260–263 Nucleic Acids Research, 2002, Vol. 30, No. 1 

# ASTRAL compendium enhancements

**John-Marc Chandonia[1], Nigel S. Walker[2], Loredana Lo Conte[3], Patrice Koehl[4], Michael Levitt[4] and Steven E. Brenner[1,2,*]**

[1]Berkeley Structural Genomics Center, Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [2]Department of Plant and Microbial Biology, 111 Koshland Hall, University of California, Berkeley, CA 94720-3102, USA, [3]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK and [4]Department of Structural Biology, D-109 Fairchild, Stanford University, Stanford, CA, USA

Received September 18, 2001; Accepted September 20, 2001

## ABSTRACT

**The ASTRAL compendium provides several databases and tools to aid in the analysis of protein structures, particularly through the use of their sequences. It is partially derived from the SCOP database of protein domains, and it includes sequences for each domain as well as other resources useful for studying these sequences and domain structures. Several major improvements have been made to the ASTRAL compendium since its initial release 2 years ago. The number of protein domain sequences included has doubled from 15 190 to 30 867, and additional databases have been added. The Rapid Access Format (RAF) database contains manually curated mappings linking the biological amino acid sequences described in the SEQRES records of PDB entries to the amino acid sequences structurally observed (provided in the ATOM records) in a format designed for rapid access by automated tools. This information is used to derive sequences for protein domains in the SCOP database. In cases where a SCOP domain spans several protein chains, all of which can be traced back to a single genetic source, a 'genetic domain' sequence is created by concatenating the sequences of each chain in the order found in the original gene sequence. Both the original-style library of SCOP sequences and a new library including genetic domain sequences are available. Selected representative subsets of each of these libraries, based on multiple criteria and degrees of similarity, are also included. ASTRAL may be accessed at http://astral.stanford.edu/.**

## BACKGROUND

The Protein Data Bank (PDB) is a centralized repository of protein structures (1,2) containing over 13 000 entries in March 2001. The SCOP database (3,4) provides a manually curated set of domains from all PDB entries, classified in a hierarchy indicating different levels of structural and evolutionary relationships between the domains. SCOP thus provides a broad survey of all known protein folds, detailed information about relatives of proteins of known structure, and a framework for classification of additional structures as they are solved.

Many tools for bioinformatic analysis rely on sequence information, but the nature of PDB files makes it challenging to accurately extract the sequence corresponding to a given domain definition. ASTRAL addresses this issue by providing an explicit mapping between the PDB ATOM and SEQRES records, which is used to derive databases of sequences corresponding to the SCOP domains, as described in the original paper (5). Also available are subsets of selected representative domains created using different thresholds and measures of similarity. To choose the highest quality representatives for these subsets, Summary PDB ASTRAL Check Index (SPACI) scores are used to provide a first order guide to the resolution, *R*-factor and stereochemical accuracy of each crystallographically determined structure.

In the two years since ASTRAL was released, the number of domain sequences included has doubled from 15 190 to 30 867, and several additional databases that require manual curation have been included.

## RAPID ACCESS FORMAT (RAF) SEQUENCE MAPPINGS

The original release of ASTRAL included mappings between ATOM and SEQRES records in PDB chains represented in ASTRAL. These CIFMAP files were automatically generated from the output of the pdb2cif program (6), which is now provided for every entry in the PDB. Due to errors and inconsistencies in the underlying data (7), fully automated mappings produce errors. Therefore, the CIFMAP files often contained bugs, some of which were described in the notes accompanying each ASTRAL release. In the latest release, we have manually corrected bugs in the mappings for 638 protein chains, eliminating all known bugs caused by errors in the automated mapping. We did not attempt to correct errors occurring in the original PDB files; in cases of discrepancy between the PDB and the pdb2cif output, the original PDB file served as the final arbiter for making manual corrections. In the curation process, some errors in the PDB files were discovered. For 26 domains, additional residues appear in observed sequence (ATOM records), but were omitted from

*To whom correspondence should be addressed at: Department of Plant and Microbial Biology, 111 Koshland Hall, University of California, Berkeley, CA 94720-3102, USA. Tel: +1 510 643 9131; Fax: +1 208 279 8978; Email: brenner@compbio.berkeley.edu

