putative catalytic residues, and the necessity to obtain a compact, well-folded structure was taken into account. Thereby, the sequence/structure space could be explored beyond the alignment variants reported by FR servers. Importantly, only one region was shifted at a time to avoid interference of effects from different parts of the model structure. Usually, the limits of the shift in either direction were dictated by the criterion of at least partial overlap of predicted vs observed secondary structure elements. For all these alternative alignments, new models were built (again, using both SWISSMODEL and MODELLER) and evaluated using VERIFY3D.

In the fifth step, all models were superimposed using SWISSPDBVIEWER and the "FRankenstein's monster" model was built from best-scoring segments. An additional criterion for selection of loops was the similarity of their conformation to the corresponding loops in at least one of the template structures. Typically, the core elements were taken from the models built using SWISSMODEL, while the loops (especially the problematic ones) were often taken from the models built by MODELLER. The final model was obtained after manual adjustment of conformation of selected side-chains followed by limited energy minimization with GROMOS96[24] (200 cycles of steepest descent) to remove steric clashes between sidechains from different fragments.

The same multistep modeling protocol was used by the five novice members of the team. They used the same set of FR alignments, but conducted their analysis independently, especially with respect to the choice of fragments to construct hybrid models (in all steps) and generation of alternative alignments (in step 4). The only difference was that the manual adjustment of side-chains in the final "FRankenstein's monster" model (carried out by the expert, step 5) was replaced by automated rotamer selection using SCWRL.[25] As a result, up to 5 alternative models were generated by the student members of the *GeneSilico* group. They were compared with the model obtained by the expert and all were assessed with VERIFY3D. All models were treated as equal, i.e. no higher weight was assigned to the model built by the expert. If one model with an outstanding score emerged, it was selected for submission. Otherwise, a hybrid model was constructed from the best-scoring parts and the side-chains were re-modeled using SCWRL. It is noteworthy that in all cases, the fold selected by the 5 novice members agreed with that identified by the expert, however for some targets either the expert or some members of the team concluded (in step 2) that modeling is unfeasible and no good candidate for submission can be proposed. In these cases, the selection of the best model or the construction of the hybrid for the final submission by the *GeneSilico* group was made based on less than 6 models. In a few cases, the *GeneSilico* group submitted more than one model, if no confident decision could be made based on the VERIFY3D evaluation.

Furthermore, the final model and all the well-scoring parts of the intermediate models were used to calculate the average residue-residue separation distances (submitted as the RR prediction category). The secondary structure of the final models was inferred according to DSSP and combined with the independent sequence-based predictions (obtained from CAFASP or calculated in-house) to generate the output in the SS format. For our own alignments, we made SS predictions based on consensus of JPRED,[26] PSIPRED,[27] SSPRO,[28] PROF,[29] and SAMT02.[6] For targets with no reliable models of the tertiary structure, the independent SS prediction was based solely on alignment-based predictions. *Bujnicki-Janusz* and *GeneSilico* calculated their RR and SS predictions independently. *Bujnicki-Janusz* (group 020) has also submitted order/disorder (DR) predictions, based on combination of SS prediction (in particular the presence of long "coil" regions), analysis of structural divergence, R-factors and conformational variability in template structures, and identification of compositionally-biased sequence regions.

RESULTS AND DISCUSSION

We were interested in analyzing the relative ranking of the three groups (242, 020, and 517) and the official assessment of their performance in relation to each other and to the other groups. It is noticeable that in all categories (CM, FR, and NF) both the expert (020) and the group of non-experts using the expert's protocol and having access to the expert's models (517) consistently outperformed group 242 i.e. CAFASP-consensus-like selection of crude FR models. The models built according to the elaborate refinement protocol (Figure 1) were very good, much better than the unrefined models selected from between the pre-computed CAFASP results, suggesting that this protocol may be a valuable tool for protein predictors and is probably worth automating in the future.

