Nucleic Acids Research, 2004, Vol. 32, Database issue ***D189–D192***
DOI: 10.1093/nar/gkh034

The ASTRAL Compendium in 2004

John-Marc Chandonia[1], Gary Hon[2], Nigel S. Walker[3], Loredana Lo Conte[4], Patrice Koehl[5], Michael Levitt[5] and Steven E. Brenner[1,2,*]

[1]Berkeley Structural Genomics Center, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [2]Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA, [3]Institute of Molecular Biology, University of Oregon, Eugene, OR 97403, USA, [4]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK and [5]Department of Structural Biology, D-109 Fairchild, Stanford University, Stanford, CA 94305, USA

Received September 11, 2003; Revised and Accepted September 16, 2003

ABSTRACT

The ASTRAL Compendium provides several databases and tools to aid in the analysis of protein structures, particularly through the use of their sequences. Partially derived from the SCOP database of protein structure domains, it includes sequences for each domain and other resources useful for studying these sequences and domain structures. The current release of ASTRAL contains 54 745 domains, more than three times as many as the initial release 4 years ago. ASTRAL has undergone major transformations in the past 2 years. In addition to several complete updates each year, ASTRAL is now updated on a weekly basis with preliminary classifications of domains from newly released PDB structures. These classifications are available as a stand-alone database, as well as integrated into other ASTRAL databases such as representative subsets. To enhance the utility of ASTRAL to structural biologists, all SCOP domains are now made available as PDB-style coordinate files as well as sequences. In addition to sequences and representative subsets based on SCOP domains, sequences and subsets based on PDB chains are newly included in ASTRAL. Several search tools have been added to ASTRAL to facilitate retrieval of data by individual users and automated methods. ASTRAL may be accessed at http://astral.stanford.edu/.

BACKGROUND

The Protein Data Bank (PDB) is a centralized repository of protein structures (1) containing over 22 000 entries in August 2003. The SCOP database (2,3) provides a manually curated set of domains from all PDB entries, classified in a hierarchy indicating different levels of structural and evolutionary relationship 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 SCOP domains, as described previously (4,5). These Rapid Access Format (RAF) maps are manually curated to eliminate errors in automatic parsing of PDB files, and to translate chemically modified amino acids back to the original sequence. The RAF maps are used to derive databases of sequences corresponding to each domain and PDB chain included in SCOP. Representative subsets of these full sequence sets are also available, chosen according to different thresholds and measures of sequence similarity.

Recent improvements to ASTRAL include the creation of PDB-style coordinate files for each SCOP domain. Sequences are now provided for each PDB chain as well as for SCOP domains; representative subsets of PDB chains are also provided. The highest quality representative in each subset is now chosen using Aberrant Entry Re-Ordered SPACI (AEROSPACI) scores rather than the SPACI scores (4) used previously; PDB entries manually annotated by the SCOP authors as aberrant are penalized so that they are less likely to be chosen as the representative structure for a given subset. Genetic domain sequences for multi-chain SCOP domains, introduced in a previous release of ASTRAL (5), are now the default. Residues appearing in PDB files which have been chemically modified after translation are replaced by the original sequence where possible in both the RAF maps and ASTRAL sequences. Many of these replacements are done automatically using the table reported previously (5); others are extracted using manual or automated curation from comments in the PDB file.

Although several complete releases of ASTRAL are produced each year, synchronized to new SCOP releases, the number of new protein structures that become available between releases of SCOP continues to increase. For example, an additional 1646 proteins (5247 domains) were added in the 5 months between the release of ASTRAL 1.63 and the current

*To whom correspondence should be addressed at Department of Plant and Microbial Biology, 461A 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

