Sequence-structure mapping errors in the PDB: OB-fold domains

ČESLOVAS VENCLOVAS,[1,2] KRZYSZTOF GINALSKI,[3,4] AND CHULHEE KANG[5]

[1]Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94551, USA
[2]Institute of Biotechnology, LT-2028 Vilnius, Lithuania
[3]Department of Biochemistry, University of Texas, Southwestern Medical Center, Dallas, Texas 75390, USA
[4]BioInfoBank Institute, 60–744 Poznań, Poland
[5]School of Molecular Biosciences, Washington State University, Pullman, Washington 99164-4660, USA

(RECEIVED January 15, 2004; FINAL REVISION March 19, 2004; ACCEPTED March 22, 2004)

Abstract

The Protein Data Bank (PDB) is the single most important repository of structural data for proteins and other biologically relevant molecules. Therefore, it is critically important to keep the PDB data, as much as possible, error-free. In this study, we have analyzed PDB crystal structures possessing oligonucleotide/oligosaccharide binding (OB)-fold, one of the highly populated folds, for the presence of sequence-structure mapping errors. Using energy-based structure quality assessment coupled with sequence analyses, we have found that there are at least five OB-structures in the PDB that have regions where sequences have been incorrectly mapped onto the structure. We have demonstrated that the combination of these computation techniques is effective not only in detecting sequence-structure mapping errors, but also in providing guidance to correct them. Namely, we have used results of computational analysis to direct a revision of X-ray data for one of the PDB entries containing a fairly inconspicuous sequence-structure mapping error. The revised structure has been deposited with the PDB. We suggest use of computational energy assessment and sequence analysis techniques to facilitate structure determination when homologs having known structure are available to use as a reference. Such computational analysis may be useful in either guiding the sequence-structure assignment process or verifying the sequence mapping within poorly defined regions.

Keywords: structure quality assessment; sequence register errors; sequence analysis; molecular modeling; X-ray crystallography; SSB protein; PDB

Experimentally determined three-dimensional (3D) structures of proteins are of great importance and of broad interest. The knowledge of protein structures is critical in understanding and/or modifying their molecular function. Structural data are also invaluable in understanding the physical basis of protein folding and stability. Therefore, it is very desirable that 3D structures deposited into the Protein Data Bank (PDB; Berman et al. 2000) are as much as possible free from errors. The PDB structural data are extensively used in a number of derivative databases and in the development of various computational biology approaches, including protein structure prediction methods. Once a flawed protein structure gets into the PDB, it may propagate into other public databases and affect the interpretation of structure-function relationship for the whole protein family.

There are a number of methods developed to date for the detection of "unusual" protein structures that are often indicative of structural flaws within specific regions of the protein chain (e.g., Luthy et al. 1992; Laskowski et al. 1993; Sippl 1993; Hooft et al. 1996). However, in most cases when poor structural quality scores are obtained for a short region, it is difficult to judge whether this is because of

Reprint requests to: Česlovas Venclovas, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, L-448, PO Box 808, Livermore, CA 94551, USA; e-mail: venclovas@llnl.gov; fax: (925) 422-2282.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.04634604.

unusual conformation, residue packing, the lack of normally bound ligand, and so forth, or whether this is due to an error in the structure determination. On the other hand, there is a subset of structural errors that oftentimes not only can be unambiguously identified, but also the recipe for their correction can be provided. These are the sequence-structure mapping or sequence-register shift errors. These particular errors arise from the incorrect assignment of residue side chains for a protein sequence segment within the electron density map. As a result, the position of the backbone for this segment usually is not significantly affected, but side chains are shifted along the backbone and regions adjacent to the segment have insertions/deletions.

In this study we have analyzed oligonucleotide/oligosaccharide binding (OB)-fold (Murzin 1993) domains in the PDB for the presence of sequence-structure mapping errors. We show that there are at least five OB-structures in the PDB that have regions where sequences have been incorrectly mapped onto the structure. We also demonstrate that the structure quality assessment together with sequence analyses may be effective both in detecting and correcting fairly inconspicuous sequence-register shift errors. Guided by the results of computational analysis, we have revisited X-ray data for one of these entries and deposited the updated structure with the PDB.

Results

An OB-fold domain, which consists of five antiparallel β-strands forming a closed or partly opened barrel (Fig. 1; Murzin 1993), is found in a large number of proteins, and currently encompasses nine superfamilies in the SCOP database (Murzin et al. 1995). Initially, we calculated ProsaII Z-scores for all PDB structures of OB-fold domains as defined in SCOP and directly available from the ASTRAL database version 1.63 (Brenner et al. 2000). However, most NMR structures produced poor Z-scores, in accord with a number of studies that found structures solved by NMR spectroscopy normally to be of lower precision than high-resolution X-ray structures (e.g., Abagyan and Totrov 1997; Doreleijers et al. 1998; Ratnaparkhi et al. 1998). Because of the inherent difficulty of using the same scale to compare the quality of NMR and X-ray structures and because only a handful OB-fold structures are solved by NMR, we decided to limit our analysis to only crystal structures of OB-domains, the set containing 842 individual protein chains. The resulting plot of ProsaII Z-score values and their dependency on the chain length is shown in Figure 2.