It should be mentioned that *Bujnicki-Janusz* tended to depart from the rigorous protocol more often than *GeneSilico* and sometimes selected the models for submission based on his intuition rather than the VERIFY3D score. Since *GeneSilico* had the access to *Bujnicki-Janusz*'s models and more consistently used the refinement procedure, we expected *GeneSilico* to consistently outperform *Bujnicki-Janusz*. We also expected that our strategy, focused on generation of possibly best FR alignments, will result in superior performance in the FR category rather than in the CM category. However, in the CM category *Bujnicki-Janusz* and *GeneSilico* scored among the top groups, while in the FR category both groups obtained quite good scores (with the mutual position in the ranking depending on the evaluation method), but not as good as the absolute top groups (see the assessors' papers in this issue of Proteins). It was very surprising for us that we performed remarkably well in the CM category rather than in the FR category. We find it noteworthy that in the FR category we were outperformed, among the others, by two other groups (453 and 006), who also used the consensus FR approach

and VERIFY3D scoring, albeit with one important modification: the key step in their strategy of model selection was to compare the models with each other using the 3DJury system and thereby to identify the most commonly occurring fold and superfamily (see the articles by Rychlewski and Ginalski in this issue of Proteins). Detailed comparison of our and their models in the CM and FR categories is beyond the scope of this article. However, we rationalize our failure to "win" the FR category by our inferior performance in selection of the correct fold or superfamily for modeling. On the other hand, our refinement strategy allowed us to produce exceptionally good alignments between the target and multiple template structures in most cases where we successfully identified the correct fold and the best set of templates, hence opening the door to success in the CM category. In the following section of this article we will try to analyze in detail a few of the predictions that were particularly successful due to rigorous application of a few key steps of our protocol, but also revealed a few remarkable shortcomings or failures. The GDT_TS scores were obtained from the CASP5 website (http://predictioncenter.llnl.gov/casp5/Casp5.html)

Target T0172

T0172, a hypothetical MraW methyltransfrase, is a member of the Rossmann-like fold methyltransferase (RFM) superfamily.[30] RFMs are characterized by the presence of a common catalytic domain with nine motifs (numbered I-VIII & X) and variable insertions and/or fusions with unrelated domains. The region corresponding to the cofactor-binding site (motifs X & I-III) is relatively well-conserved, while the neighborhood of the catalytic site (motifs IV-VIII) exhibits strong divergence at the sequence and structure level (i.e. the motifs IV-VIII are different in each subfamily of RFMs). Most of FR servers (both in CAFASP and in our metaserver) correctly identified the N-terminal region of T0172 (aa 1-109, motifs X & I-IV), but differed greatly with respect to the alignment of the remaining 190 aa. This uncertain region was much longer than the typical substrate-binding part of the RFM catalytic domain and we suspected that T0172 contains an additional domain fused with the common core. The pattern of predicted secondary structures ($\alpha\beta\alpha\beta\alpha\beta\alpha\beta-\alpha\alpha\alpha\alpha\alpha-\alpha\beta\alpha\beta\alpha\beta$) confirmed this conjecture–an atypical series of α-helices set apart the repetitive $\alpha\beta$ units typical for the RFM fold. We resubmitted the C-terminal region of T0172 as a separate FR query, which, unsurprisingly, reported a match to the catalytic part of the RFM fold. We have also submitted the helical domain and the fused N- and C-termini as additional queries. FR servers reported a perfect match between the fused termini and various RFM domains. Interestingly, the top-scoring template both for the separated termini and the "fusion" variant was not an experimentally-solved structure, but our own homology model of RNA:m^5C methyltransferase Sun (1j4f in PDB). On the other hand, according to FR servers, the helical insertion exhibited no significant similarity to known structures. Accordingly, group 242 submitted the fully automated FR model built by Pmodeller based exclusively on the 1j4f template.

