© 2000 Oxford University Press

The ASTRAL compendium for protein structure and sequence analysis

Steven E. Brenner*, Patrice Koehl and Michael Levitt

Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126, USA

Received September 1, 1999; Revised and Accepted October 13, 1999

ABSTRACT

The ASTRAL compendium provides several databases and tools to aid in the analysis of protein structures, particularly through the use of their sequences. The SPACI scores included in the system summarize the overall characteristics of a protein structure. A structural alignments database indicates residue equivalencies in superimposed protein domain structures. The PDB sequence-map files provide a linkage between the amino acid sequence of the molecule studied (SEQRES records in a database entry) and the sequence of the atoms experimentally observed in the structure (ATOM records). These maps are combined with information in the SCOP database to provide sequences of protein domains. Selected subsets of the domain database, with varying degrees of similarity measured in several different ways, are also available. ASTRAL may be accessed at http://astral.stanford.edu/

BACKGROUND

Three-dimensional coordinate structures provide insight into proteins' function, mechanism and evolution. The growth of structural information available (1–3) has caused its influence to pervade molecular biology, and now homology-based methods can be used to provide an outline model for nearly half the proteins in a completely sequenced genome (4). Structural genomics efforts explicitly aim to increase that figure (5,6).

The Protein Data Bank (PDB) is a centralized resource providing free access to protein structures (7), and it is augmented by a host of domain and classification databases such as 3Dee, CATH, DDBASE, Entrez and SCOP, which imbue the structures with context and analysis (8–13). Among these, the SCOP database is unique for being a fully curated manual classification indicating distinct levels of relationship, and hand-defined protein domain definition. Provision of both the classification and domain definition are important both for recognizing structures of distant homologs based on sequence and for understanding the functional, structural and evolutionary context for these proteins.

Unfortunately, the nature of PDB files often makes it challenging to accurately provide a linkage between the biological sequence and reported structure of a given protein. Identifying domains within that sequence is a further task. Finally, the majority of domain sequences in the PDB are very similar to others, and it is frequently helpful to reduce the redundancy by selecting high-quality representatives.

To address these issues and aid the use of protein structures and their associated sequences, the ASTRAL compendium augments SCOP with tools, resources and libraries. At present, the principal resources provided by ASTRAL are sequence databases corresponding to the domains of structures in the SCOP database. Also available are selections from these databases intended to provide high-quality subsets with low redundancy at desired degrees of similarity. In order to choose proteins for these selections, we use the Summary PDB ASTRAL Check Index (SPACI) which provides a first-order estimate of the resolution and regularity of crystallographically determined protein structures. ASTRAL also provides a mapping between the PDB ATOM and SEQRES fields, based on the alignments provided by the pdb2cif program (14). An added feature of ASTRAL is a library of structural alignments from SCOP 1.38, produced by STRUCTAL (15).

SUMMARY PDB ASTRAL CHECK INDEX (SPACI)

The PDB contains coordinate entries of varying quality that may contain irregularities (16). The SPACI score is intended to provide an approximate measure to report these characteristics of structures that have been determined by X-ray crystallography. This score is useful to provide a crude overview of structure quality, and it is particularly valuable for selecting a representative from large numbers of PDB entries for very closely related protein structures.

The SPACI score incorporates three different quantities: the resolution of the original data, how well the model fits the data (R-factor), and stereochemical check parameters which indicate how well the structure complies with standard molecular geometry:

$$\text{SPACI} = \frac{1}{\text{Resolution}} + (0.1 - \text{R-factor}) + \text{SCS}$$

The first term of the SPACI index is the reciprocal of the resolution. This term usually dominates, and consequently SPACI scores of 0.4 or greater typically represent good structures.