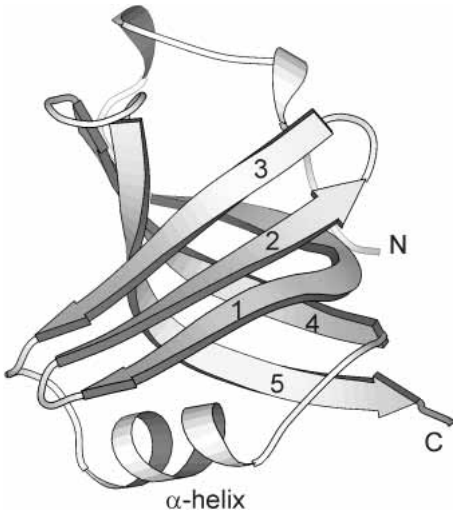

Figure 1. An example of a canonical structure of the OB-fold based on the N-terminal domain of archaeal aspartyl-tRNA synthetase (1b8a; Schmitt et al. 1998). Five β-strands (numbered) form a barrel capped with an α-helix.

To help with the selection of protein chains (points in the plot) for further study, we first performed a simple regression analysis of the ProsaII Z-score distribution as a function of sequence length. However, the structural representation of OB-domains in PDB is highly biased. Some of these domains are represented by multiple PDB entries, each containing several polypeptide chains, whereas others are represented only by a single chain. Therefore, for the regression analysis, we took only those protein chains that are <95% identical in sequence to each other from the corresponding subset of the ASTRAL 1.63 database. The distribution of Z-scores for the resulting nonredundant set containing 106 OB-domains is shown in Figure 2A. The regression line corresponds to the scores expected for the sequence length x. Next, we sorted the points of the complete set (842 chains) according to their residuals (the vertical distance from the regression line) and picked 100 protein chains with the largest residuals for a more detailed analysis. In this set, we specifically focused on potential sequence-structure mapping errors. We reasoned that errors of this kind are usually unambiguous and therefore a straightforward solution to correct them can be offered. As a result, we have identified sequence-structure mapping errors in 12 OB-fold protein chains (Fig. 2B) originating from five PDB entries. All of these errors are described in more detail following.

Escherichia coli *single-stranded DNA binding (SSB) protein (1qvc, 1eqq)*

In the version 1.63 of SCOP and ASTRAL databases, there are three crystal structures of *E. coli* SSB determined both in a free state (1kaw [Raghunathan et al. 1997], 1qvc [Matsumoto et al. 2000]) and bound to ssDNA (1eyg [Raghunathan et al. 2000]). On comparison of these structures with DALI, we have noticed the discrepancy of the sequence-structure mapping between 1qvc and the two other structures. The 1qvc structure has been solved with a significantly higher resolution (2.2 Å) compared with either 1eyg