Bujnicki-Janusz and *GeneSilico* continued with "FRankensteinization" of the T0172 model according to the protocol shown in Figure 1. The hybrid model was constructed by merging best-scoring N- and C-terminal fragments of the crude FR models. It was superimposed onto the composite template, including 1j4f as well as all experimentally solved RFM structures, providing the initial alignment for comparative modeling. The comparative model scored poorly in regions corresponding to motif III and VII-VIII. We generated 20 alternative target-template alignments by shifting the "unfit" sequence in either direction. Two important criteria guiding this refinement step were: 1) the match between the secondary structure predicted for T0172 and observed in the template structures and 2) the match between certain hydrophobic amino acids commonly observed in RFM structures, which are located in the crossover regions and probably stabilize the folding of the key supersecondary structural elements. The same strategy was previously used by one of us to successfully predict the structural details of another RFM superfamily member.[31] For the final modeling of the catalytic domain of T0172 (i.e. domain T0172_1) *Bujnicki-Janusz* used MODELLER with additional constraints on predicted secondary structure. This allowed to model a small α-helix in the motif VIII loop, which was absent from the template structures. After scoring of all models, the non-experts (*GeneSilico*) selected the backbone of the expert's model and rebuilt all sidechains with SCWRL. This rebuilt model exhibited marginally better VERIFY3D score and was submitted by *GeneSilico* according to the aim of elimination of as much of the experts' subjectivity as possible. As for the helical domain, both *Bujnicki-Janusz* and *GeneSilico* decided to select one of the models generated *ab initio* by ROBETTA and available from the CAFASP server. However, *Bujnicki-Janusz* selected the model of the helical domain based on his intuition, while *GeneSilico* selected the model, which scored best according to VERIFY3D. *GeneSilico* submitted combined domains as model_1, while *Bujnicki-Janusz* submitted the catalytic domain as model_1 and the same domain_1 recombined with domain 2 to yield a full-length protein as model_2.

Figure 2 shows model_1 of T0172_1 submitted by *Bujnicki-Janusz*, which almost perfectly aligns with the experimental solution. Only our two groups (020 and 517) managed to accurately predict all major secondary structure elements in this domain with no sequence shifts. However, our models of the helical domain (T0172_2) turned out to be incorrect. It is noteworthy that the VERIFY3D criterion used by *GeneSilico* allowed to identify a model with a somewhat better GDT_TS score than the model intuitively selected by *Bujnicki-Janusz*. At the time of the writing the results of detailed assessment of side chain predictions were not available and we could not compare the performance of *Bujnicki-Janusz*'s expertise in side chain modeling vs *GeneSilico*'s conventional use of SCWRL. According to the GDT_TS analysis, the "monster" built by *Bujnicki-Janusz* was the best T0172 model in CASP5 (model_2: 1st score 46.59, model_1 3rd score 44.45), while the model submitted by *GeneSilico* was the second