*To whom correspondence should be addressed at present address: Department of Plant and Microbial Biology, 461A Koshland, University of California, Berkeley, CA 94720-3102, USA. Tel: +1 510 462 9999; Fax: +1 208 279 8978; Email: brenner@compbio.berkeley.edu

The Stereochemical Check Score (SCS) combines scores (called WCK1-4 and PCK1-3 here) provided by WHATCHECK (17) and PROCHECK (18):

$$\text{SCS} = 0.1 - (0.1{*}\text{WCKS}) - (0.1{*}\text{PCKS})$$

$$\text{WCKS} = \left(1 - \frac{1}{4}\left(\frac{\text{WCK1} + 8}{16} + \frac{\text{WCK2} + 8}{12} + \frac{\text{WCK3} + 5}{7} + \frac{\text{WCK4} + 10}{15}\right)\right)$$

$$\text{PCKS} = \left(\frac{(\text{PCK1} + \text{PCK2} + \text{PCK3}) - 3}{6}\right)$$

Details of this score are provided on the ASTRAL website. The SPACI score only provides a rough estimate of the quality of the structure, and omits useful information such as R-free (19), so it is no substitute for knowing details of the structure determination and model complexity for a particular purpose. However, a similar score has been used since 1992 without any obvious problems (20), and a version of SPACI has been widely used in the SCOP database since 1994.

SPACI scores are valid *only* for crystal structures for which all of the parameters from PROCHECK and WHATCHECK are available.

PDB SEQUENCE MAPS

To build homology models and make other use of protein structures in analyzing sequence, it is necessary to relate coordinate information to the residue sequences. One of the difficulties of using protein structure data is that the identification of residues in PDB files follows no uniform simple set of rules. Numbering can have insertions, gaps, and need not even be monotonic. In addition, historical PDB files are non-compliant with the standard file format in a variety of ways. A further complication is that not all atoms in the molecule may be visible to the experimentalist determining the structure. So, the SEQRES records portion of the PDB file that contains the sequence of the molecule being studied may not bear a straightforward relationship with the ATOM records that provide three-dimensional coordinates for each atom visible to the experimentalist. The macromolecular Crystallographic Information Format (mmCIF) (21) attempts to remedy these problems by providing a linear sequence corresponding to the molecule being studied and it maps the atoms observed onto the linear sequence.

The Research Collaboratory for Structural Bioinformatics, which operates the PDB, has produced a program called pdb2cif which attempts to read PDB-format files and produce CIF-formatted files from these (14). Because of errors in file formats, pdb2cif does not succeed in producing a correct output in all cases, but we have found that it produces the correct alignment for >99% of PDB files and the program is well-maintained. The ASTRAL resource provides files extracted from the CIF files produced by pdb2cif which summarize the alignment between the SEQRES and ATOM records.

THE SCOP DOMAIN SEQUENCES

Because domains are the fundamental units of protein structure and evolution, the SCOP database, in common with most other structure databases, divides all proteins into their constituent domains for classification. The sequences of domains are useful for matching structures with sequences (22–25). They have also been used extensively in evaluations of sequence comparison (25–29), and in understanding work in structural biology (2,30). However, because of the vagaries of PDB files, extracting sequences for these domains, based on the residue numbers, is not always trivial. The ASTRAL database provides two types of sequences for domains of proteins in the SCOP database. One provides residue sequences corresponding to the ATOM records (for each residue for which atoms are located) within the range specified for the SCOP domain. The second, and usually preferred, set of domain sequences are produced according to the SEQRES records (for all residues in the molecule as experimentally studied), with boundaries determined using the PDB sequence maps from pdb2cif.

Most SCOP domains correspond to a single contiguous region of sequence. In these cases, the ASTRAL identifier for the sequence is the same as the SCOP domain identifier. When the SCOP domain is discontinuous in sequence, but all in one chain, then the several segments of sequence are concatenated, with an ‘X’ character separating the segments. Some SCOP domains span multiple chains. In these cases, a separate ASTRAL sequence entry is made for each chain. The identifiers for these sequences are produced by replacing the initial ‘d’ (‘domain’) of the SCOP identifier with an ‘e’ (‘element’) and appending the chain identifier.