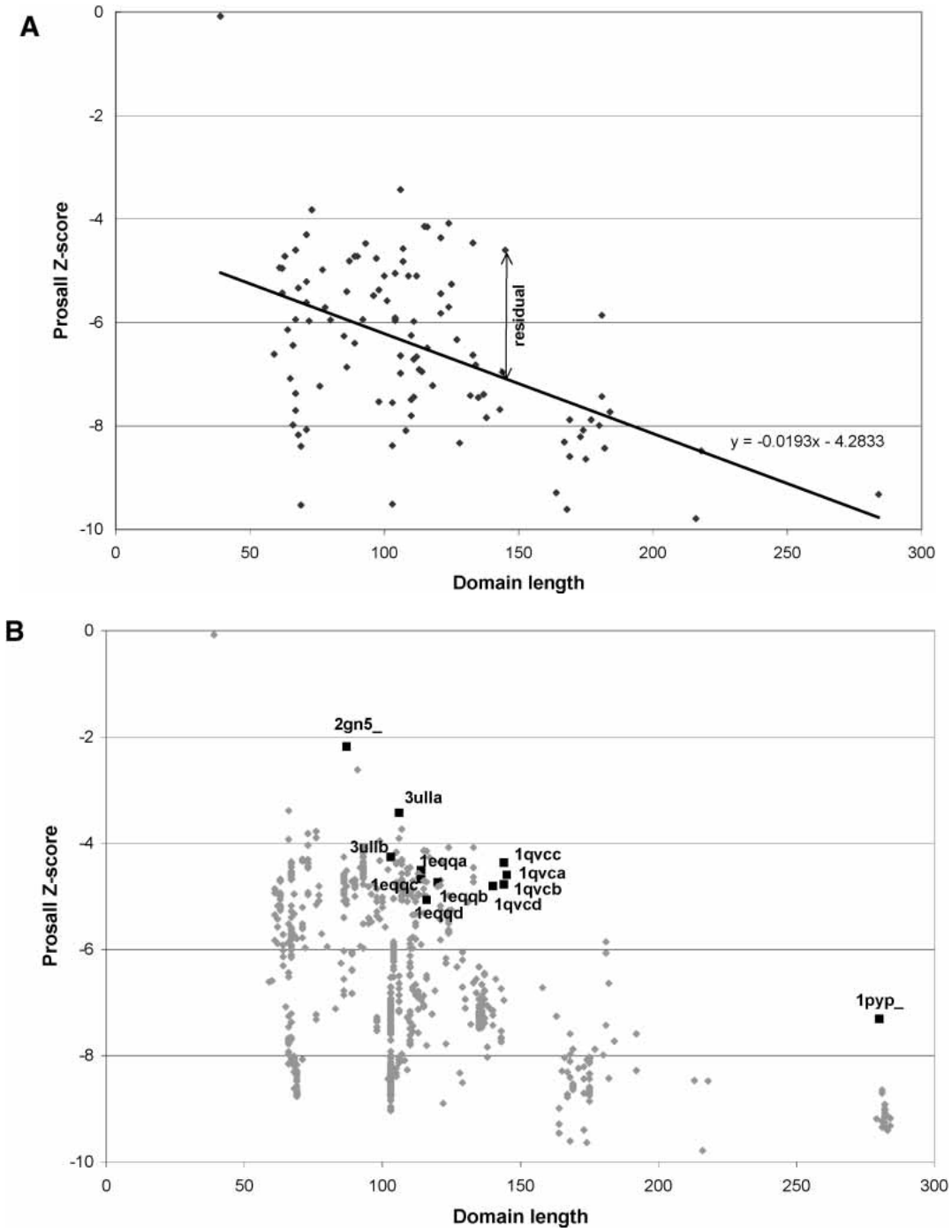

Figure 2. ProsaII Z-score distribution for OB-fold domains plotted as a function of protein sequence length. (*A*) Regression analysis based on a nonredundant set (at 95% sequence identity) of OB-domains. Both the regression line and its equation are displayed. (*B*) A complete set of OB-domains from ASTRAL 1.63 plus chains of the 1eqq entry. Protein chains containing sequence-register shift errors are represented as filled squares with their ASTRAL codes indicated.

(2.8Å) or 1kaw (2.9Å). However, to our surprise, we have found that it is 1qvc, the highest-resolution structure for the *E. coli* SSB to date, that has an apparent sequence-structure mapping error. The error is caused by a one-residue register shift and is confined to the last two β-strands (in a canonical description, both of these strands are equivalent to the fifth β-strand) of the OB-domain (Figs. 1, 3). This conclusion is backed by the fact that all four individual chains of 1qvc produce worse ProsaII Z-scores than any chain of the other two PDB entries for this protein. The SSB protein chains in

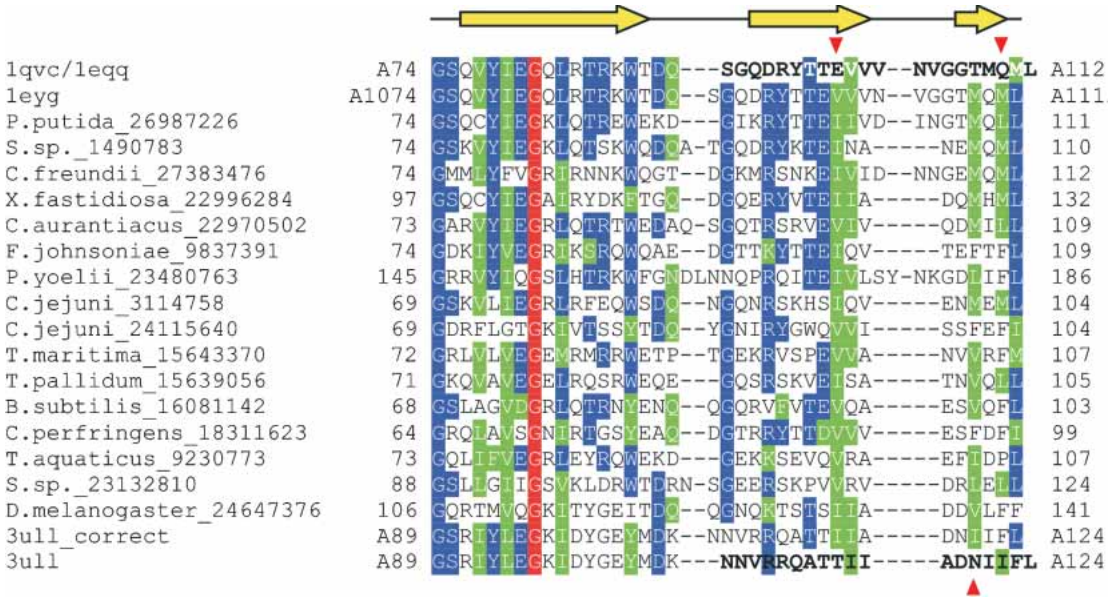

Figure 3. Multiple sequence alignment for selected single-stranded DNA binding (SSB) proteins indicating sequence-structure mapping errors for *E. coli* SSB (1qvc/1eqq) and human mitochondrial SSB (3ull). Each sequence is denoted either by the species name followed by the NCBI gene identification number or by the PDB code. Conserved residues shared by at least 50% of all sequences in the whole family are highlighted in blue (identical) and green (similar). Invariant glycine is highlighted in red. Regions affected by sequence-register shift errors are indicated in bold, and residues that deviate most from the conserved hydrophobic pattern of aligned SSB proteins are denoted by red triangles.