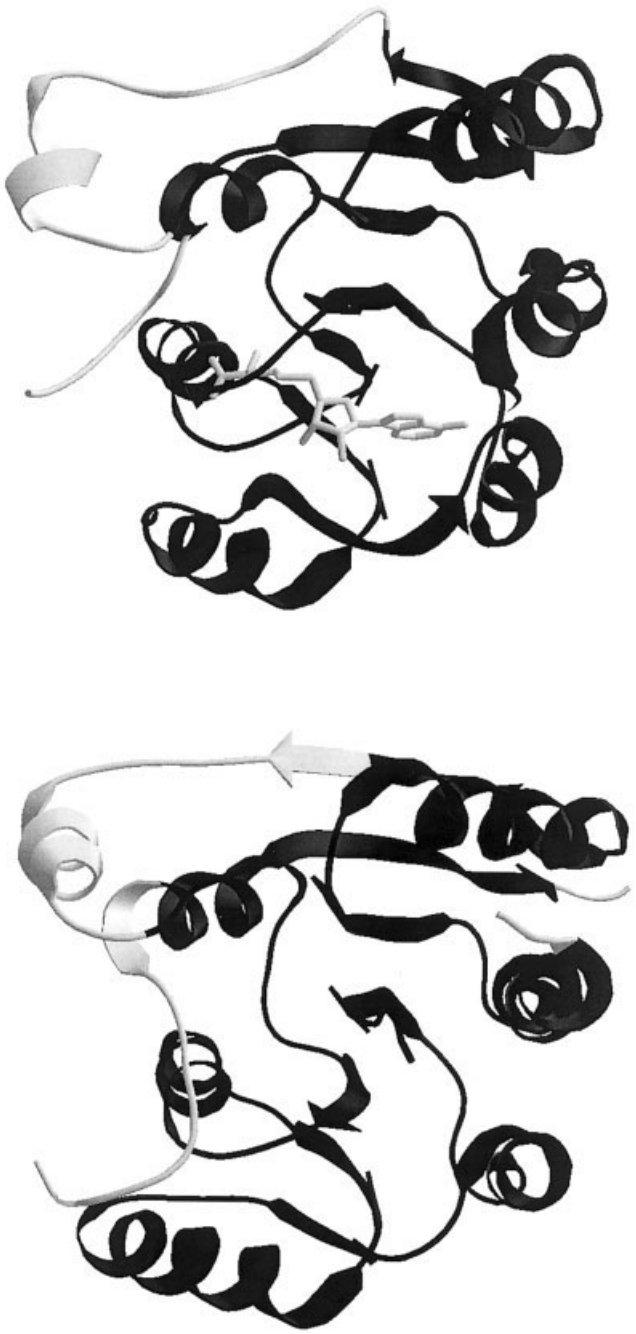

Fig. 2. Cartoon diagram of the T0172 (domain 1) structure solved by X-ray crystallography (top) compared to the blind model submitted by *Bujnicki-Janusz* (bottom; GDT_TS score 44.45). The cofactor molecule is shown in wireframe representation to indicate the most conserved part of the RFM fold. Superimposable regions that generate correct alignment (i.e. corresponding to accurate parts of the model) are colored in black. Regions that are non-superimposable (RMS $>$ 4 Å) or produce incorrect alignment, are shown in gray.

best (GDT_TS score of 46.50). This result also underscores the fact that adding a completely wrong insertion to make the model "complete" (in terms of length), and hence generation of partially misfolded structure (model_2 of *Bujnicki-Janusz*) can significantly **improve** the score over the submission of an almost-perfect (but incomplete) model (model_1 of *Bujnicki-Janusz*). Hence, it is probably possible to "cheat" to achieve better ranking in CASP by extending correct, but incomplete models with bizarre extensions that even could be deliberately misfolded.

Target T0183

T0183, a hypothetical protein TM1559 from *T. maritima*, was an easy CM target, with one challenging element: according to the FR analysis, it displayed obvious similarities to TIM-barrel Class I aldolases, but SS prediction revealed that its extended N-terminus is likely to form an α-helix, not observed in any of the template structures. Among the pre-computed CAFASP models that scored reasonably well according to VERIFY3D, only the models built by ROBETTA contained the N-terminal helix, apparently added "*ab initio*" to the homology-modeled core. Consequently, the *GeneSilico-servers-only* group selected the best-scoring ROBETTA's model for submission as the potentially best fully-automated prediction. The "Frankenstein" modeling of the conserved core of the TIM-barrel was trivial due to the relative homogeneity of intermediate models generated in the 1st step of our protocol. Hence, only a few alternative alignments were tested in the 5th step. The final "FRankenstein's monster" for the target T0183 was built from the core of the best-scoring intermediate model (based mainly on the 1jcj structure) combined with the N-terminus taken from the ROBETTA's model and with the loops that were closest to the template structure (*Bujnicki-Janusz*) or scored best according to VERIFY3D (*GeneSilico*). Regardless of the significant differences in conformations of the loops, both "monsters" turned out to be the best models for the target T0183 in CASP5 (with the GDT_TS scores of 81.38 for *Bujnicki-Janusz* and 80.16 for *GeneSilico*), simply because only these models included the N-terminal helical extension, which was more or less correctly predicted by ROBETTA (Fig. 3). The ROBETTA model itself, on the other hand, was misfolded in the C-terminus and was ranked only as the 22nd (among 1st models) with the GDT_TS score of 76.11. This case proves that the "FRankenstein" protocol can outperform the traditional comparative modeling protocol by including elements built *ab initio*. It also demonstrates that even a partly misfolded model can provide excellent parts to build a superior model. What we did not have during CASP5, was the possibility to run ROSETTA or another *ab initio* method to extend the termini or fill-in missing (or evidently misfolded) internal parts of some of our other models. However, we had to rely only on pre-computed *ab initio* models available from the CAFASP website.