The ASTRAL sequence files contain only domains classified in classes 1–7 of the SCOP database, and thus excludes peptides and fragments, designed proteins and non-proteins. Furthermore, it also excludes sequences which are less than 20 residues in length and sequences for which >20% of the residues are unknown or not identified. In a small number of cases, we have found that pdb2cif is unable to produce a valid mapping between the ATOM and SEQRES files, leading to incorrect PDB sequence maps. In these cases, the SEQRES-based sequences have some error, and a file summarizing known problems is available.

SEQUENCE SUBSET SELECTIONS

Because the PDB has large numbers of similar structures, selected subsets of PDB sequences have been produced to remove undesired redundancy (31,32). These subsets can sample all of the different structures in the PDB with only a fraction of the entries, and they remove bias due to over-represented structures. Subsets provided by ASTRAL database are special because they are based on SCOP domains and they explicitly incorporate structure quality at each step of the selection (rather than using a blanket threshold). At present, users of ASTRAL wishing to use subsets may select among one selection mechanism, three similarity criteria, and a wide variety of similarity threshold cutoffs. The current selection mechanism called ‘greedy SPACI’ suggested by Tim Hubbard (33), uses the algorithm described in (28) with SPACI scores used to rank structures.

The three similarity criteria presently available in ASTRAL are ‘BLAST identity in both’ (BIB), ‘*E*-value in 100,000,000 residues’ (E100M) and ‘SCOP classification’ (SC). Cutoffs for BIB range from 10 to 95% identity in a BLAST (34) gapped alignment, as a fraction of the average length of the sequences. In addition, a 100% BIB identity cutoff removes only domains entirely identical in sequence to others in the selected subset.

Because percentage identity is a poor measure of sequence similarity (28), an *E*-value based similarity measure is also available. A complication is that *E*-values depend upon database size. This means that *E*-values from the 'all by all' comparison will underestimate the significance of pairs in a smaller subset database. For this reason, we use a similarity measure based on the *E*-value of the matches, where score is based on a database of 100 000 000 residues [roughly the size of SWISS-PROT and TrEMBL (35) today, and about 50 times larger than the current complete ASTRAL SEQRES database]. When the *E*-values for a pair of sequences differ depending upon which was used as a query, the lower (more significant) *E*-value is used for the pair. Thresholds cutoffs range from $E = 10$ to $E = 10^{-50}$.

The SCOP classification similarity criterion uses the SCOP hierarchy to assess similarity between protein domains, and results in only the highest SPACI-scoring domain from a given level being selected. The threshold cutoffs available for this criterion are Class (CL), Fold (CF), Superfamily (SF), Family (FA) and Protein & Species (SP). So, for example, the superfamily-threshold subset would contain a single representative of each superfamily in SCOP, selected to have the highest SPACI score.

STRUCTURAL ALIGNMENT OF SCOP DOMAIN STRUCTURES

The SCOP database is based on using the sequence and structural similarity of proteins of known three-dimensional structure. While sequence comparisons in SCOP are done automatically, structural comparisons rely on the expert knowledge of Alexey Murzin. We have used our method of structure alignment, STRUCTAL (15), and compared it with the manual gold-standard (36). Our automatic comparisons have two advantages: (i) they give a statistical score that measures the probability that the observed match could have arisen by chance and (ii) they include a detailed alignment in which residues of the compared proteins are aligned on the basis of structure, not sequence. By explicitly providing the alignment, ASTRAL allows users to see how different proteins compare at the detailed structural level.

ACKNOWLEDGEMENTS

S.E.B. is an Alfred P. Sloan and US Department of Energy Postdoctoral Fellow in Computational Molecular Biology. This work is funded by DOE grant DE-FG-03-95ER62135.

