Protein Science (1997), 6:246–248. Cambridge University Press. Printed in the USA.

FOR THE RECORD

LPFC: An internet library of protein family core structures

ROBERT SCHMIDT,[1] MARK GERSTEIN,[2] AND RUSS B. ALTMAN[1]

[1]Section on Medical Informatics, Stanford University, MSOB X-215, Stanford, California 94305-5479
[2]Department of Structural Biology, Stanford University, Fairchild Laboratories, Stanford, California 94305-5400

(RECEIVED May 28, 1996; ACCEPTED October 11, 1996)

Abstract: As the number of protein molecules with known, high-resolution structures increases, it becomes necessary to organize these structures for rapid retrieval, comparison, and analysis. The Protein Data Bank (PDB) currently contains nearly 5,000 entries and is growing exponentially. Most new structures are similar structurally to ones reported previously and can be grouped into families. As the number of members in each family increases, it becomes possible to summarize, statistically, the commonalities and differences within each family. We reported previously a method for finding the atoms in a family alignment that have low spatial variance and those that have higher spatial variance (i.e., the "core" atoms that have the same relative position in all family members and the "non-core" atoms that do not). The core structures we compute have biological significance and provide an excellent quantitative and visual summary of a multiple structural alignment. In order to extend their utility, we have constructed a library of protein family cores, accessible over the World Wide Web at http://www-smi.stanford.edu/projects/helix/LPFC/. This library is generated automatically with publicly available computer programs requiring only a set of multiple alignments as input. It contains quantitative analysis of the spatial variation of atoms within each protein family, the coordinates of the average core structures derived from the families, and display files (in bitmap and VRML formats). Here, we describe the resource and illustrate its applicability by comparing three multiple alignments of the globin family. These three alignments are found to be similar, but with some significant differences related to the diversity of family members and the specific method used for alignment.

Keywords: database integration; globins; protein cores; protein families; structural variability; world wide web

Our core-defining procedure takes a multiple structural alignment of macromolecules as input and computes an average structure while identifying each of the aligned atoms as either "core" or "non-core" (Altman & Gerstein, 1994; Gerstein & Altman, 1995a, 1995b). Our definition of core is entirely structural; that is, core atoms are those whose positions are essentially invariant throughout the structures of a given family. Non-core atoms are present in all family members, but have positions that vary more widely within the family. (The final category of atoms are those that are not present in all family members, and thus not part of a multiple structural alignment.) We have applied our core-finding procedure successfully to the globin and immunoglobulin families, using benchmark hand-derived multiple alignments (Gerstein & Altman, 1995a, 1995b). The structural core correlates remarkably well with a number of important biological features of these families (including, for example, the identification of functional sites, correlation with experimental evidence on folding, and genomic organization).

The Library of Protein Family Cores (LPFC) is a large-scale extension of such analyses. For the first release of the LPFC, we have used the multiple alignments generated by two automatic alignment procedures: (1) that of Overington and colleagues [as contained in the HOMALDB database (Overington et al., 1993; Šali & Overington, 1994)] and (2) that of Holm and Sander [as contained in the FSSP database (Holm & Sander, 1994)]. The LPFC affords access to the results of our core computations [as raw data, images, or VRML (Bell et al., 1995)], as well as the original computer programs used to perform these computations. The LPFC code can be used to generate an accessible library of core structure for any multiple alignment method.

In order to demonstrate the ease of both visual and quantitative comparison of multiple alignment strategies using LPFC, we illustrate a comparison of the multiple alignments of the globins as reported in the FSSP database, the HOMALDB database, and a very reliable hand alignment (Lesk & Chothia, 1980; Gerstein et al., 1994b).

Results and discussion: The LPFC (version 1.0) contains information about average core structures for more than 280 protein families (some of which overlap), taken from both the HOMALDB (47 families) and FSSP (233 families) databases, as well as two "gold-standard" manual alignments of the globins and immunoglobulins (Kabat et al., 1983). The number of structures per family ranges from 3 (for many families) to 18 (the globin family), with

Reprint requests to: Russ B. Altman, Section on Medical Informatics, Stanford University, MSOB X-215, Stanford, CA 94305-5479; e-mail: russ.altman@stanford.edu.

an average of 6.9 proteins per family. Each protein family contains an average of 123 aligned residues, of which 87 are low variance, core atoms. The vast majority of core atoms belong to alpha-helical or beta-strand segments, with loops tending to be less conserved across family members and tending to have higher spatial variance within the families. We have linked the LPFC to other relevant protein structural resources such as scop (Murzin et al., 1995), the Protein Motions Database (Gerstein, 1995), SWISS-PROT (Bairoch & Boeckmann, 1992), and others. LPFC can be accessed with a generic universal resource locator (URL) using a single Protein Data Bank (PDB) identifier or list of identifiers, as described in the Electronic Appendix.

