BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 4 2005, pages 551–553
doi:10.1093/bioinformatics/bti049

Seq2Struct: a resource for establishing sequence-structure links

Allegra Via, Andreas Zanzoni and Manuela Helmer-Citterich*

Department of Biology, Centre for Molecular Bioinformatics, University of Rome Tor Vergata, 00133 Rome, Italy

Received on May 18, 2004; revised and accepted on September 22, 2004
Advance Access publication September 28, 2004

ABSTRACT

Summary: Several methods for establishing cross-links between Protein Data Bank (PDB) structures or Structural Classification of Proteins (SCOP) domains and Swiss-Prot + TrEMBL sequences (or vice versa) rely on database annotations. Alternatively, sequence alignment procedures can be used. In this study, we describe Seq2Struct, a web resource for the identification of sequence-structure links. The resource consists of an exhaustive collection of annotated links between Swiss-Prot + TrEMBL and PDB + SCOP database entries. Links are based on pre-established highly reliable thresholds and stored in a relational database, which has been enhanced using annotations derived from Swiss-Prot, PDB, SCOP, GOA and DSSP databases. The Seq2Struct database contents, supported by a WWW web interface, can be queried both online and downloaded.

Availability: The Seq2Struct resource, with related documentation, is available at http://surface.bio.uniroma2.it/seq2struct/

Contact: seq2struct@cbm.bio.uniroma2.it

INTRODUCTION

Comprehensiveness and reliability of links between Swiss-Prot + TrEMBL (Boeckmann *et al*., 2003) sequences and the Protein Data Bank (PDB) structures (Berman *et al*., 2000) or Structural Classification of Proteins (SCOP) domains (Murzin *et al*., 1995) are classical problems in both molecular and computational biology. Several methods for establishing sequence-structure correspondences, such as those provided by SRS (http://srs.embl-heidelberg.de:8000/srs5/) or EBI MSD (http://www.ebi.ac.uk/msd/) resources, rely on database annotations. Full annotation of sequences (e.g. TrEMBL) and structures (e.g. PDB), however, is often sparse and not always reliable. Cross-references with SCOP are not available in Swiss-Prot + TrEMBL and are mediated by the PDB entries in MSD.

Alternatively, programs like BLAST at NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) for sequence similarity searches can be used. BLAST accepts a number of different types of input: a sequence in FASTA format, a bare sequence or a number of sequence identifiers (e.g. NCBI gi, Swiss-Prot + TrEMBL identification number, ID, or accession number, AC). A single sequence can be searched in several databases, such as Swiss-Prot, GenBank CDS translations (Benson *et al*., 2004), PIR (Wu *et al*., 2003) and PDB sequences. In the case of PDB, protein sequences are extracted from the SEQRES lines of the PDB files. The entire SCOP database cannot be searched and the structure codes are not accepted as input.

Concerning sequence-structure cross-links, the resources described above are not complete in number and lack potentially interesting information, such as the quality of the alignment associated with the cross-reference.

METHODS AND RESULTS

In this note, we describe a user-friendly web resource consisting of an exhaustive database of annotated links between Swiss-Prot + TrEMBL entries and either PDB chains or SCOP domains, together with the tools required for querying it. The resource provides the known cross-references of a query, plus a list of annotated supplementary links including all proteins sharing a significant sequence identity with the query. A sequence-structure pair is identified by means of a sequence alignment procedure together with a set of boundary conditions imposed on the alignments. The collection of sequence-structure links is then enhanced with data extracted from Swiss-Prot, PDB, SCOP, GOA (Camon *et al*., 2004; Harris *et al*., 2004) and DSSP (Kabsch and Sander, 1983) databases. In the case of structures, protein sequences are extracted from the atom coordinates. SCOP coordinate files are retrieved from the ASTRAL database (Chandonia *et al*., 2004). The database similarity search program BLASTP (Altschul *et al*., 1997) is used to search all the structures against the complete Swiss-Prot + TrEMBL database. For each structure, hits are retained if they display at least 92% sequence identity in non-gapped regions, at least 25 residues are aligned, no more than 15% of the gaps are present in the alignment, and the E-value is $<10^{-6}$. The procedure made it possible to establish a large number of new sequence-structure

*To whom correspondence should be addressed.

