BIOINFORMATICS DISCOVERY NOTE

Vol. 18 no. 10 2002
Pages 1391–1395

Fold-recognition detects an error in the Protein Data Bank

Janusz Bujnicki[1], Leszek Rychlewski[2] and Daniel Fischer[3,*]

[1]Bioinformatics Laboratory, International Institute of Molecular and Cell Biology, ks. Trojdena, 4 02-109, Warsaw, [2]BioInfoBank Institute, ul. Limanowskiego 24A, 60-744, Poznan, Poland and [3]Bioinformatics, Department of Computer Science, Ben Gurion University, Beer-Sheva, 84015, Israel

Received on January 9, 2002; revised on March 18, 2002; accepted on March 19, 2002

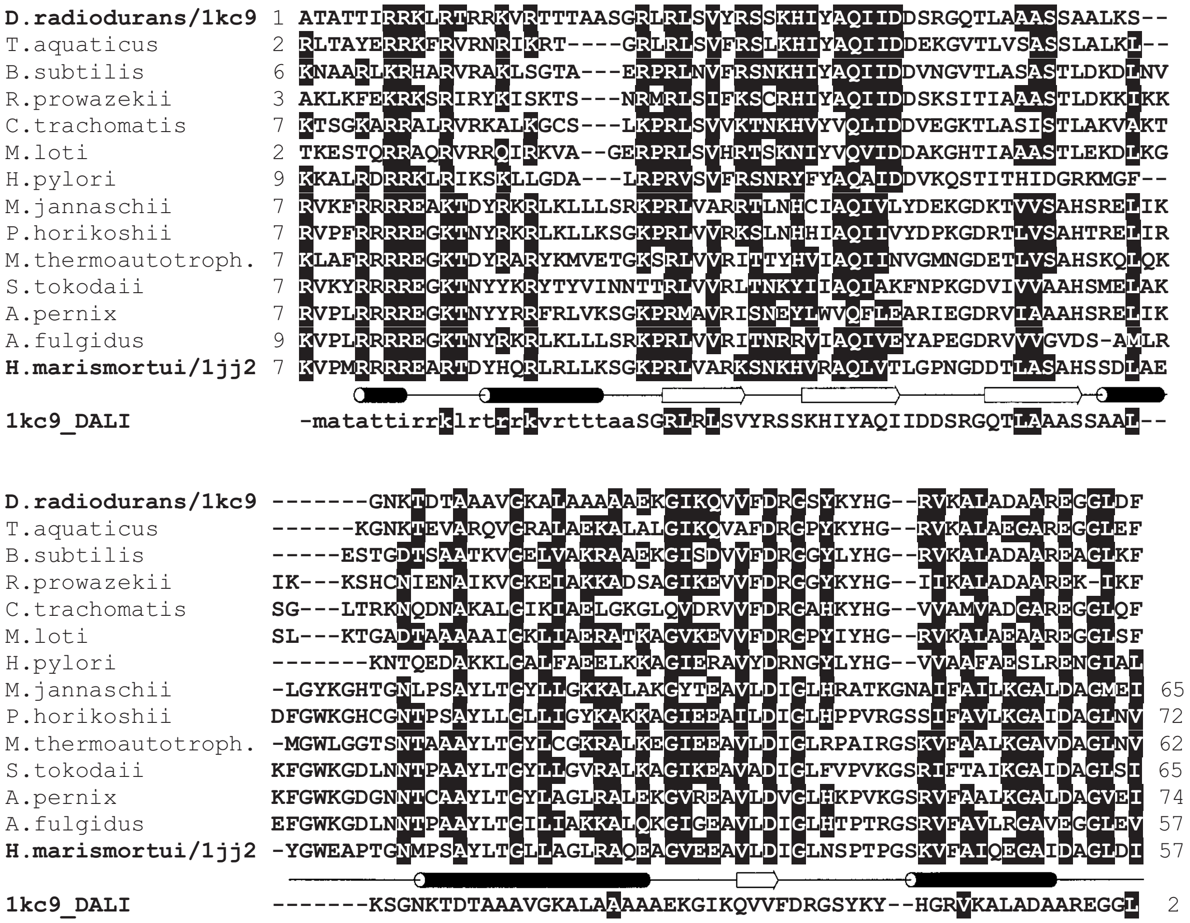
Determining the 3D-structure of a protein by means of X-ray crystallography, requires the fitting of the amino acid sequence into an electron-density map. This can be a difficult, time-consuming task if the map is of low resolution, showing little data for the side-chains. Here we show how fold recognition allowed us to detect an error in a recently deposited PDB entry, and propose a method that can be of aid in the fitting of a sequence into a low-resolution electron-density map. The described procedure is the result of an interesting anomaly detected during the protein structure prediction benchmarking experiment, LiveBench (Bujnicki *et al.*, 2001a).