Figure 1 shows graphical summaries of three globin core computations, produced by our core finding procedure and displayed with the *proteanD* program (Altman et al., 1995). The one from FSSP has 13 structures and 104 aligned α-carbons, with an "ellipsoid of variation" on average 2.5 $Å^3$ in volume; the one from HOMALDB has 18 structures and 112 aligned α-carbons, with an ellipsoid volume of 1.26 $Å^3$; and the hand alignment of 8 structures has 115 aligned α-carbons, with a volume of 1.09 $Å^3$. The first qualitative observation of the three globin alignments is that the regions with the larger green ellipsoids are quite similar throughout all three alignment methods. In particular, the region of helix F, the helix that contains the distal histidine known to be critical for oxygen binding, is a region of high structural variability across the family. In addition, some regions near helix F and some surface helical ends have high variability. We have shown previously that the areas of non-core atoms in the globins are those that are critical in determining oxygen-binding properties, are on different exons than the core atoms, and represent parts of the structure that fold relatively late (Altman & Gerstein, 1994; Gerstein & Altman, 1995b).

The three globin structural alignments we used are interesting because they share very few specific globin structures. All three alignment methods produce visually similar average structures. The inclusion of a wider range of evolutionarily related proteins will, in general, tend to increase the overall volume of variability and may lead to significant differences in the rank order of volumes within the family. In the case of the three globin alignments analyzed here, the rank order of atomic variability is somewhat variable across the methods, with a correlation coefficient of only approximately 0.2. The average volumes observed also show some variability, reflecting both the number of structures used to define the family in the multiple alignment, as well as the range of different structures that are included in the alignment.

The ellipsoid representation used for display within LPFC does not become cluttered as the number of structures increases, and this allows visual inspection of core structures as a check on the multiple alignments. Possible errors in alignments become obvious visually, because the ellipsoids are forced to take on highly eccentric shapes in order to include atoms that are aligned erroneously. However, the variability evident in structural alignments can come from two different processes: the adaptive changes in protein structure over evolution or the intrinsic flexibility of any particular class of protein molecules. Alignments that show very high variability may actually be the result of aligning structures in different conformational states. In fact, the two alignments with the greatest variability in LPFC, those of the serpins and the adenylate kinases (serpin and adk families in HOMALDB), both manifest the effects of motions (Stein & Chothia, 1991; Gerstein et al., 1993). The Protein Motions Database provides comprehensive information about known protein motions (Gerstein et al., 1994a; Gerstein, 1995; http://hyper.stanford.edu/~mbg/ProtMotDB).

The amino acid positional uncertainties computed with our method are not related to the crystallographic *B*-factor in any simple way. There are many examples of internal segments that are very well resolved in each of the constituent family members, but that have high variability across the family. Similarly, there are some examples of external surface elements that are part of the family core. The *B*-factors of well-resolved proteins, such as we have used in computing these cores, tend to be significantly smaller than the variabilities we compute across protein families.

