The HSSP database of protein structure–sequence alignments and family profiles

Chris Dodge[1], Reinhard Schneider[2,+] and Chris Sander[1,2,*]

[1]European Bioinformatics Institute, EMBL-EBI, Genome Campus, Cambridge CB10 1SD, UK and
[2]European Molecular Biology Laboratory, EMBL-HD, D-69012 Heidelberg, Germany

Received October 17, 1997; Accepted October 17, 1997

ABSTRACT

HSSP (http://www.sander.embl-ebi.ac.uk/hssp/) is a derived database merging structure (3-D) and sequence (1-D) information. For each protein of known 3D structure from the Protein Data Bank (PDB), we provide a multiple sequence alignment of putative homologues and a sequence profile characteristic of the protein family, centered on the known structure. The list of homologues is the result of an iterative database search in SWISS-PROT using a position-weighted dynamic programming method for sequence profile alignment (MaxHom). The database is updated frequently. The listed putative homologues are very likely to have the same 3D structure as the PDB protein to which they have been aligned. As a result, the database not only provides aligned sequence families, but also implies secondary and tertiary structures covering 33% of all sequences in SWISS-PROT.

INTRODUCTION

HSSP (homology-derived structures of proteins) is a derived database merging information from three-dimensional structures and one-dimensional sequences of proteins. The added value in the database stems from the evolutionary observation that protein sequences can vary considerably while maintaining the same overall 3-D structure. One can therefore group sequence-similar proteins into families of structural homologues. If the 3-D structure of only one family member is known, then by implication one can derive the basic 3-D structure, or fold, of all family members.

To exploit this principle, we align, for each protein of known 3-D structure in the Protein Data Bank (1), all its likely sequence homologues. As a result, HSSP is not only a database of aligned sequence families, but also a database of implied secondary and tertiary structures. Likely secondary structures can be carried over directly from the PDB protein to each homologue. Tertiary structure models can be built by fitting the sequence of the homologue, as aligned, into the 3-D template of the protein of known structure (sequence inserts, however, are very difficult to model in 3D).

Relative to the experimentally derived structural information in PDB, HSSP increases the number of effectively known protein structures several fold. The database is useful for analyzing residue conservation in structural context, for defining structurally meaningful sequence patterns and, in general, for studying protein evolution, folding and design.

CONTENT AND FORMAT OF THE DATABANK

For each protein in PDB, with identifier xxxx (such as: 1PPT, 5PCY), there is an ASCII (text) file xxxx.HSSP which contains: (i) the primary sequence of the protein of known structure, along with the derived secondary structure and solvent accessibility calculated from the coordinates using DSSP (2); (ii) aligned sequences of a few or tens or hundreds of sequences from the SWISS-PROT database (3) deemed structurally homologous to this protein; (iii) sequence variability, using two different measures, at each position in the multiple sequence alignment; and (iv) occupancy, i.e., the number of sequences that span this position. Alignments were produced using a modified Smith–Waterman dynamic programming algorithm, allowing gaps, and likely homologues were selected applying a well-tested threshold for structural homology. Some details of the methods are given elsewhere (4).

For example, the dataset 1PPT.HSSP (data not shown, but available on the Web site) contains more than 30 aligned sequences of pancreatic hormones, neuropetides Y and peptides YY from different species. Residue Y27 (Tyr) is in an α-helix (H), has a solvent accessibility of 56 Å^2 and has a variablity of 0, i.e., it is strictly conserved as Tyr in all sequences. The alignments could be used to build explicit 3-D models of each of the homologous sequences. Such models would be quite accurate in the core regions (helices and strands), but less accurate in loop regions. If the 3-D-structure of one of the aligned sequences is known experimentally, a pointer to that structure in PDB is given in the column STRID (structure identifier).

As there is considerable redundancy in the Protein Data Bank, i.e., several datasets in PDB represent the same structural family, the sequence families in HSSP overlap. For example, there are separate files for hemoglobin and myoblobin, which have ~30–35% identical residues, so that proteins homologous to both

*To whom correspondence should be addressed at: EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Tel: +44 1223 494 609; Fax: +44 1223 494 468; Email: sander@embl-ebi.ac.uk

+Present address: LION bioscience AG, Im Neuenheimer Feld 517, D-69120 Heidelberg, Germany hemoglobin and myoglobin appear in both files. Sequence-identical chains in the PDB entry are removed so that the xxxx.hssp files only contain sequence-unique chains.

DISTRIBUTION

Anonymous FTP

If you have access to Internet you can obtain HSSP by anonymous ftp (File Transfer Protocol) from ftp.embl-ebi.ac.uk in directory:/pub/databases/hssp or using a world wide web browser from ftp://ftp.embl-ebi.ac.uk/pub/databases/

World Wide Web (WWW)

The HSSP database and HSSP-related information and data are accessible from http://www.sander.embl-ebi.ac.uk/

The program (MaxHom) that generates the alignments is currently not available for distribution. Request for alignments based on structures not in the Protein Data Bank may be sent to C. Sander by email. Results will be mailed back, capacity permitting. Priority will be given to new 3-D structures.

Conditions

Academic redistribution of single files or of the entire database is permitted, provided that dataset integrity is strictly maintained. No inclusion in other databases or datasets, academic or other, without explicit permission of the authors. All commercial rights reserved. Not to be used for classified research. Users are asked to refer to this paper and ref. 4 in reporting results based on use of the database.