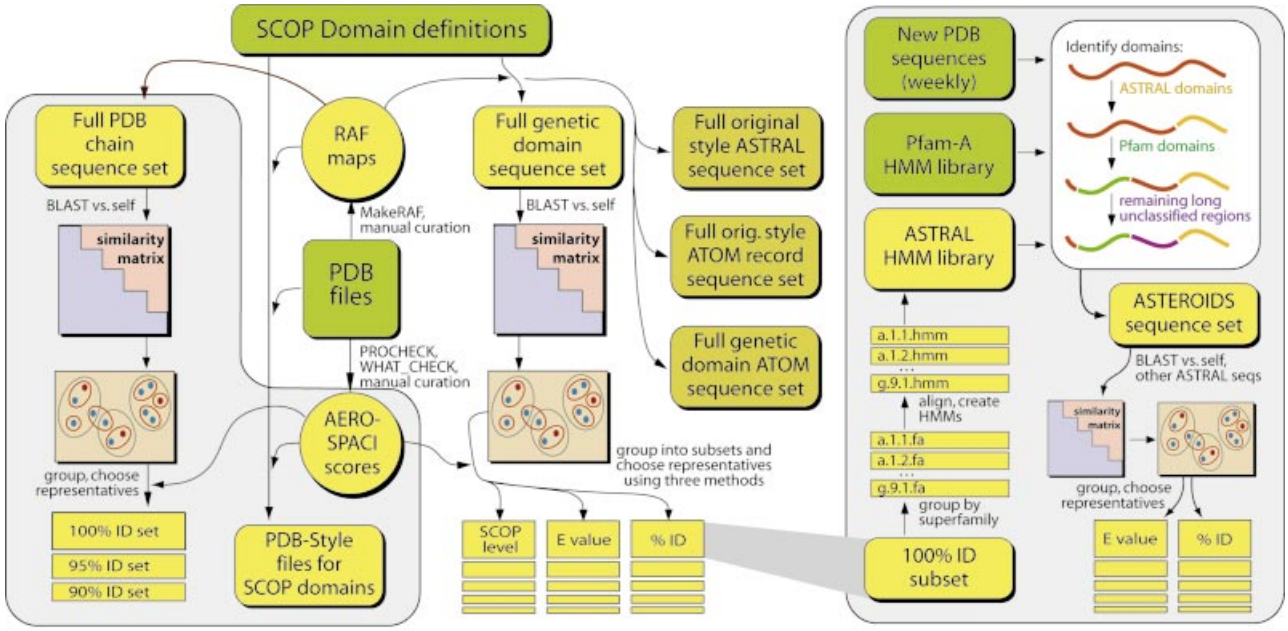

Figure 1. Data flow in ASTRAL. Primary data sources are shown in green. Primary ASTRAL databases are shown in light yellow. Less commonly used resources are shown in darker yellow. Resources added recently are outlined in light blue. Using the RAF maps, four complete sequence sets are created for every 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. 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, and subsets are created using three similarity criteria and various thresholds. Representatives are chosen according to AEROSPACI scores, described in the text. PDB chain sequence sets are derived from the SEQRES records of every PDB chain in SCOP; selected subsets are created at 90–100% ID thresholds. PDB-style files are derived from the RAF maps and SCOP domain definitions. At each new release of ASTRAL, all non-redundant sequences from each SCOP superfamily are aligned using MAFFT (10). A hidden Markov model (7) (HMM) is created from the multiple sequence alignment for each superfamily using HMMER (6). These HMMs are used to predict domains in the sequences of newly released PDB entries on a weekly basis. HMMs from the Pfam-A database are also used to predict domains in regions of the sequences not identified by HMMs derived from SCOP superfamilies. Unassigned regions of at least 50 consecutive residues are also predicted to be potential domains. The predicted domains (ASTEROIDS) are available in a single file, as well as optionally available integrated into representative subsets selected according to two similarity criteria (BLAST E-value and % identity) at various thresholds.

version, 1.65, compared with 1540 proteins (5170 domains) added in the previous 6 months since version 1.61. As the new data represent ~10% of the total number of domains in ASTRAL, it is important to rapidly incorporate these new structures. A major new feature in ASTRAL is integration of preliminary domain classifications of newly released structures using hidden Markov models (6,7) trained on superfamilies of previously classified domains.

CURATED MAPPINGS

The RAF maps (5) provide explicit mappings between the sequence of PDB chains studied (the SEQRES records) and the experimentally observed atoms (the ATOM records) for every PDB chain in SCOP. Manual curation ensures that the mapping presented in the RAF file is an exact representation of the data in the original PDB file, even when the PDB file itself is erroneous. In cases where residues have been post-translationally modified, efforts are made to represent the original sequence in the RAF maps. Many standard chemical modifications are translated automatically, as described previously (5). A great amount of additional manual and automatic curation has been added in recent ASTRAL releases. Hundreds of additional translations are parsed from comments in SEQADV records, in cases where a residue is annotated as 'modified'. Several thousand more translations are manually curated from comments in the PDB files that indicate which amino acid was chemically modified to derive a non-standard heterogen. In some cases, a single heterogen is derived from multiple amino acids, e.g. the chromophores of luminescent proteins which are cyclizations of three adjacent amino acids; these are mapped to multiple residues in the RAF maps and sequences. All non-standard residue translations are documented on our website in a format that is easily parsed by humans or automated methods. The RAF format is designed to be rapidly accessed in various computer languages, and we will soon release open source Perl modules to facilitate development of software that interacts with the RAF database.

ASTRAL SCOP REPRESENTATIVE SUBSETS

An overview of the ASTRAL build process is shown in Figure 1. 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 previously (5), and the other two (the original-style sequence sets) use the prior method of splitting each multi-chain domain into multiple sequences (4). 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. Genetic domain sequence sets mapped from SEQRES records are now the default ASTRAL sequences.