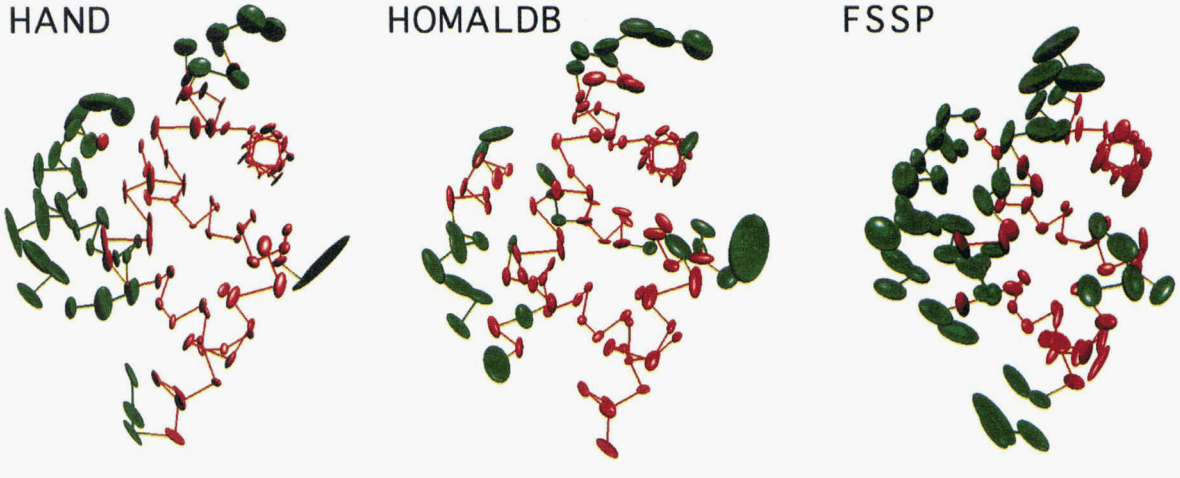

Fig. 1. Summary of LPFC entries for globins from hand alignment, HOMALDB database, and FSSP database. Each atom in the alignment is drawn with an ellipsoid enclosing the positions assumed among family members. Green ellipsoids correspond to high variability positions, and red ellipsoids correspond to lower variability positions. The globin families, as described in the text, are shown in a standard orientation (looking down the axis of Helix B in the upper right, and with Helix F at the far left). Despite detailed differences, it is apparent that all three alignment techniques identify a core region in the center of the molecules comprising helices A, B, G, and H (with other helices represented to a variable degree). The three methods show high variability in sections entering loops, as well as in the region around Helix F, which packs the heme group against the rest of the scaffolding. Each of these images can be viewed over the internet with a VRML viewer.

Supplementary material in Electronic Appendix: We have included an Electronic Appendix that contains the details of how the LPFC is generated, as well as tables summarizing the resulting statistics of the protein families

Acknowledgments: R.B.A. is a Culpeper Medical Scholar and is supported by NIH LM-05652 and LM-06422. Computing resources were provided by the CAMIS NIH LM-05305. M.G. is a Damon-Runyon Fellow (DRG-1272). We thank L. Holm and C. Sander for use of the FSSP database and for assistance in processing the FSSP files. We thank A. Šali and J.P. Overington for providing the HOMALDB database.

References

Altman R, Gerstein M. 1994. Finding an average core structure: Application to the globins. *Proc Second Int Conf Intell Sys Mol Biol*. Menlo Park, California: AAAI Press.

Altman RB, Hughes C, Gerstein MB. 1995. Methods for displaying macromolecular structural uncertainty: Application to the globins. *J Mol Graphics 13*:142–152.

Bairoch A, Boeckmann B. 1992. The Swiss-Prot protein-sequence data-bank. *Nucleic Acids Res 20*:2019–2022.

Bell G, Parisi A, Pesce M. 1995. *The virtual reality modeling language (VRML) 1.0 specification*. Mountain View, California: Silicon Graphics.

Gerstein M. 1995. A protein motions database. *PDB Quarterly Newsletter 73*:2.

Gerstein M, Altman RB. 1995a. Using a measure of structural variation to define a core for the globins. *Comput Applic Biosci 11*:633–644.

Gerstein M, Altman RB. 1995b. Average core structures and variability measures for protein families: Application to the immunoglobulins. *J Mol Biol 251*:161–175.

Gerstein M, Lesk AM, Chothia C. 1994a. Structural mechanisms for domain movements. *Biochemistry 33*:6739–6749.

Gerstein M, Schulz G, Chothia C. 1993. Domain closure in adenylate kinase: Joints on either side of two helices close like neighboring fingers. *J Mol Biol 229*:494–501.

Gerstein M, Sonnhammer E, Chothia C. 1994b. Volume changes on protein evolution. *J Mol Biol 236*:1067–1078.

Holm L, Sander C. 1994. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res 24*:204–210.

Kabat EA, Wu TT, Bilofsky H, Reid-Milner M, Perry H. 1983. *Sequences of proteins of immunological interest*. Washington, DC: National Institutes of Health.

Lesk AM, Chothia CH. 1980. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *J Mol Biol 136*:225–270.

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

Overington J, Zhu Z, Šali A, Johnson M, Sowdhamini R, Louie G, Blundell T. 1993. Molecular recognition in protein families: A database of aligned three-dimensional structures of related proteins. *Biochem Soc Trans Part 3*:597–604.

Šali A, Overington J. 1994. Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci 3*:1582–1596.

Stein P, Chothia C. 1991. Serpin tertiary structure transformation. *J Mol Biol 221*:615–621.