CONTENT AND SIZE OF THE CURRENT RELEASE

The content and size of the HSSP database is of course tightly coupled to the development of the databases of protein 3D structures (PDB) and sequences (e.g., SWISS-PROT). An overview of the increase in size is given in Table 1. Interestingly, ~20 000 out of 59 000 known sequences (SWISS-PROT release 34) are putative homologues of known structures and therefore have an implied known 3-D structure.

The complete set of data files currently requires ~600 Mb of disk storage. Updates of the database are done on a regular basis.

LIMITATIONS

Accuracy of reported alignments

In general, the alignments in HSSP are based almost entirely on sequence information and therefore may deviate from alignments based on comparison of known 3-D structures in local detail, especially in terms of placement of gaps. In these cases, the sequence alignment may correctly represent conservation in the evolutionary chain of events connecting the two sequences while structural alignment may reflect a local structural rearrangement as a result of mutations in sequence positions spatially near the conserved residues. Alignments, whether based on sequences or structures, are often uncertain in loop regions.

Definition of variability

In using variability scores, the user should be aware that low occupancy positions (few alignments span that position) have ill-determined variability values; in the limit of zero occupancy the variability is undefined and set to zero. For some purposes, the user may choose to use only positions with occupancy larger than, say, five proteins.

RELATED DATABASES AND INFORMATION SERVICES

The following databases and information services are also available from the Sander and Holm groups at EMBL-EBI, with network access provided by the same mechanisms as for HSSP (FTP and WWW access, see above).

Table 1.

HSSP release (month/year)	No. of HSSP data sets	No. of SWISS-PROT entries	Total no. of alignments in the HSSP database	No. of unique alignments and fraction of SWISS-PROT in the HSSP database
05/91	488	20 024	37 715	3065 (15.3%)
02/92	621	22 654	43 266	3498 (15.4%)
04/92	652	23 742	45 140	4556 (19.2%)
09/92	736	25 044	49 784	4825 (19.2%)
02/93	694	28 154	54 043	5370 (19.1%)
07/93	1361	29 955	104 837	7197 (24.0%)
10/93	1532	31 808	123 810	7642 (24.0%)
04/94	1959	36 000	148 175	9554 (26.5%)
08/94	2158	38 303	154 590	10 136 (26.5%)
08/95	3158	43 470	241 518	11 762 (27.0%)
09/96	4189	52 205	317 231	15 140 (29.0%)
10/97	5745	59 021	485 527	20 025 (33.9%)

DSSP, a database of secondary structure, solvent accessibility and other information derived from 3-D structures in the Protein Data Bank (2).
http://www.sander.embl-ebi.ac.uk/dssp/
personal email: sander@embl-ebi.ac.uk

FSSP, a database of protein structure families, based on 3-D alignments of protein structures, and a dictionary of structural domains (folding motifs).
http://www2.embl-ebi.ac.uk/dali/fssp/ and
http://www2.embl-ebi.ac.uk/dali/domain/
personal email: holm@embl-ebi.ac.uk

PDBselect, a representative subset of sequence-unique proteins of known 3-D structure selected from the Protein Data Bank (5).
http://www.sander.embl-heidelberg.de/pdbsel/
personal email: hobohm@embl-heidelberg.de

PredictProtein, an electronic mail server that provides a predicted secondary structure and solvent accessibility profile for any protein sequence with homologues in SWISS-PROT. Rated at 72% sustained three-state accuracy (6,7).
http://www.embl-heidelberg.de/predictprotein/
personal email: rost@embl-heidelberg.de

PropSearch, performs searches in sequence databases using amino acid composition and other non-sequential properties of a protein sequence as input (8).
http://www.sander.embl-heidelberg.de/propsearch/
personal email: hobohm@embl-heidelberg.de

GeneQuiz, results of automated protein sequence analysis for completely sequenced genomes.
http://www.sander.embl-ebi.ac.uk/genequiz/
personal email: genequiz@embl-ebi.ac.uk

GPCRDB, information system for G-protein coupled receptors.
http://swift.embl-heidelberg.de/7tm/
personal email: vriend@embl-heidelberg.de

Dali, an electronic mail server that performs a 3D similarity search in the Protein Data Bank, given the atomic coordindates of a 3D protein model as input (9).
http://www2.embl-ebi.ac.uk/dali/
personal email: holm@embl-ebi.ac.uk

Special software is available to construct 3-D models by homology based on the information in HSSP files, such as *WHATIF* by Gert Vriend (10) or *MaxSprout/Torso* by Liisa Holm and Chris Sander (11).

Report any problems with the HSSP database to the authors by electronic mail: sander@embl-ebi.ac.uk or dodge@embl-ebi.ac.uk.

REFERENCES

1 Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
2 Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
3 Bairoch,A. and Boeckmann,B. (1992) *Nucleic Acid Res.*, **20**, 2019–2022.
4 Sander,C. and Schneider,R. (1991) *Proteins*, **9**, 56–68.
5 Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) *Protein Sci.*, **3**, 409–417.
6 Rost,B., Schneider,R. and Sander,C. (1993) *Trends Biochem. Sci.*, **18**, 120–123.
7 Rost,B., Schneider,R. and Sander,C. (1994) *Comput. Applic. Biosci.*, **10**, 53–60.
8 Hobohm,U. and Sander,C. (1995) *J. Mol. Biol.*, **251**, 390–399.
9 Holm,L. and Sander,C. (1993) *J. Mol. Biol.*, **233**, 123–138.
10 Vriend,G. (1990) *J. Mol. Graphics*, **8**, 52–56.
11 Holm,L. and Sander,C. (1992) *Proteins*, **14**, 213–223.