Protein structure prediction is aimed at generating approximate 3D structure models for target protein sequences of unknown structure. In particular, fold recognition or threading is aimed at those targets sharing little or no significant sequence similarity to any protein of known structure (Fischer *et al.*, 1996). The LiveBench continuous benchmarking program assesses the performance of automatic fold-recognition servers by submitting as prediction targets the sequences of newly released PDB entries with no clear sequence similarity to previously released proteins. After collecting the predicted models from the participating servers, the predictions are assessed by comparing them to the experimental structures.

Among the target sequences recently considered, were a number of chains of the newly released PDB entry 1kc9, which describes the crystal structure of the large ribosomal subunit from the bacterium *D. radiodurans* (Harms, 2001). This is a C_α-only structure with a resolution of 3.1 Å. Here we focus on the M chain, which corresponds to the structure of residues 2-114 of the ribosomal protein L18, herein referred to as 1kc9_M. All the fold-recognition methods (Bujnicki *et al.*, 2001b) participating in LiveBench (including iterated PSI-BLAST searches) suggested with very high confidence that a previously determined structure is compatible with that of 1kc9_M (see the corresponding LiveBench web-page at http://BioInfo.PL/Meta/target.pl?V=4&T=0&id=5058). All these hits corresponded to the N-terminal region of the structure of the ribosomal protein L18 from the archaeon *H. marismortui* (PDB codes 1jj2_M, a full-atom entry, and 1ffk_K, a C_α-only entry; Ban *et al.* 2000), the sequence of which is 32% identical to 1kc9_M. Although this level of similarity is often considered close to the so-called 'twilight-zone' (a PSI-BLAST search on NCBI using default parameters hit 1ffk in the first round with a not very significant e-value of 0.04, but subsequent rounds assigned to 1ffk very significant e-values—10^{-8} in the second round), the fold-recognition results strongly indicated that these two proteins have diverged from a common ancestor, and that the 3D structure of 1kc9_M is likely to be very similar to that of *H. marismortui*. For example, the fold-recognition servers ORFeus (Pas *et al.*, 2002) and INBGU (http://www.cs.bgu.ac.il/~bioinbgu; Fischer 2000) identified 1jj2 at rank one with scores of 18.9 and 115.6, respectively, which are significantly higher than the suggested confidence thresholds above which it is extremely unlikely to obtain a wrong prediction (>7.0 for ORFeus and >30 for INBGU). Furthermore, the fold-recognition meta-predictors PCONS (Lundstrom *et al.*, 2001) and 3DS3 ('Shotgun on 3', D. Fischer, unpublished), available via the meta server at http://bioinfo.pl/meta, also identified 1jj2 at rank one with scores of 4.6 and 100, respectively, well above their suggested confidence thresholds (>2.0 for PCONS and >30 for 3DS3). Consequently, 3D models of *D. radiodurans*'s ribosomal protein L18 based on the *H. marismortui* structure as parent are expected to be correct, if the sequence-to-structure alignment detected by fold recognition is accurate (Figure 1).