1qvc contain an extended C-terminal region missing from other structures of this protein. Initially, we considered a possibility that this nonglobular segment may be the actual reason for low ProsaII scores. However, the scores remained poor on its truncation, confirming that the problem resides within the OB-barrel structure. Importantly, ProsaII profiles for 1qvc exhibit a high-energy region, which coincides with the region of the residue mapping discrepancy. Inspection of the structure coupled with the analysis of the multiple sequence alignment immediately reveals the reason for that. Two structural positions (100 and 110 in 1qvc) are normally occupied by hydrophobic residues (Fig. 3), contributing their side chains to the packing of the OB-barrel interior. However, because of the sequence shift in 1qvc, these, mostly inaccessible to solvent, positions are filled with charged (Glu) and polar (Gln) residues, respectively, resulting in energetically unfavorable interactions.

Recently, one additional crystal structure of *E. coli* SSB protein (1eqq) has been released from the PDB. Although at the time of this study the structure had not yet been included into SCOP (or ASTRAL), we have analyzed it, too (Fig. 2B). It turns out that this newly released SSB structure has the identical sequence-register shift error as 1qvc (Fig. 3). Interestingly, despite major flaws, both structures have reasonably good R values (1qvc, R: 0.247; 1eqq, R: 0.213). However, these seemingly good values apparently are achieved only because of a significant overrefinement, as indicated by much higher corresponding R_{free} values (1qvc, R_{free}: 0.317; 1eqq, R_{free}: 0.339).

Human mitochondrial single-stranded DNA binding protein (3ull)

Error detection

The most elusive error that we have detected in the analyzed set of OB-fold domains was present in the structure of human mitochondrial SSB protein (HsmtSSB). The 3ull entry (Yang et al. 1997) has two protein chains in the asymmetric unit and is the only PDB entry for this protein at 2.4 Å resolution with relatively good R (0.195) and R_{free} (0.237) values, indicative of a well-refined structure. Thus, it was quite unexpected to see poor ProsaII Z-scores for 3ull (especially for the chain A) in comparison with other OB-domains of similar length (Fig. 2B). More detailed assessment of the structure with the ProsaII profiles has indicated that there are several high-energy regions, including one at the C terminus (Fig. 4). Next, we explored whether a residue mapping error might be responsible for the poor ProsaII evaluation results. Because this has been the only structure for HsmtSSB, we could not use the direct comparison method as for the *E. coli* SSB structures. Thus, we first produced a structure-based alignment between HsmtSSB and the correct structure (1eyg) of the *E. coli* SSB

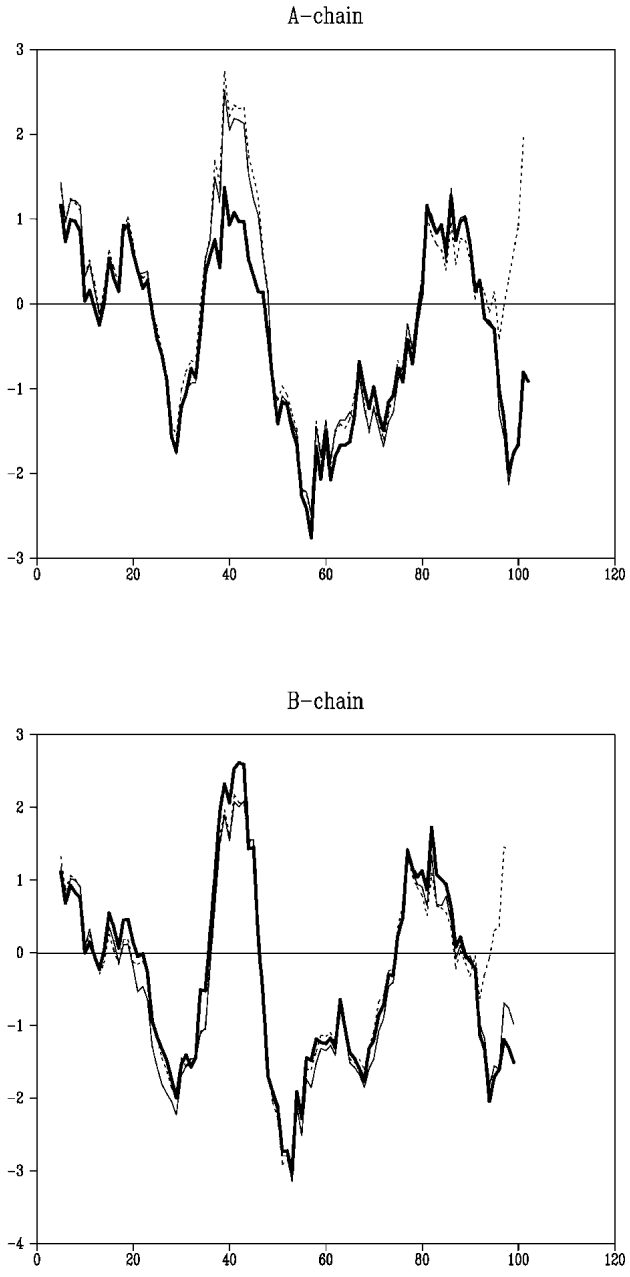

Figure 4. ProsaII energy profiles for the A- and B-chains of the original 3ull structure (dotted line), a computational model built from the corrected sequence register (thin solid line), and a revised X-ray structure (bold solid line). The X-axis denotes residue positions along the chain, and the Y-axis denotes energy values averaged over a sliding 10-residue window.