Target T0130

The nucleotidyltransferase (NTase) superfamily[32] is characterized by a compact fold with three universally conserved carboxylate residues and unrelated structural decorations observed in different subfamilies. PSI-BLAST searches revealed that T0130 is homologous to NTases, but failed to indicate the statistical similarity of its

Fig. 3. Cartoon diagram of the T0183 structure solved by X-ray crystallography (top) compared to the blind model submitted by *Bujnicki-Janusz* (bottom; GDT_TS score 81.38). Superimposable regions that generate correct alignment (i.e. corresponding to accurate parts of the model) are colored in black. Regions that are non-superimposable (RMS $>$ 4 Å) or produce incorrect alignment, are shown in gray.

sequence to any NTase of known structure. Nevertheless, all FR servers reported the mammalian poly(A) polymerases (1fa0 and 1f5A), NTase superfamily members, as the best-scoring templates. Other NTase structures were reported with substantially lower scores, suggesting that T0130 is most closely related to PAPs. This notion was corroborated by our analysis of VERIFY3D scores for the preliminary models. Remarkably, we found that only a small portion of the T0130 sequence was aligned in the same way by different servers. The FR alignments for the N-terminal helix and for the C-terminal half of this small protein (114 aa) varied greatly. Moreover, only two catalytic residues were reproducibly identified by fully-automated methods. We found that one of the carboxylates (D79) conserved among T0130 and its closest homologs (found in the unfinished genomes rather than in the nr database) was aligned with the catalytic carboxylate only in model_1 of ROBETTA available from the CAFASP website, but not in the FR models. Consequently, the ROBETTA model was selected for submission by *GeneSilico-servers-only* (group 242).

Starting from the initial FRankenstein's monster-based consensus (including ROBETTA-like position of the third carboxylate), *Bujnicki-Janusz* and *GeneSilico* generated (independently) over 40 alternative models by progressively shifting the target-template alignments in the uncertain regions (in one region at a time). The best-scoring variants of the N-terminal and C-terminal segments were combined with the unambiguously modeled core and merged to produce the final model. The model submitted by *GeneSilico* was very similar to that submitted by *Bujnicki-Janusz*, although it was built independently. Figure 4 shows that the refinement protocol allowed us to build a model of T0130 with all but two secondary structure elements perfectly aligned to the corresponding elements in the experimental solution (including regions with practically no sequence similarity between the target and the template). However, we have completely failed to predict a β-strand observed in the C-terminus of the native structure. This β-strand is present in the structure of the kanamycin NTase (1kny), used as a relatively minor component of the composite template. However, this β-strand is absent from the PAP structure, and PAP was reported as the unequivocally best template by all FR servers. Moreover, the C-terminal β-strand was not found in any of the preliminary FR models, even those based on 1kny. We were also misled by the fact that most of SS prediction methods suggested that the C-terminal region of T0130 assumes helical, not extended conformation. *A posteriori* analysis suggested that if we included the C-terminal strand from 1kny as the major component of the composite template, we could generate models with locally better VERIFY3D score. Hence, from this failure we learned that if non-identical secondary structures are observed in the template superfamily, one should explore the model space based on all possible scaffolds, i.e. build the "FRankenstein's monster" not only from fragments generated by FR methods, but to use all templates to build additional fragments. Nonetheless, we would like to underscore that thanks to the correctly aligned residues over 75% of the sequence length, even without one of the secondary structure elements, our submissions were among the best models built for T0130 in CASP5 (according to the GDT_TS analysis, the model submitted by *Bujnicki-Janusz* was ranked as the 1st).