The quality of the fold-recognition 3D models was assessed using the experimental 3D structure of 1kc9 (see the above url for a summary of the fold-recognition models and their assessment). The overall RMSD of the

*To whom correspondence should be addressed.
E-mail: dfischer@cs.bu.ac.il



```
D.radiodurans/1kc9 1 ATATTIRRKLRTRRKVRTTTAASGRLRLSVYRSSKHIYAQIIDDSRGQTLAAASSAALKS--
T.aquaticus        2 RLTAYERRKFRVRNRIKRT----GRLRLSVFRSLKHIYAQIIDDEKGVTLVSASSLALKL--
B.subtilis         6 KNAARLKRHARVRAKLSGTA---ERPRLNVFRSNKHIYAQIIDDVNGVTLASASTLDKDLNV
R.prowazekii       3 AKLKFEKRKSRIRYKISKTS---NRMRLSIFKSCRHIYAQIIDDSKSITIAAASTLDKKIKK
C.trachomatis      7 KTSGKARRALRVRKALKGCS---LKPRLSVVKTNKHVYVQLIDDVEGKTLASISTLAKVAKT
M.loti             2 TKESTQRRAQRVRRQIRKVA--GERPRLSVHRTSKNIYVQVIDDAKGHTIAAASTLEKDLKG
H.pylori           9 KKALRDRRKLRIKSKLLGDA---LRPRVSVFRSNRYFYAQAIDDVKQSTITHIDGRKMGF--
M.jannaschii       7 RVKFRRRREAKTDYRKRLKLLLSRKPRLVARRTLNHCIAQIVLYDEKGDKTVVSAHSRELIK
P.horikoshii       7 RVPFRRRREGKTNYRKRLKLLKSGKPRLVVRKSLNHHIAQIIVYDPKGDRTLVSAHTRELIR
M.thermoautotroph. 7 KLAFRRRREGKTDYRARYKMVETGKSRLVVRITTYHVIAQIINVGMNGDETLVSAHSKQLQK
S.tokodaii         7 RVKYRRRREGKTNYYKRYTYVINNTTRLVVRLTNKYIIAQIAKFNPKGDVIVVAAHSMELAK
A.pernix           7 RVPLRRRREGKTNYYRRFRLVKSGKPRMAVRISNEYLWVQFLEARIEGDRVIAAAHSRELIK
A.fulgidus         9 KVPLRRRREGKTNYRKRLKLLLSRKPRLVVRITNRRVIAQIVEYAPEGDRVVVGVDS-AMLR
H.marismortui/1jj2 7 KVPMRRRREARTDYHQRLRLLKSGKPRLVARKSNKHVRAQLVTLGPNGDDTLASAHSSDLAE

1kc9_DALI            -matattirrklrtrrkvrtttaaSGRLRLSVYRSSKHIYAQIIDDSRGQTLAAASSAAL--


D.radiodurans/1kc9 -------GNKTDTAAAVGKALAAAAAEKGIKQVVFDRGSYKYHG--RVKALADAAREGGLDF
T.aquaticus        ------KGNKTEVARQVGRALAEKALALGIKQVAFDRGPYKYHG--RVKALAEGAREGGLEF
B.subtilis         -----ESTGDTSAATKVGELVAKRAAEKGISDVVFDRGGYLYHG--RVKALADAAREAGLKF
R.prowazekii       IK---KSHCNIENAIKVGKEIAKKADSAGIKEVVFDRGGYKYHG--IIKALADAAREK-IKF
C.trachomatis      SG---LTRKNQDNAKALGIKIAELGKGLQVDRVVFDRGAHKYHG--VVAMVADGAREGGLQF
M.loti             SL---KTGADTAAAAAIGKLIAERATKAGVKEVVFDRGPYIYHG--RVKALAEAAREGGLSF
H.pylori           -------KNTQEDAKKLGALFAEELKKAGIERAVYDRNGYLYHG--VVAAFAESLRENGIAL
M.jannaschii       -LGYKGHTGNLPSAYLTGYLLGKKALAKGYTEAVLDIGLHRATKGNAIFAILKGALDAGMEI 65
P.horikoshii       DFGWKGHCGNTPSAYLLGLLIGYKAKKAGIEEAILDIGLHPPVRGSSIFAVLKGAIDAGLNV 72
M.thermoautotroph. -MGWLGGTSNTAAAYLTGYLCGKRALKEGIEEAVLDIGLRPAIRGSKVFAALKGAVDAGLNV 62
S.tokodaii         KFGWKGDLNNTPAAYLTGYLLGVRALKAGIKEAVADIGLFVPVKGSRIFTAIKGAIDAGLSI 65
A.pernix           KFGWKGDGNNTCAAYLTGYLAGLRALEKGVREAVLDVGLHKPVKGSRVFAALKGALDAGVEI 74
A.fulgidus         EFGWKGDLNNTPAAYLTGILIAKKALQKGIGEAVLDIGLHTPTRGSRVFAVLRGAVEGGLEV 57
H.marismortui/1jj2 -YGWEAPTGNMPSAYLTGLLAGLRAQEAGVEEAVLDIGLNSPTPGSKVFAIQEGAIDAGLDI 57

1kc9_DALI          -------KSGNKTDTAAAVGKALAAAAAEKGIKQVVFDRGSYKY--HGRVKALADAAREGGL 2
```