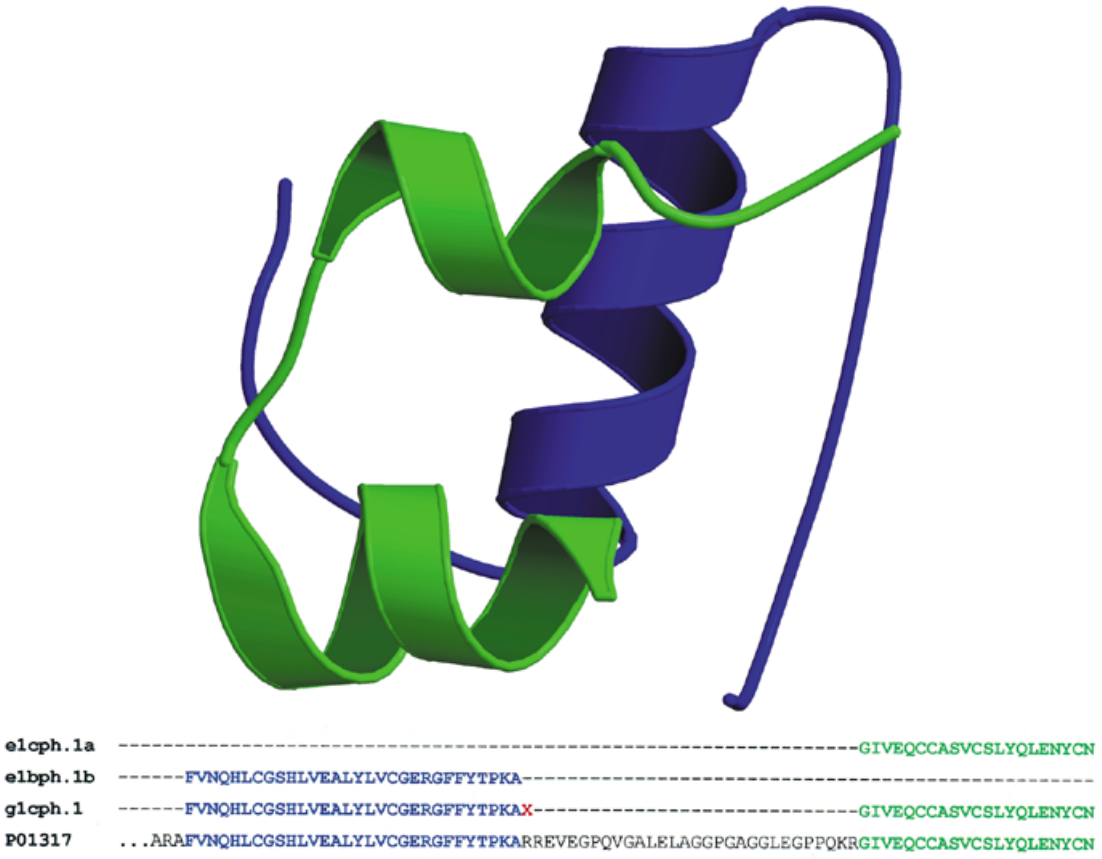


**Figure 1.** Genetic versus original-style domain sequences. d1cph1 is a multi-chain domain from the PDB file 1CPH, cow insulin (9). The a chain is shown in green; the b chain is shown in blue. In the original-style sequence sets, the domain is split into two sequences: e1cph.1a and e1cph.1b. In the genetic domain sequence sets, the sequences are concatenated into a single entry, g1cph.1, with the chain sequences in the correct order and separated by the letter 'X' (red). Sequence alignment of the three chains with a region of the insulin precursor protein [P01317 from SWISS-PROT (8)] is shown. The figure was created using RasMol (12), MOLSCRIPT (13) and Raster3D (14).

the record of the sequence being studied (SEQRES records). For an additional 83 domains, the sequences given in the SEQRES records and ATOM records do not match. While the differences are sometimes small (Asp/Asn), some are more significant (Glu/Pro). A file listing these differences is available on our web site.

As a robust alternative to the CIFMAP files, we now distribute the sequence mappings in RAF, a file format designed for rapid access in most computer languages. Details of the format are available on the web site. In addition to being the definitive set of maps used in SCOP and ASTRAL, the database of RAF maps is intended as a general purpose, manually curated resource for PDB users. It includes mappings for all PDB chains represented in the first seven classes in SCOP. The old format CIFMAP files are also available in the current release, but manual corrections are not included in these maps, and their use has been deprecated.

## GENETIC DOMAINS

A SCOP domain may include fragments from different PDB chains. In most cases, these appear to be the product of a single gene. For example, insulin and many proteases are products of post-translational cleavage of single precursor chains; each chain is given a different identifier in the PDB file. It is not trivial to reassemble the fragments in the order found in the original gene sequence, as the order in which the chains are presented in the PDB file often has no correlation with the original genetic order. In previous versions of SCOP, multiple chains in each of these 'genetic domains' have almost always been listed in the correct order in the SCOP domain definitions. In the current release, the correct order has been ensured through manual curation in collaboration with the SCOP authors. The order of chains in each genetic domain was determined by aligning the sequences against SWISS-PROT (8), followed by manual inspection. In the original-style ASTRAL sequences, there is a separate entry for each chain included in a SCOP domain, with the initial 'd' in the SCOP identifier replaced by an 'e'. Figure 1 shows the example of d1cph1, a multi-chain domain from a structure of cow insulin (9). This domain is split into the one sequence for each chain (a and b), e1cph.1a and e1cph.1b. In the genetic domain sequence sets, the sequences are concatenated into a single entry with the initial 'd' in the domain name replaced by a 'g'. In the insulin example, the d1cph.1 domain is given as a single sequence, g1cph.1, with the chain sequences in the correct order (b before a) and