protein, which is the closest homolog in the PDB, sharing ~34% identical residues. Then we derived a second alignment between these two proteins using a sequence-only method (PSI-BLAST-ISS, see Materials and Methods). Once we contrasted the two alignments obtained by different means, we found that they differ in the C-terminal region by the one-residue shift (Fig. 3), suggesting the presence of a sequence-structure mapping error in 3ull. In the following step, we generated 3D models for both chains of HsmtSSB with MODELLER using the original structure as a template and the new alignment at the C terminus. In the ProsaII evaluation, new models for both chains fared notably better than the original 3ull structure. ProsaII profiles showed that the energy "spike" at the C terminus has disappeared (Fig. 4). Accordingly, ProsaII Z-scores improved from −3.43 to −4.41 for the A-chain and from −4.26 to −5.27 for the B-chain.

Revision of the X-ray structure

The findings of the computational analysis have prompted us to reanalyze the original crystallographic data for the HsmtSSB structure. The region including residues 105–111 in both chains of the structure constitutes a disordered loop, which does not have a well-defined electron density. On the basis of the suggestion from computational results, the residues corresponding to 107–124 were shifted by one position through the introduction of an addition residue in this disordered loop.

These newly generated models for the two chains were refined with the original X-ray data using X-PLOR (Brünger 1992), producing an initial R-value of 0.27 (2.4 Å). After several cycles of positional refinement, simulated annealing refinement, and the temperature factor refinement, we were able to fit the substituted residues to the electron density. Backbones for residues 110–124 of the original and new models are nearly superimposable with the RMSD value of 0.41 Å. The final structure has an R-value of 0.191 (R_{free} = 0.23) and the RMS deviations (from the standard geometry) are 0.016 Å for bonds and 3.6° for angles, which is very similar to the values for the original structure.

The calculated $2F_o$-F_c and SA-omit maps between residues 108 and 125 show a similar quality in both the original and the revised structure. For example, Arg 110 and Arg 111 fit similarly into both models because corresponding residue positions do not have any electron density beyond Cβ atoms. The electron density of the SA-omit map corresponding to the residue 123 fits better with phenylalanine instead of isoleucine, and the asymmetric shape of the density corresponding to the residue 116 favors the side chain of isoleucine rather than threonine, thus supporting the revised model. But the side chain of the residue 112 shows an elongated density that fits better with arginine instead of glutamine and thus favors the old model.

Although X-ray data overall show a slight preference for the new model, because of the low resolution or intrinsic flexibility in the C-terminal region, it would be difficult to decide which model is the correct one based solely on the original electron density map. The unusually high number of duplets composed of identical or very similar residues (Fig. 3) in this fairly short sequence region of HsmtSSB does not help the situation either. On the other hand, only

one of the alternative sequence-structure mappings can be expected to represent the folded structure of HsmtSSB. It would be highly unlikely that the HsmtSSB C-terminal region would be able to adopt two structurally similar conformations, yet have a major difference in the sequence register. Indeed, energy considerations (Fig. 4) coupled with the evolutionary analysis (Fig. 3) in addition to the electron density map, make the revised HsmtSSB structure clearly favored over the original one. Both the ProsaII Z-scores (−4.87 for chain A; −4.98 for chain B) and energy profiles for the revised structure show significant improvement (Fig. 4) and are comparable to computationally derived models. It is interesting to note that both ProsaII and PSI-BLAST-ISS were effective in making such a clear distinction between the two alternatives within the problematic region despite the fact that in a multiple sequence alignment only a single position within the original 3ull structure (Asn 120) notably deviates from the conservation pattern of aligned SSB proteins (Fig. 3).

The coordinates of the revised HsmtSSB structure have been deposited with the PDB (1s3o).

Inorganic pyrophosphatase, budding yeast (1pyp)

The budding yeast inorganic pyrophosphatase, like the *E. coli* SSB protein, is represented in PDB by multiple structures, some determined at a very high resolution (e.g., 1e9g [Heikinheimo et al. 2001]; 1.15 Å resolution, R value: 0.136). Among all of the structures for this enzyme, only the earliest one (1pyp) determined at 3 Å-resolution (Arutiunian et al. 1981) has displayed a significantly worse ProsaII Z-score. A subsequent superposition of 1pyp and a high resolution structure (1e9g) revealed that the amino acid sequence has been mapped incorrectly in several regions of the 1pyp structure (Fig. 5).

Gene 5 DNA binding protein from bacteriophage fd (2gn5)

Unlike the earlier cases, the 2.3 Å resolution structure of the gene 5 DNA binding protein from bacteriophage fd (2gn5; Brayer and McPherson 1983) has been suspected to have problems in a number of previous reports. Both the structure quality assessment methods (Morris et al. 1992; Sippl 1993) and a comparison with the results of an NMR study (Folkers et al. 1991) pointed to the presence of structural flaws in 2gn5. This is not surprising in the view of massive sequence register errors present in this structure. In the sequence-independent structural superposition of 2gn5 with the higher resolution structure of the same protein (1vqb [Skinner et al. 1994]; 1.8 Å resolution) the protein backbone differences are within a reasonable range (RMSD = 1.6 Å for 81 Cα atoms), but only 51% of residues are identical in the structurally equivalent positions. In other words, close to half of the residues in the 2gn5 structure have been assigned incorrect positions.