Table 1. Number of links as cross-referenced by Swiss-Prot + TrEMBL and as provided by Seq2Struct

Links[a]	Sptr cross-referenced[b]	Seq2Struct Total[c]	100 id[d]	98 ≤ id < 100[d]	96 ≤ id < 98[d]	92 ≤ id < 96[d]
Swiss-Prot/PDB	22 660	157 561	36 464	30 344	28 546	62 207
TrEMBL/PDB	1451	324 624	31 201	85 869	96 279	111 275
Swiss-Prot/SCOP	0	161 200	39 388	29 235	28 786	63 791
TrEMBL/SCOP	0	369 840	48 245	88 767	107 907	124 921

Data refer to the Swiss-Prot release 44.0, TrEMBL 27.0 and SCOP release 1.65. PDB data refer to July 7, 2004.
[a]Pairs of databases for which links have been established.
[b]Number of links established by the cross-references as provided by Swiss-Prot and TrEMBL.
[c]Total number of links provided by the Seq2Struct resource.
[d]Number of links provided by the Seq2Struct resource within a range of sequence identity (percentage).

links: those between sequences and SCOP domains (not directly referenced in other databases at the moment), plus a relatively high number of missing links between the Swiss-Prot + TrEMBL databases and the PDB (Table 1). Information is collected in the two text files: sptr_pdb.txt (PDB chains/Swiss-Prot + TrEMBL pairs) and sptr_scop.txt (SCOP/Swiss-Prot + TrEMBL pairs). Each file, for each structure-sequence link, reports, among other information (see the documentation), the Swiss-Prot + TrEMBL sequence range of aligned residues and a quality score (QS). The QS is a number of '*' symbols between one and four and is correlated to sequence identity (if there are gaps, the sequence identity 'outside gaps' is considered instead). Four '*' (i.e. '****') are assigned to hits displaying 100% sequence identity. Other QS values are described in detail in the Seq2Struct documentation. This information is stored in a PostgreSQL relational database (http://www.postgresql.org/). The query forms are written in Python programming language (http://www.python.org/). Supplementary annotation (such as protein description, cross-references, GO terms and structure determination experimental method) is integrated into the Seq2Struct database. Secondary structure and solvent accessibility information from the DSSP database are also available in the residue-by-residue mapping table (see below). The user can query the database by submitting either a structure (PDB or SCOP) code or a Swiss-Prot + TrEMBL accession code and specifying a target database.

The core output consists of a list of sequence-structure links corresponding to the query protein submitted. For each link, two different output web pages are provided: the user can choose between the 'simple mode' view where the information of all the sequence-structure pairs corresponding to a single query is displayed in a unique table, and the 'detailed mode' view where the information of each single sequence-structure pair is displayed in a separate table. The 'detailed mode' also provides links to GO annotations, to the sequence alignment ('alignment' button) and to a table containing the sequence alignment in the form of a residue-by-residue specific mapping ('residue mapping' button). In addition, this table displays secondary structure and solvent accessibility information. The output, the sptr_pdb.txt and the sptr_scop.txt text files can be downloaded as tab-delimited text files, which are easily parseable with computer programs. The user can choose to sort the output by using one or more parameters, such as QS (default), *E*-value, sequence identity, sequence similarity, percentage of query's aligned residues and the percentage of target aligned residues.

Compared with other databases, the Seq2Struct resource allows the identification of many links besides those that are detected by cross-references. This is particularly relevant in the case of TrEMBL sequences, for which annotation is sparse or absent. Moreover, Seq2Struct provides direct correspondences between Swiss-Prot + TrEMBL and Astral SCOP entries, including the SCOP domain annotations. Such correspondences, as provided by MSD, are mediated by the PDB entries corresponding to SCOP IDs.

The Seq2Struct database contents are updated monthly by using the most recent releases of Swiss-Prot + TrEMBL, PDB and Astral SCOP databases. Supplementary updates will be provided following the major releases of the databases. A more detailed description of the resource, its motivation and some examples clarifying the improvements of Seq2Struct with respect to already existing web resources can be found on the documentation page.

ACKNOWLEDGEMENTS

We thank Teresa Colombo and Arnaud Ceol for providing useful data. We acknowledge the support from Telethon, FIRB 'Bioinformatica per la Genomica e la Proteomica' and GENEFUN.

REFERENCES

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.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res.*, **32** (Database issue), D23–D26.

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.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32** (Database issue), D262–D266.

Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32** (Database issue), D189–D192.

Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32** (Database issue), D258–D261.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

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.

Wu,C.H., Yeh,L.L., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Ledley,R.S., Kourtesis,P., Suzek,B.E., Vinayaka,C.R., Zhang,J. and Barker,W.C. (2003) The Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347.