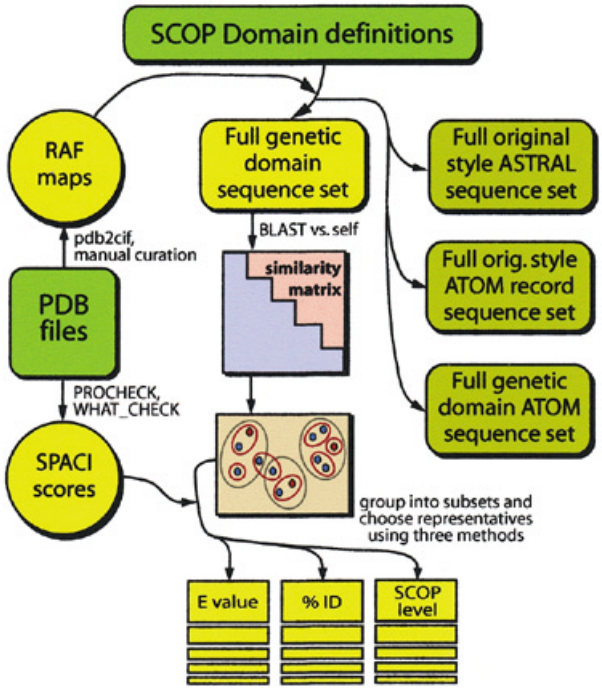


**Figure 2.** Data flow in ASTRAL. Primary data sources, the PDB and SCOP domain definitions, green; primary ASTRAL databases, light yellow; less commonly used resources, darker yellow. Using the RAF maps, four complete sequence sets are created for every domain in the first seven classes of the SCOP database. Two sets (the genetic domain sets) include the genetic domain sequences, and the other two (the original-style sequence sets) use the prior method of splitting each multi-chain domain into multiple sequences. For each of these methodologies, one complete sequence set is derived from sequences in the PDB ATOM records, and another from sequences in the SEQRES records. The SEQRES sets (for both genetic domain and original-style methods) are used to derive representative subsets. Each set is fully compared against itself using BLAST (10), and subsets are created using three similarity criteria and various thresholds. Representatives are chosen according to SPACI scores, a first order guide to the resolution, *R*-factor, and stereochemical accuracy of crystallographically determined structures (5).

separated by the letter 'X'. In the current version of ASTRAL, all fragments in a single domain were found to be products of a single gene. However, this may not always be the case, so future versions of ASTRAL may include 'd', 'e' and 'g' type domains in a single sequence set.

## REPRESENTATIVE SUBSETS

Since the majority of the domains in the PDB are very similar to others, it is helpful to reduce the redundancy by selecting high quality representatives at different levels of similarity. The process of selecting these representative subsets is shown in Figure 2.

Using the RAF maps, four complete sequence sets are created for every domain in the first seven classes of the SCOP database. Two sets (the genetic domain sets) include the genetic domain sequences described above, and the other two (the original-style sequence sets) use the prior method of splitting each multi-chain domain into multiple sequences. For each of these methodologies, one complete sequence set is derived from sequences in the PDB ATOM records, and another from sequences in the SEQRES records. We expect the genetic domain sets, in particular the one derived from SEQRES records, to be more commonly used than sets created using the prior methodology, so the original-style sets may be deprecated in future versions of ASTRAL.

The SEQRES sets (for both genetic domain and original-style methods) are used to derive representative subsets. As shown in Figure 2, each set is fully compared against itself using BLAST (10), and subsets are created using the three similarity criteria (BLAST *E*-values, sequence identity and SCOP classification) described previously (5). Representatives are chosen according to SPACI scores, which are derived from the resolution and *R*-factor of crystallographically determined structures, as well as the output of the programs PROCHECK (11) and WHAT_CHECK (7). Future SPACI scores may incorporate updated programs and additional source data (15). This selection procedure will be enhanced in future versions of ASTRAL to reflect mutated or misfolded structures.

The most frequently requested representative subsets are those filtered at a 40% level of sequence identity (ID) and a 95% ID level. These two subsets are highlighted on our web site. Since the original release of ASTRAL, the 40% ID subset has increased in size from 1947 domain sequences to 3613. The 95% ID subset has increased in size from 3285 sequences to 6146. Both of these represent an increase in the order of 85%, in contrast to the 103% increase in the total number of domains in the full sequence sets.