The ProsaII evaluation, coupled with the selection scheme for further analysis (based on residuals calculated during the regression analysis), proved to be an effective tool for uncovering potential errors in OB-domains. After ProsaII Z-scores were sorted according to their residuals, nine protein chains from all five flawed PDB entries were within 30 "most unusual" structures. Some of other poor Z-scores turned out to be simply the result of incompleteness of a structure assignment. If extensive chain regions are missing from the 3D structure, the structural integrity is affected and therefore the structure does not score as well during the energy evaluation. A good example of such a case is a partly disordered S1 RNA-binding domain of polynucleotide phosphorylase (1e3p; Symmons et al. 2000) for which coordinates of only 39 residues are assigned (top left in Fig. 2). Poor scores are also obtained when OB-domains have noncompact insertions/extensions beyond the canonical β-barrel framework (see Fig. 1 for the canonical OB-fold). Such noncompact regions are often favorably interacting with other domains in multidomain proteins, and Z-score becomes favorable if either the region is removed or the complete multidomain structure is evaluated. For consistency, we have not modified boundaries of ASTRAL domains, and that is why ProsaII Z-scores for a number of apparently correct structures look poor. If one ignores the noise from these correct structures that have their integrity compromised, the erroneous structures in the Figure 2 plot provide an empirical baseline for a scrutiny of any new OB-structure that scores similarly or worse.

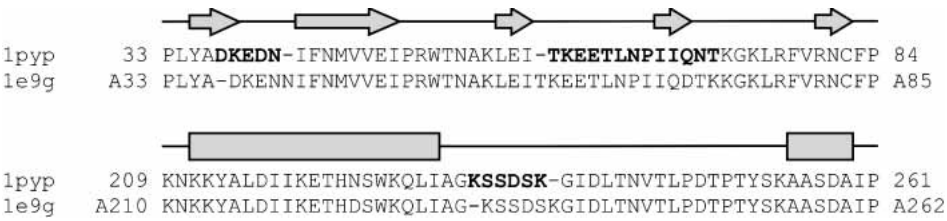

Figure 5. Structure-based sequence alignment between the two versions of the budding yeast inorganic pyrophosphatase structure: 1pyp and a high-resolution structure of the same protein (1e9g). Sequence register errors within 1pyp are denoted in bold.

Although we have only focused on OB-fold domains, we believe that the computational approach presented here can be applied for the detection of sequence register errors in protein structures of virtually any fold. Independently of which folds are analyzed, the overall procedure can be summarized as follows: (1) detection of "unusual" structures using computational quality assessment such as ProsaII Z-scores and energy profiles; (2) producing a structure-based alignment (e.g. DALI) between the "suspected" structure and a structure of a homologous protein; (3) generating a reliable structure-independent alignment (e.g., with PSI-BLAST-ISS) that includes the same pair of proteins; (4) identification of potential sequence-register shifts, manifesting themselves as differences between the structure-derived alignment and the one derived without explicitly using structural information; (5) producing a 3D model of a "suspected" structure using a suggested alternative sequence-structure mapping and contrasting the corresponding model with the initial structure using both energy (ProsaII Z-scores, energy profiles, visual inspection) and evolutionary (multiple sequence alignments) considerations. Perhaps it should also be emphasized that this general scheme can easily accommodate methods for structure quality assessment, structure superposition, and sequence alignment other than those we have used in this study.

Discussion

The PDB is one of the most important primary resources of biological information, and its data are increasingly used to derive new secondary databases, and develop and test various computational methods. In such a PDB-centric world of structural biology data, errors present in PDB structures may easily propagate into derivative databases and/or negatively affect computational efforts in biological research. The chances for these errors to spread increase greatly when the incorrect structure is either considered to be the representative structure or it is the only available structure for the entire protein family. Unfortunately, sequence-mapping errors in the structures of OB-fold domains provide such an example. During the most recent worldwide "blind" testing of protein structure prediction methods (CASP5; Moult et al. 2003), one of the sequences submitted as modeling targets was the SSB protein from *Mycobacterium tuberculosis* (Saikrishnan et al. 2003; T0151, see http://predictioncenter.llnl.gov/casp5 for details). The *E. coli* SSB protein turned out to be the closest available structural template that could be used to model T0151 by comparison. Many predictor groups automatically selected the *E. coli* SSB structure having the best resolution (1qvc, 2.2 Å), being a structural representative in the FSSP database (Holm and Sander 1996), yet having sequence-structure mapping errors. Analysis of the prediction results for this CASP5 target showed that over 95% of predictor groups failed to produce correct alignment at the C-terminal region of T0151, corresponding to the incorrect sequence-structure mapping region in 1qvc (Venclovas 2003). As a result, it is impossible to objectively compare the performance of different modeling methods for this particular prediction target.