Target T0170

T0170 (FF domain of HYPA/FBP11) was classified as a FR/NF target, however we have modeled it using essentially the same protocol as all CM targets. The experimentally solved structure shares only two of its four helices

Fig. 4. Cartoon diagram of the T0130 structure solved by X-ray crystallography (top) compared to the blind model submitted by *Bujnicki-Janusz* (bottom; GDT_TS score 59.25). Superimposable regions that generate correct alignment (i.e. corresponding to accurate parts of the model) are colored in black. Regions that are non-superimposable (RMS $>$ 4 Å) or produce incorrect alignment, are shown in gray.

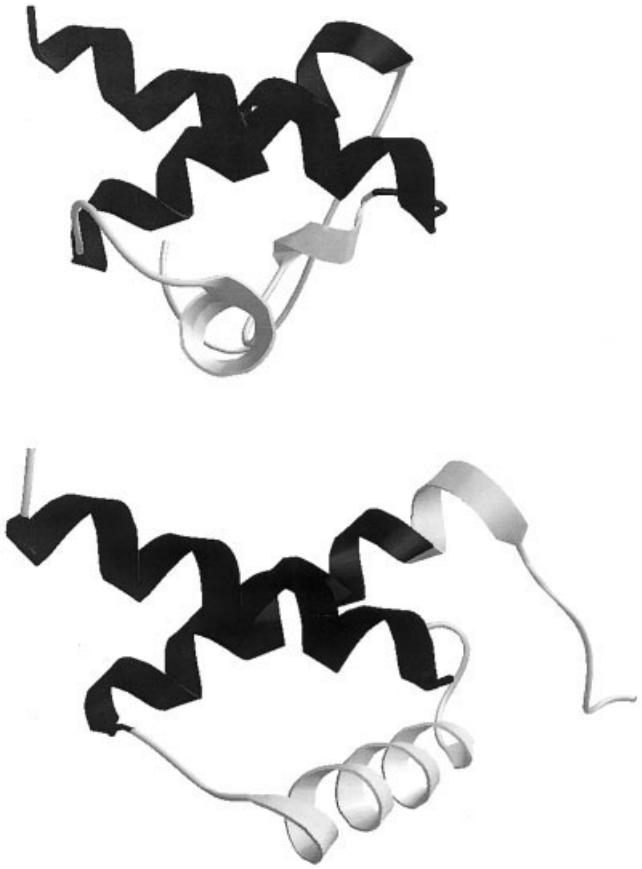

Fig. 5. Cartoon diagram of the T0170 structure solved by NMR (top; GDT-TS score 53.26) compared to the blind model submitted by *GeneSilico* (bottom). Superimposable regions that generate correct alignment (i.e. corresponding to accurate parts of the model) are colored in black. Regions that are non-superimposable (RMS $>$ 4 Å) or produce incorrect alignment, are shown in gray.