## NEW TRANSLATION TABLE

Chemically modified residues are now included in our translation table which maps the three-letter codes found in PDB files to one-letter codes in our sequences. The translation table for modified residues is given in Table 1. This information was extracted from the Het group dictionary given on the PDB site, although modified amino acids are not always distinguished in this file from other types of prosthetic groups. Several PDB files used HETATM codes for modified amino acids that were inconsistent with the standard table; however, in these cases the notation was explained in the PDB headers and the correct sequences were manually entered into the RAF maps.

## FORMAT CHANGES

In order to facilitate parsing of the sequence headers, the FASTA header for each ASTRAL sequence has been modified to always include a region code, identifying a residue range and chain ID. If the domain spans an entire PDB entry with no chain ID, the region code is represented as (–). The old classification page numbers in SCOP have been replaced by new SCOP Concise Classification String (sccs) identifiers (4), and this change is also reflected in the headers of FASTA files distributed in ASTRAL. Details on parsing the headers are available on our web site.

## ACKNOWLEDGEMENTS

We thank J. Michael Sauder for suggesting the translation of non-standard amino acids. This work is supported by grants from the NIH (1-P50-GM62412, 1-K22-HG00056 and GM1455),

**Table 1.** Translation table for chemically modified amino acids

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2as | d | 3ah | h | 5hp | e | acl | r | aib | a | alm | a | alo | t | aly | k | arm | r | asa | d |
| asb | d | ask | d | asl | d | asq | d | aya | a | bcs | c | bhd | d | bmt | t | bnn | a | buc | c |
| bug | l | c5c | c | c6c | c | ccs | c | cea | c | chg | a | cle | l | cme | c | csd | a | cso | c |
| csp | c | css | c | csw | c | cxm | m | cy1 | c | cy3 | c | cyg | c | cym | c | cyq | c | dah | f |
| dal | a | dar | r | das | d | dcy | c | dgl | e | dgn | q | dha | a | dhi | h | dil | i | div | v |
| dle | l | dly | k | dnp | a | dpn | f | dpr | p | dsn | s | dsp | d | dth | t | dtr | w | dty | y |
| dva | v | efc | c | fla | a | fme | m | ggl | e | glz | g | gma | e | gsc | g | hac | a | har | r |
| hic | h | hip | h | hmr | r | hpq | f | htr | w | hyp | p | iil | i | iyr | y | kcx | k | llp | k |
| lly | k | ltr | w | lym | k | lyz | k | maa | a | men | n | mhs | h | mis | s | mle | l | mpq | g |
| msa | g | mse | m | mva | v | nem | h | nep | h | nle | l | nln | l | nlp | l | nmc | g | oas | s |
| ocs | c | omt | m | paq | y | pca | e | pec | c | phi | f | phl | f | pr3 | c | prr | a | ptr | y |
| sac | s | sar | g | sch | c | scs | c | scy | c | sel | s | sep | s | set | s | shc | c | shr | k |
| soc | c | sty | y | sva | s | tih | a | tpl | w | tpo | t | tpq | a | trg | k | tro | w | tyb | y |
| tyq | y | tys | y | tyy | y | agm | r | gl3 | g | smc | c | asx | b | cgu | e | csx | c | | |

The one-letter code of the base amino acid is shown to the right of the three-letter code for the corresponding chemically modified amino acid.

the Department of Energy (DE-FG03-95ER62135) and the Searle Scholars Program (01-L-116).

## REFERENCES


1. 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.
2. 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. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 245–248.
3. 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.
4. Lo Conte,L., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
5. Brenner,S.E., Koehl,P. and Levitt,M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **28**, 254–256.
6. Bernstein,H., Bernstein,F. and Bourne,P.E. (1998) pdb2cif: translating PDB entries into mmCIF format. *J. Appl. Crystallog.*, **31**, 282–295.
7. Hooft,R.W.W., Vriend,G., Sander,C. and Abola,E.E. (1996) Errors in protein structures. *Nature*, **381**, 272.
8. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL. *Nucleic Acids Res.*, **28**, 45–48.
9. Gursky,O., Badger,J., Li,Y. and Caspar,D.L. (1992) Conformational changes in cubic insulin crystals in the pH range 7–11. *Biophys. J.*, **63**, 1210.
10. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
11. 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.
12. Sayle,R.A. and Milner-White,E.J. (1995) RasMol: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–376.
13. Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, **24**, 946–950.
14. Merritt,E.A. and Bacon,D.J. (1997) Raster3D: photorealistic molecular graphics. *Methods Enzymol.*, **277**, 505–524.
15. Hooft,R.W.W., Sander,C., Scharf,M. and Vriend,G. (1996) The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value. *Comput. Appl. Biosci.*, **12**, 525–529.