On the basis of the analysis of crystal structures of protein domains possessing OB-fold, we can at least roughly estimate the extent of significant structural flaws present in the entire PDB. If we add the recently released incorrect structure for the *E. coli* SSB protein (1eqq) to the ASTRAL 1.63 set of OB-domains, the error rate at the level of individual chains for this limited subset of the PDB runs at 1.4% (12/846 chains). Note that in this study we only concerned ourselves with the kind of structural errors that can be unambiguously pinned down and a straightforward solution to correct them can be provided. There also might be other, less obvious, structural errors that nevertheless affect the quality of protein structural data. We cannot exclude a possibility that assigning coordinates for X-ray data of OB-fold domains is more error prone than for structures of many other folds. Nevertheless, on the basis of our findings, it would not be unreasonable to estimate that up to 1% of protein chains in the PDB might have significant structural errors.

On the other hand, it is encouraging that we have not found any sequence-register shift errors in high-resolution structures. The highest resolution at which errors were found to be present is 2.2 Å (1qvc). Although this does not necessarily mean that there are no errors within structures solved at a higher resolution than that, it suggests that it might be reasonable to use 2 Å resolution as a cutoff for filtering out most, if not all, faulty protein structures.

What can be done to prevent incorrect structures from making their way into the PDB in the first place? Obviously, the heaviest burden lies on the experimentalists (X-ray crystallographers and NMR spectroscopists) to provide correct interpretation of structural data. On the other hand, computational methods that do not use the underlying experimental data can provide an increasing help in this task. As shown in this work and also in another recent study (Bujnicki et al. 2002), the protein structure evaluation combined with sequence analysis techniques can be especially effective in detecting and correcting sequence-structure mapping errors. Because of continuous improvement in computational methods and an explosive increase in volume of protein sequence data, it is often possible to produce reliable alignments even at the very low level of sequence homology (Venclovas 2003). Importantly, the improving ability to generate reliable alignments can also facilitate the experimental structure determination. It might be used to either guide the sequence assignment process or verify sequence-structure mapping within poorly defined regions, even if the only available structural reference is a remotely related protein.

Materials and methods

The energies for 3D structures of OB-domains were estimated using ProsaII (Sippl 1993), a method based on empirical mean force potentials. ProsaII can produce both overall energy evaluation (Z-score) and a profile that displays energy properties of the protein structure as a function of the amino acid sequence position. The more negative the Z-score value (which is protein-length dependent), the lower the estimated energy, and thus the protein structure is less likely to contain errors. Likewise, positive values in energy profiles point to strained sections of the chain resulting from either unusual local packing or surface composition of the structure, whereas negative values indicate stable parts of the protein.

Structure-based alignments were generated from pairs of structures superimposed in a sequence-independent manner with DALI (Holm and Sander 1993).

The region-specific estimation of sequence-based alignment reliability for a pair of sequences was done with the previously developed PSI-BLAST Intermediate Sequence Search (PSI-BLAST-ISS) procedure (Venclovas 2001). Briefly, in this procedure, a set of proteins (~50–150) homologous to both sequences are used individually as seeds to generate corresponding PSI-BLAST (Altschul et al. 1997) profiles, usually not exceeding five iterations. Using the SEALS package (Walker and Koonin 1997) and in-house Perl scripts, the alignments between the two sequences of interest are extracted from each of these PSI-BLAST output files and compared. The result of this procedure is a multiple sequence alignment, where the first sequence is aligned with a number of copies of the second sequence in the way that corresponds to different PSI-BLAST output files. If the second sequence in most instances is aligned to the first one in the identical way (a single major alignment variant), the region is considered to be aligned reliably.

Alignments within the sequence family (high level of homology) were generated by first collecting close homologs using a standard BLAST search and subsequently aligning them using T-coffee (Notredame et al. 2000) with only minimal manual adjustments. Molecular models were built automatically from the sequence-structure alignments with MODELLER (Šali and Blundell 1993).

Acknowledgments

This research was performed in part under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48. Č.V. also acknowledges support from Howard Hughes Medical Institute.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Abagyan, R.A. and Totrov, M.M. 1997. Contact area difference (CAD): A robust measure to evaluate accuracy of protein models. *J. Mol. Biol.* **268:** 678–685.

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

Arutiunian, E.G., Terzian, S.S., Voronova, A.A., Kuranova, I.P., Smirnova, E.A., Vainstein, B.K., Hohne, W.E., and Hansen, G. 1981. X-ray-diffraction study of inorganic pyrophosphatase from bakers-yeast at a 3 Å resolution. *Dokl. Akad. Nauk SSSR* **258:** 1481–1485.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28:** 235–242.

Brayer, G.D. and McPherson, A. 1983. Refined structure of the gene 5 DNA binding protein from bacteriophage fd. *J. Mol. Biol.* **169:** 565–596.

Brenner, S.E., Koehl, P., and Levitt, M. 2000. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* **28:** 254–256.

Brünger, A.T. 1992. X-PLOR A system for X-ray crystallography and NMR, Version 3.1. Yale University, New Haven, CT.

Bujnicki, J., Rychlewski, L., and Fischer, D. 2002. Fold-recognition detects an error in the Protein Data Bank. *Bioinformatics* **18:** 1391–1395.

Doreleijers, J.F., Rullmann, J.A., and Kaptein, R. 1998. Quality assessment of NMR structures: A statistical survey. *J. Mol. Biol.* **281:** 149–164.