Fig. 1. The multiple sequence alignment of ribosomal L18 proteins shows extensive conservation (1kc9_M homologs/ Bacteria—top 7, 1jj2_M homologs/Archaea—bottom 7), while the structure alignment of 1kc9_M and 1jj2_M, having a 2-residue shift, shows little conservation. Conserved residues are highlighted. The number of terminal residues omitted for clarity is indicated for each sequence. The secondary structure of 1jj2_M is shown as tubes and arrows. The structural alignment of 1kc9_M and 1j22_M reported by DALI is shown at the bottom, with unambiguously superimposed C_α atoms indicated in uppercase in the 1kc9_DALI sequence. The alignment obtained by the fold-recognition servers (see Bujnicki *et al.* 2001b, and references therein and also http://BioInfo.PL/Meta/target.pl?V=4&T=0&id=5058) largely corresponds to the multiple sequence alignment, while the structural alignment (1kc9_DALI) has a global shift of 2 residues.

models and the native structure was relatively high: 6.6 Å over 109 residues, indicating that the predicted models significantly differ from the native structure. In addition to measuring differences using RMSD, LiveBench uses other quality assessment tools for evaluation. One of these tools is MAXSUB (Siew *et al.*, 2000), a sequence-dependent assessment measure, that assigns scores to a model in the range 0.0 to 1.0. A MAXSUB score of 0.0 indicates that the model has few residues that can be superimposed to the native structure, while a positive score indicates that some significant structural similarity exists. A MAXSUB score of 1.0 indicates a perfect match. For reference, models with MAXSUB scores above 0.2 correspond to predictions considered to be 'correct' in previous CASP experiments (Siew *et al.*, 2000; CASP3, 1999; CASP4, 1999). For the *D. radiodurans* predicted fold-recognition models, MAXSUB assigned a score of 0.0, with only 19 'well-predicted' residues, suggesting that the quality of the models was very poor. Given the extremely high confidence scores obtained by the servers, we were very surprised that the predicted models were of such low quality.

Structural comparison by DALI (Holm and Sander, 1993) confirmed that 1jj2_M and 1ffk_K are highly similar to 1kc9_M (106 superimposable C_α atoms with an RMSD of 3.1 Å; see the above URL for details). A C_α model for 1kc9_M built using 1ffk_K as parent and DALI's alignment as guide, obtains a MAXSUB score of 0.60, with a subset of 86 residues that superimpose onto 1kc9_M with an RMSD of 1.9 Å, indicating that this can be considered to be a high quality fold-recognition model. Thus, despite the fact that our fold-recognition methods correctly identified a compatible structure for 1kc9_M, the sequence-to-structure alignments contained inaccuracies, as judged by comparing the predicted models with the native structure of 1kc9_M. We observed that all the predicted models were based on sequence–structure alignments that generally differed by a 2-residue shift when compared to DALI's structural alignment (Figure 1).

This fact was of great concern, because we have not observed errors of this magnitude for predicted models with such high confidence scores. We reasoned that there could be two possible explanations: (a) the experimental structure of either 1kc9_M or 1jj2_M had a 2-residue shift, or (b) the particular sequence of 1kc9_M 'deceived' all our methods resulting in an overall 2-residue shift in the sequence–structure alignments.