with POU homeodomains (1au7, 1akh) formed by three helices. The homeodomain structures were reported by FR servers, albeit with low scores. Nonetheless, only these templates allowed us to generate preliminary models with a reasonably folded, globular core (Fig. 5), which scored quite well according to VERIFY3D. One of such models (Pmodeller model_2) was submitted by *GeneSilico-servers-only* (group 242). As with the other targets, the "FRankenstein's monster" was built from the best-scoring fragments, generated after extensive testing of the alternative target-template alignments. According to the automatic evaluation (GDT_TS score 53.26), the "monster" generated by *GeneSilico* turned out to be the best "model_1" for T0170 in CASP5 (and 5th if all models are considered). The model built by *Bujnicki-Janusz* was scored as 10th among the 1st models (GDT_TS score 44.92), slightly worse than the model selected by *GeneSilico-servers-only* (GDT-TS score 46.38). The difference between all three models was in the backbone coordinates, because *GeneSilico* used the best-scoring template 1au7, while *Bujnicki-Janusz* used both 1au7 and 1akh.

Both *GeneSilico* and *Bujnicki-Janusz* failed to predict the two internal helices of T0170, because there was no corresponding structure present in the templates used for modeling. The structure fragment, which was homology-modeled using the middle α-helix of 1au7 and 1akh, superimposed poorly onto the two small helices observed in the experimentally solved structure. We were aware of the potentially poor quality of our models in this region at the time of their submission, because VERIFY3D scores for the middle part of T0170 remained low despite exten-

sive "FRankensteinization", while the terminal parts (helices) scored quite well in the final models. It is noteworthy that among all T0170 models (1-5) submitted to CASP5, ROBETTA's model_4 turns out to be definitely superior to our CM models - it correctly predicts all helices with correct orientation and sequence alignment (GDT_TS score 64.85). As mentioned above, we believe that one of the major shortcomings of our protocol that needs to be addressed in the future, was the lack of genuine *ab initio/de novo* folding method, which could help us explore the fold space beyond that restricted by the template structures. It could be informative to test whether running folding simulations using ROSETTA[33] (or other *ab initio* methods) with our "FRankenstein's monsters" as starting models (in analogy to Pcons2 models used by the automated ROBETTA server in CASP5), with constraints on well-scoring regions and more freedom on poorly-scoring regions, would result in a systematical improvement of their quality. In our experience, the cases of partially confident models, which contain parts yearning to be refined by *ab initio* methods are quite common in the research practice of molecular biologists that plan the experimental work based on predicted protein structures. We believe that our protocol, which combines elements of CM and FR, has a great potential to become a useful tool for generation of high quality homology models and that it can be further improved (to generate even better models) by adding *ab initio*–like refinement of less confident regions.

CONCLUSIONS

We developed a novel approach for protein structure prediction, which combines various aspects of traditional homology modeling, fold recognition and *ab initio* protein structure prediction. According to the official CASP5 evaluation, this approach turned out to be very successful in the CM category (where *Bujnicki-Janusz* and *GeneSilico* ranked among the few "winning" groups) even though we expected it to be more successful in the FR category (where we were outperformed by other groups using novel consensus FR predictors). The novelty of the "FRankenstein" method is in the focus on "vertical" recombination of structure fragments, typical for the *ab initio* field, rather than "horizontal" sequence alignments, which are typical for comparative modeling. Since we were outperformed by consensus predictors in the FR category, we plan to add the consensus-based ranking to our scoring system for template selection and weighting. Thus, we regard CASP5 primarily as a learning experience, from which we determined the strengths and weaknesses of our modeling protocol and as a source of inspirations for new components to be included in our strategy. It is probably noteworthy that all steps of our protocol beyond fold recognition and generation of the initial set of models were manual. The hand intervention in exploring the alignment space in uncertain regions, as well as superposition and merging of models were very successful in improving the quality of CM and FR predictions, but also extremely labor-intensive (over 24 non-expert-hours/model). Nonetheless, most of the steps leading to generation of the "FRankenstein's model" are automatable. Our goal for the CASP6 experiment will be to build a fully automated prediction pipeline based on the protocol described in this article and to test its performance in CAFASP4 versus the combined strength of human experts and computers.