The SEQRES sets (for both genetic domain and original-style methods) are used to derive representative subsets. As shown in Figure 1, each set is fully compared against itself using BLAST (8), and subsets are created using the three similarity criteria (BLAST E-values, sequence identity and SCOP classification) described previously (4). Representatives are chosen according to AEROSPACI scores, which are derived from calculated SPACI scores and manual annotation by SCOP authors. SPACI scores, a first order guide to the resolution, *R*-factor and stereochemical accuracy of crystallographically determined structures, have been described previously (4). AEROSPACI scores add an additional penalty of –2.0 to structures annotated as chimeric, circularly permuted, disordered, missing large regions, erroneous, misfolded, mistraced, mutant or truncated. Theoretical structures, which are not present in SCOP but still assigned AEROSPACI scores, are assigned an additional penalty of –5.0.

PDB CHAIN SEQUENCE SETS

A set of sequences is created which includes the sequence of every PDB chain in SCOP, based on SEQRES records. Selected subsets are also derived from this set using the same method as used to derive SCOP domain subsets. Because PDB chains often contain multiple domains, we create subsets only at high sequence identity (90–100% ID); lower thresholds would produce incorrect results in cases where several multi-domain proteins share a single common domain.

CONTINUOUS UPDATES

ASTEROIDS (ASTral newER pOtentIal Domain Set) is a set of sequences of newly released PDB entries, divided into domains and optionally available integrated into the ASTRAL representative subsets. Because new PDB files are available each week, ASTEROIDS are created using a fully automated method for predicting domains similar to those already classified in the manually curated databases SCOP and Pfam (9).

An overview of the ASTEROIDS build process is shown on the right side of Figure 1. The 100% ID representative subset of genetic domain sequences is grouped by superfamily. Sequences from each superfamily are aligned with MAFFT (10), using the fftnsi algorithm and all default options. A hidden Markov model (7) (HMM) is created from the multiple sequence alignment for each superfamily using the HMMER (6) tools hmmbuild and hmmcalibrate (with all default options). These HMMs are built once during a full release of ASTRAL, and then used to predict domains in the sequences of newly released PDB entries on a weekly basis, using an E-value cutoff of 10^{-4}. HMMs from the Pfam-A database (9) are also used to predict domains in the remaining unassigned regions of sequence; in these cases, an E-value equal to the 'trusted cutoff' or 10^{-4}, whichever is more significant, is used to assign domain predictions. Overlaps of up to 10 residues between multiple HMMs are allowed, and regions of sequence matching several domains are assigned to the one with the more significant E-value. Longer overlaps result in the entire domain prediction being rejected, and automatically flagged for later manual review to prevent further erroneous predictions. After domain assignment using HMMs, any unassigned region of at least 50 consecutive residues is also predicted to contain at least one potential domain, and included in the ASTEROIDS set. All ASTEROIDS are assigned SCOP sid-style identifiers (3) beginning with the letter 'u'; for example, u1abcd1 would be the first of several predicted domains in chain D of the PDB entry 1ABC. The FASTA headers for the ASTEROID sequences indicate the chain and region boundaries, the source of the domain prediction (ASTRAL superfamily, Pfam or remaining unassigned region), and the version of the database and E-value of the prediction for domains identified using HMMER. ASTEROID sequences are integrated into representative ASTRAL subsets selected according to two similarity criteria (BLAST E-value and % identity) at a variety of thresholds using previously described methods for creating representative subsets (4). ASTEROID sequences are assigned AEROSPACI scores of –9.99; other PDB chains have AEROSPACI scores ranging from –5.9 to 1.91, so an ASTEROID is only chosen as a structural representative if no similar sequence already classified in SCOP is available. Representative subsets of SCOP/ASTRAL domains are available with or without ASTEROIDS, and all ASTEROIDS may also be downloaded in a single FASTA file. Multiple alignments and HMMs for ASTRAL superfamilies are also available.

PDB-STYLE FILES

To facilitate use of ASTRAL by structural biologists, we provide PDB-style files containing coordinates for each SCOP domain. These files also contain REMARK records documenting the original PDB file used as a data source, as well as information on the domain's classification in ASTRAL and SCOP, such as identifiers and AEROSPACI scores.

IMPROVED SCRIPTS

CGI scripts are now provided which retrieve individual sequences and PDB-style files. Both genetic domain and original-style sequences can be retrieved, as well as sequences derived from either SEQRES or ATOM records. Data may be searched using a variety of identifiers, including PDB codes, SCOP sid identifiers and SCOP sccs identifiers (3).

ACKNOWLEDGEMENTS

This work is supported by grants from the NIH (1-P50-GM62412, 1-K22-HG00056) and the Searle Scholars Program (01-L-116), and by the US Department of Energy under contract DE-AC03-76SF00098.

REFERENCES

1. 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.
2. 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.

3. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP Database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
4. Brenner,S.E., Koehl,P. and Levitt,M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **28**, 254–256.
5. Chandonia,J.M., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2002) ASTRAL compendium enhancements. *Nucleic Acids Res.*, **30**, 260–263.
6. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
7. Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.
8. 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.
9. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
10. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.