To test hypothesis (a) we have attempted to assess the quality of 1kc9_M using Eisenberg *et al.*'s 3D-structure evaluation program VERIFY3D (Eisenberg *et al.*, 1997) at http://www.doe-mbi.ucla.edu/Services/Verify_3D (other similar programs could have been used as well). VERIFY3D assigns a compatibility score to each residue of a full-atom protein structure. This score measures the compatibility of the model with its sequence. Negative or <0.1 scores are indicative of potential problems. Because 1kc9_M is a C_α-only structure, we first generated a full-atom model for 1kc9_M using Holm & Sander's MAXSPROUT program (Holm and Sander, 1991http://www.ebi.ac.uk/dali/maxsprout). The overall VERIFY3D score (Figure 2) for 1kc9_M was low (0.13), with some regions with negative scores. Subsequently, we generated a second full-atom model for residues 4–112, referred to here as 1kc9_M_shifted, using the coordinates of 1kc9_M (except for the last two), but with a shift of 2 positions in the amino acid identities. The VERIFY3D overall score for this full-atom model was 0.27, significantly better (+0.14) than that of 1kc9_M, indicating that 1kc9_M may contain some problems.

As a control of our procedure and of the MAXSPROUT and VERIFY3D programs, we applied the exact same steps to the C_α-only structure of 1ffk_K. The VERIFY3D overall score for this full-atom model was 0.34, whereas the VERIFY3D score of a model generated from a 2-residue shifted structure of 1ffk_K was only 0.18, a difference of 0.16. This also indicates that the native 1ffk_K appears to be a correct structure. A further control for MAXSPROUT was carried out by removing all non-C_α atoms from the full-atom 1jj2_M and by applying the same steps as above. The MAXSPROUT generated full-atom model for 1jj2_M had an overall VERIFY3D score of 0.33 while the native full-atom structure of 1jj2_M had a score of 0.40. These controls demonstrate that the procedure used above to assess the quality of a C_α model are reliable, and that 1kc9_M is of rather low quality, lower than what could have been expected for a 3.1 Å structure. The difference in quality between 1kc9_M and 1kc9_M_shifted is of the same order as the difference observed between the native 1ffk_K and its shifted version, suggesting that the native 1ffk_K is a correct model, that 1kc9_M may be incorrect and that 1kc9_M_shifted is a better model.

Further support for the correctness of the fold-recognition sequence–structure alignments produced by the servers can be gathered on evolutionary grounds. We collected close homologs (first round PSI-BLAST hits with e-values below 10^{-5}) for 1kc9_M and for 1jj2_M and computed a multiple sequence alignment (shown in Figure 1). The conservation of most positions in the multiple alignment is evident. The first and last rows in the multiple alignment correspond to the sequences of 1kc9_M and 1jj2_M, respectively. Their alignment is essentially identical to the sequence–structure alignment obtained by fold-recognition. This is not surprising, because most modern fold-recognition methods make use of evolutionary information of homologous sequences in the recognition process. Figure 1 shows that the structural alignment of 1kc9_M (denoted as 1kc9_DALI) and 1jj2_M has very few conserved positions, suggesting again that the former may be shifted by 2 residues.

Finally, when we compared the fold-recognition predicted models with 1kc9_M_shifted, an overall RMSD of 4.1 Å was obtained, with a relatively high MAXSUB score of 0.49 and 73 residues superimposable residues. This suggests that if 1kc9_M_shifted is the correct structure, the predicted models are of relatively high accuracy, comparable to that obtained by structural comparison. Encouraged by these findings, we analyzed a number of other chains in the 1kc9 entry where our methods strongly indicated similar potential problems. Using the fold-recognition alignments and other modeling bioinformatics methods, we have constructed theoretical models for the L13, L14, L15, L16, L18, L23, and L24 proteins of the *D. radiodurans* ribosome complex. These full atom models have been deposited into the Protein Data Bank under the entry 1gs2, and will be described elsewhere (Bujnicki *et al.*, 2002). Subsequently, we communicated these findings to the authors of 1kc9, who confirmed that the 1kc9 contained shifts (A. Yonath, personal communication) and promptly released a replacement entry (PDB code: 1kpj). A number of chains in 1kpj had been retraced, suggesting that most shifts of the original entry might have been due to some technical difficulty. All the seven chains for which our methods identified potential problems were revised in 1kpj. Superposition of 1gs2_M onto 1kpj_M resulted in an RMSD of 1.6 Å, indicating that the former is a highly accurate model, and that the retracing of the latter is consistent with our findings