ACKNOWLEDGMENTS

We thank Gordana Maravić for critical reading of the manuscript and two anonymous reviewers for useful suggestions. J.M.B is an EMBO/HHMI Young Investigator and a fellow of Foundation for Polish Science. I.A.C., M.F., M.A.K, and J.A.S. were supported by the Polish State Committee for Scientific Research (I.A.C., M.F., and J.A.S. by grant 6P04A01124, M.A.K. by grant 3P05A02024).

REFERENCES

1. Venclovas Zemla A, Fidelis K, Moult J. Comparison of performance in successive CASP experiments. Proteins 2001;45 Suppl 5:163 -70 2001;45 Suppl 5:163–170.
2. Fischer D, Elofsson A, Rychlewski L, Pazos F, Valencia A, Rost B, Ortiz AR, Dunbrack RL, Jr. CAFASP2: The second critical assessment of fully automated structure prediction methods. Proteins 2001;45 Suppl 5:171–183.
3. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. LiveBench-2: Large-scale automated evaluation of protein structure prediction servers. Proteins 2001;45:184–191.
4. Tramontano A, Leplae R, Morea V. Analysis and assessment of comparative modeling predictions in CASP4. Proteins 2001;45 Suppl 5:22–38.
5. Sippl MJ, Lackner P, Domingues FS, Prlic A, Malik R, Andreeva A, Wiederstein M. Assessment of the CASP4 fold recognition category. Proteins 2001;45 Suppl 5:55–67.
6. Karplus K, Karchin R, Barrett C, Tu S, Cline M, Diekhans M, Grate L, Casper J, Hughey R. What is the value added by human intervention in protein structure prediction? Proteins 2001;45 Suppl 5:86–91.
7. Sayle RA, Milner-White EJ. RASMOL: biomolecular graphics for all. Trends Biochem Sci 1995;20:374–376.
8. LoConte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 2000;28:257–259.
9. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. Structure prediction Meta Server. Bioinformatics 2001;17:750–751.
10. Douguet D, Labesse G. Easier threading through web-based comparisons and cross-validations. Bioinformatics 2001;17:752–753.
11. Rost B. Predicting one-dimensional protein structure by profile based neural networks. Methods Enzymol 1996;266:525–539.
12. Kurowski MA, Bujnicki JM. GeneSilico protein structure prediction meta-server. Nucleic Acids Res 2003;31:3305–3307.
13. Kelley LA, McCallum CM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:501–522.
14. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pac Symp Biocomput 2000;119–130.
15. 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.
16. 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.
17. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
18. Xu J, Li M, Lin G, Kim D, Xu Y. Protein structure prediction by linear programming. Pac Symp Biocomput 2003;in press
19. 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.
20. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;25:4876–4882.
21. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
22. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature 1992;356:83–85.
23. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723.
24. Scott WRP, Hunenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Kruger P, van Gunsteren WF. The GROMOS biomolecular simulation program package. J Phys Chem 1999;103:3596–3607.
25. Dunbrack RL. Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL. Proteins 1999;Suppl 3:81–87.
26. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502–511.
27. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
28. Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 2002;47:228–235.
29. Ouali M, King RD. Cascaded multiple classifiers for secondary structure prediction. Protein Sci 2000;9:1162–1176.
30. Bujnicki JM. Comparison of protein structures reveals monophyletic origin of the AdoMet-dependent methyltransferase family and mechanistic convergence rather than recent differentiation of N4-cytosine and N6-adenine DNA methylation. In Silico Biol 1999;1:1–8, http://www.bioinfo.de/isb/1999-01/0016/.
31. Bujnicki JM. *In silico* analysis of the tRNA:m^{1}A58 methyltransferase family: homology-based fold prediction and identification of new members from Eubacteria and Archaea. FEBS Lett 2001;507: 123–127.
32. Aravind L, Koonin EV. DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history. Nucleic Acids Res 1999;27:1609–1618.
33. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE, Baker D. Rosetta in CASP4: Progress in ab initio protein structure prediction. Proteins 2001;45 Suppl 5:119–126.