REFERENCES

1. Orengo,C.A., Jones,D.T. and Thornton,J.M. (1994) *Nature*, **372**, 631–634.
2. Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1997) *Curr. Opin. Struct. Biol.*, **7**, 369–376.
3. Holm,L. and Sander,C. (1996) *Science*, **273**, 595–603.
4. Teichmann,S.A., Chothia,C. and Gerstein,M. (1999) *Curr. Opin. Struct. Biol.*, **9**, 390–399.
5. Kim,S.H. (1998) *Nature Struct. Biol.*, **5**, 643–645.
6. Sali,A. (1998) *Nature Struct. Biol.*, **5**, 1029–1032.
7. Abola,E.E., Bernstein,F.C., Bryant,S.H., Koetzle,T.F. and Weng,J. (1987) In Allen,F.H., Bergerhoff,G., and Sievers,R. (eds), *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*. Data Comission of the International Union of Crystallography, Cambridge, UK, pp. 107–132.
8. Holm,L. and Sander,C. (1998) *Nucleic Acids Res.*, **26**, 316–319.
9. Orengo,C.A., Pearl,F.M., Bray,J.E., Todd,A.E., Martin,A.C., Lo Conte,L. and Thornton,J.M. (1999) *Nucleic Acids Res.*, **27**, 275–279. Updated article in this issue: *Nucleic Acids Res.* (2000) **28**, 000–000.
10. Siddiqui,A.S. and Barton,G.J. (1995) *Protein Sci.*, **4**, 872–884.
11. Sowdhamini,R., Rufino,S.D. and Blundell,T.L. (1996) *Fold Des.*, **1**, 209–220.
12. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
13. Marchler-Bauer,A., Addess,K.J., Chappey,C., Geer,L., Madej,T., Matsuo,Y., Wang,Y. and Bryant,S.H. (1999) *Nucleic Acids Res.*, **27**, 240–243. Updated article in this issue: *Nucleic Acids Res.* (2000) **28**, 000–000.
14. Bernstein,H., Bernstein,F. and Bourne,P.E. (1998) *J. Appl. Crystallogr.*, **31**, 282–295.
15. Subbiah,S., Laurents,D.V. and Levitt,M. (1993) *Curr. Biol.*, **3**, 141–149.
16. Branden,C.I. and Jones,T.A. (1990) *Nature*, **343**, 687–689.
17. Hooft,R.W., Sander,C., Scharf,M. and Vriend,G. (1996) *Comput. Appl. Biosci.*, **12**, 525–529.
18. Morris,A.L., MacArthur,M.W., Hutchinson,E.G. and Thornton,J.M. (1992) *Proteins*, **12**, 345–364.
19. Brunger,A.T. (1992) *Nature*, **355**, 472–475.
20. Brenner,S.E. and Berry,A. (1994) *Protein Sci.*, **3**, 1871–1882.
21. Bourne,P., Berman,H.M., Watenpaugh,K., Westbrook,J. and Fitzgerald,P.M.D. (1997) *Methods Enzymol.*, **277**, 571–590.
22. Wolf,Y.I., Brenner,S.E., Bash,P.A. and Koonin,E.V. (1999) *Genome Res.*, **9**, 17–26.
23. Teichmann,S.A., Park,J. and Chothia,C. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 14658–14663.
24. Gerstein,M. and Levitt,M. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 11911–11916.
25. Gerstein,M. (1998) *Bioinformatics*, **14**, 707–714.
26. Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T.J.P. and Chothia,C. (1998) *J. Mol. Biol.*, **284**, 1201–1210.
27. Park,J., Teichmann,S.A., Hubbard,T. and Chothia,C. (1997) *J. Mol. Biol.*, **273**, 349–354.
28. Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
29. Levitt,M. and Gerstein,M. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 5913–5920.
30. Brenner,S.E. and Levitt,M. (1999) *Protein Sci.*, in press.
31. Heringa,J., Sommerfeldt,H., Higgins,D. and Argos,P. (1992) *Comput. Appl. Biosci.*, **8**, 599–600.
32. Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) *Protein Sci.*, **1**, 409–417.
33. Hubbard,T.J.P., Ailey,B., Brenner,S.E., Murzin,A.G. and Chothia,C. (1998) *Acta Crystallogr. D Biol. Crystallogr*, **54** [1 (Pt 6)], 1147–1154.
34. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
35. Bairoch,A. and Apweiler,R. (1999) *Nucleic Acids Res.*, **27**, 49–54. Updated article in this issue: *Nucleic Acids Res.* (2000) **28**, 000–000.
36. Gerstein,M. and Levitt,M. (1998) *Protein Sci.*, **7**, 445–456.