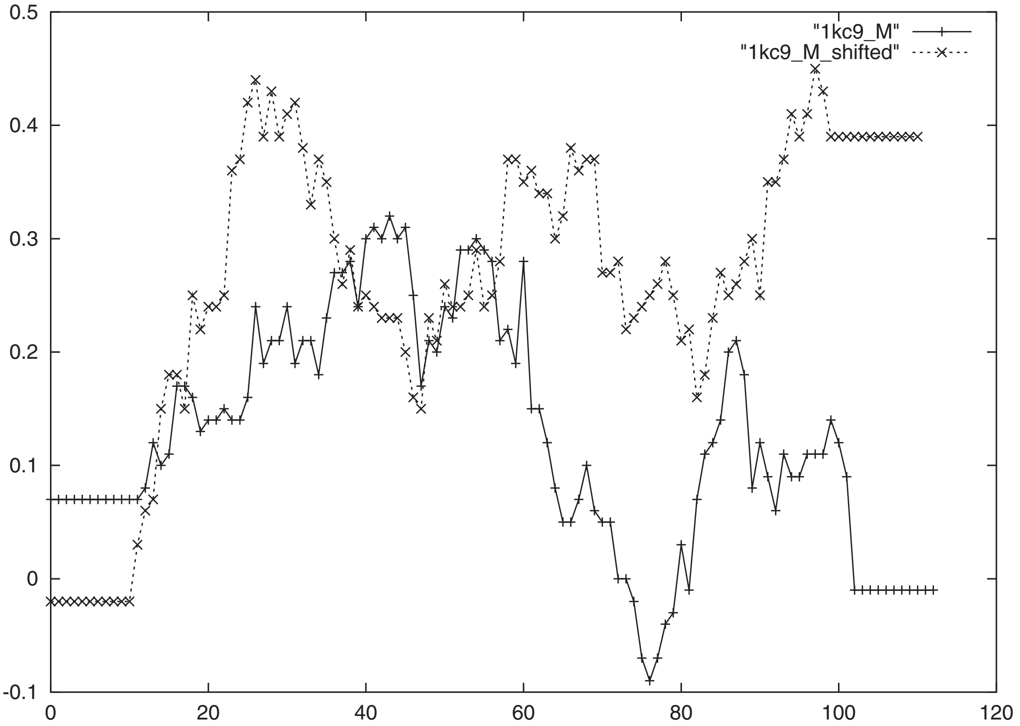


Fig. 2. VERIFY3D plots corroborate that the sequence in 1kc9_M is shifted. The '+' line corresponds to the full-atom model generated from the original entry 1kc9_M and the 'X' line corresponds to our proposed full-atom, shifted structure, 1kc9_M_shifted. The latter has higher VERIFY3D scores at almost all positions, indicative of a higher sequence–structure compatibility. The original entry 1kc9_M shows regions with scores below 0.1, with a dip below zero around residue 78. VERIFY3D scores below 0.1 are indicative of serious problems in the model. Except for the termini, 1kc9_M_shifted shows no potentially wrong regions, suggesting that it represents a better model of the ribosomal protein L18.

above. Similar results were obtained for the L13, L14 and L23 proteins. However, less encouraging results were obtained for the L15, L16 and L24 proteins, suggesting that further revision of the experimental structures may be required. Only a detailed comparison of our predicted models vis-a-vis the experimental structure factors (not yet available) will allow us to determine whether these proteins in 1kpj still contain shifts, or whether our predictions for these proteins are wrong. In any case, we conclude that our methods allowed for the rapid detection of an erroneous PDB entry, that they could have been of help to avoid problems in the original deposition, and may also help to identify other potential errors in the replacement structure and to refine the full-atom experimental ribosome structure.