Folkers, P.J., van Duynhoven, J.P., Jonker, A.J., Harmsen, B.J., Konings, R.N., and Hilbers, C.W. 1991. Sequence-specific 1H-NMR assignment and secondary structure of the Tyr41–His mutant of the single-stranded DNA binding protein, gene V protein, encoded by the filamentous bacteriophage M13. *Eur. J. Biochem.* **202:** 349–360.

Heikinheimo, P., Tuominen, V., Ahonen, A.K., Teplyakov, A., Cooperman, B.S., Baykov, A.A., Lahti, R., and Goldman, A. 2001. Toward a quantum-mechanical description of metal-assisted phosphoryl transfer in pyrophosphatase. *Proc. Natl. Acad. Sci.* **98:** 3121–3126.

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

———. 1996. Mapping the protein universe. *Science* **273:** 595–603.

Hooft, R.W., Vriend, G., Sander, C., and Abola, E.E. 1996. Errors in protein structures. *Nature* **381:** 272.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., and Thornton, J.M. 1993. Procheck—A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26:** 283–291.

Luthy, R., Bowie, J.U., and Eisenberg, D. 1992. Assessment of protein models with three-dimensional profiles. *Nature* **356:** 83–85.

Matsumoto, T., Morimoto, Y., Shibata, N., Kinebuchi, T., Shimamoto, N., Tsukihara, T., and Yasuoka, N. 2000. Roles of functional loops and the C-terminal segment of a single-stranded DNA binding protein elucidated by X-Ray structure analysis. *J. Biochem. (Tokyo)* **127:** 329–335.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G., and Thornton, J.M. 1992. Stereochemical quality of protein structure coordinates. *Proteins* **12:** 345–364.

Moult, J., Fidelis, K., Zemla, A., and Hubbard, T. 2003. Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins* **53 (Suppl. 6):** 334–339.

Murzin, A.G. 1993. OB(oligonucleotide/oligosaccharide binding)-fold: Common structural and functional solution for non-homologous sequences. *EMBO J.* **12:** 861–867.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536–540.

Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302:** 205–217.

Raghunathan, S., Ricard, C.S., Lohman, T.M., and Waksman, G. 1997. Crystal structure of the homo-tetrameric DNA binding domain of *Escherichia coli* single-stranded DNA-binding protein determined by multiwavelength x-ray diffraction on the selenomethionyl protein at 2.9-Å resolution. *Proc. Natl. Acad. Sci.* **94:** 6652–6657.

Raghunathan, S., Kozlov, A.G., Lohman, T.M., and Waksman, G. 2000. Structure of the DNA binding domain of *E. coli* SSB bound to ssDNA. *Nat. Struct. Biol.* **7:** 648–652.

Ratnaparkhi, G.S., Ramachandran, S., Udgaonkar, J.B., and Varadarajan, R. 1998. Discrepancies between the NMR and X-ray structures of uncomplexed barstar: Analysis suggests that packing densities of protein structures determined by NMR are unreliable. *Biochemistry* **37:** 6958–6966.

Saikrishnan, K., Jeyakanthan, J., Venkatesh, J., Acharya, N., Sekar, K., Varshney, U., and Vijayan, M. 2003. Structure of Mycobacterium tuberculosis single-stranded DNA-binding protein. Variability in quaternary structure and its implications. *J. Mol. Biol.* **331:** 385–393.

Šali, A. and Blundell, T.L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234:** 779–815.

Schmitt, E., Moulinier, L., Fujiwara, S., Imanaka, T., Thierry, J.C., and Moras, D. 1998. Crystal structure of aspartyl-tRNA synthetase from Pyrococcus kodakaraensis KOD: Archaeon specificity and catalytic mechanism of adenylate formation. *EMBO J.* **17:** 5227–5237.

Sippl, M.J. 1993. Recognition of errors in three-dimensional structures of proteins. *Proteins* **17:** 355–362.

Skinner, M.M., Zhang, H., Leschnitzer, D.H., Guan, Y., Bellamy, H., Sweet, R.M., Gray, C.W., Konings, R.N., Wang, A.H., and Terwilliger, T.C. 1994. Structure of the gene V protein of bacteriophage f1 determined by multiwavelength x-ray diffraction on the selenomethionyl protein. *Proc. Natl. Acad. Sci.* **91:** 2071–2075.

Symmons, M.F., Jones, G.H., and Luisi, B.F. 2000. A duplicated fold is the structural basis for polynucleotide phosphorylase catalytic activity, processivity, and regulation. *Structure Fold. Des.* **8:** 1215–1226.

Venclovas, Č. 2001. Comparative modeling of CASP4 target proteins: Combining results of sequence search with three-dimensional structure assessment. *Proteins* **(Suppl.) 5:** 47–54.

———. 2003. Comparative modeling in CASP5: Progress is evident, but alignment errors remain a significant hindrance. *Proteins* **53 (Suppl. 6):** 380–388.

Walker, D.R. and Koonin, E.V. 1997. SEALS: A system for easy analysis of lots of sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5:** 333–339.

Yang, C., Curth, U., Urbanke, C., and Kang, C. 1997. Crystal structure of human mitochondrial single-stranded DNA binding protein at 2.4 Å resolution. *Nat. Struct. Biol.* **4:** 153–157.