In summary, a number of bioinformatics tools can be of help during the structure determination process. In particular, for those cases where (i) only a low-resolution electron-density map is available, and (ii) fold-recognition and structural comparison indicate that the yet-undetermined 3D structure is similar to an already solved protein, we propose that the comparison of a structure-based alignment with a sequence-based alignment can be of aid for the better tracing of the sequence into the map. This procedure can be used as a verification step even for those cases where the structure determination applied Molecular Replacement techniques using a related structure. Interestingly, D. Jones has recently suggested that even when no obvious related structures exist, fold-recognition models may be used as Molecular Replacement phasing models (Jones, 2001). A web-server has been set up to facilitate the search for suitable threading models and to enable the comparison of experimental protein structures with fold recognition results (http://BioInfo.PL/Meta/pdb-test.pl). The server allows the user/crystallographer to perform the same analysis as the one carried out by the LiveBench program, which led to the original detection of the error in 1kc9_M.

We believe that the use of fold-recognition and other bioinformatics verification tools, such as those described above, should be applied to every newly determined structure. Fold-recognition models, albeit being only 'in-silico' models, can often provide a rich source of evolutionary, statistical and structural information. Incorporating this independent source of information into the structure determination process can clearly be beneficial, especially if the experimental data is of poor quality. With the growth in structural knowledge expanding as the result of structural genomics projects, fold-recognition models will be valuable for an increasing number of cases.

Since the submission of this manuscript, the replacement entry 1kpj has now been revised, partly in agreement with our analyses. This confirmed our suggestion that further revision of 1kpj may be required.

NOTE ADDED IN PROOF

Since the submission of this manuscript, the replacement entry 1kpj has now been replaced by a third pdb entry (11nr). In 11nr, a number of chains have been revised, partly in agreement with our analyses. This confirmed our suggestion that further revision of 1kpj may be required.

REFERENCES

Ban,N., Nissen,B.N., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.

Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001a) Livebench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci.*, **10**, 352–361.

Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001b) Structure prediction meta server. *Bioinformatics*, **17**, 750–751.

Bujnicki,J.J., Feder,M., Rychlewski,L. and Fischer,D. (2002) Errors in the D. radiodurans large ribosomal subunit structure detected by protein fold-recognition and structure validation tools. *FEBS Letters*. in press

CASP3 (1999) Critical Assessment of Protein Structure Prediction Methods (CASP), Round III. *Proteins*, Suppl. 4, see also http://Predictioncenter.lnl.gov

CASP4 (2001) Critical Assessment of Protein Structure Prediction Methods (CASP), Round IV. *Proteins*, Suppl. 5, see also http://Predictioncenter.lnl.gov

Eisenberg,D., Luthy,R. and Bowie,J.U. (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. *Meth. Enzymol.*, **277**, 396, 404.

Fischer,D., Rice,D.W., Bowie,J.U. and Eisenberg,D. (1996) Assigning amino acid sequences to 3D protein folds. *FASEB J.*, **10**, 126–136.

Fischer,D. (2000) Combining sequence derived properties with evolutionary information. *Proc. Pac. Symp. Biocomput.*, 119–130.

Harms,J.M., Schluenzen,F., Zarivach,A., Bashan,A., Gat,S., Agmon,I., Bartels,H., Franceschi,F and Yonath,A. (2001) High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell*, **107**, 679.

Holm,L. and Sander,C. (1991) Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *J. Mol. Biol.*, **218**, 183–194.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Jones,D.T. (2001) Evaluating the potential of using fold-recognition models for molecular replacement. *Acta Crys.*, **D57**, 1428–1434.

Lundstrom,J., Rychlewski,L., Bujnicki,J. and Elofsson,A. (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci.*, **10**, 2354–2362.

Pas,J., Wyrwicz,L., Bujnicki,J.M. and Rychlewski,L. (2002) ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. Submitted 2002.

Siew,N., Elofsson,A., Rychlewski,L. and Fischer,D. (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **6**, 776–785.