ANTHONY ARTALE
MED LIB NATHAN CUMMINGS CTR (S-46)
MEMORIAL SLOAN KETTERING CANCER CTR
1275 YORK AVENUE
NEW YORK, NY 10021
UNITED STATES

Telephone: 212/639-7441
Fax: 646/422-2316

ORDER NUMBER:	CI-05875297-9
Account Number:	DD704731
Delivery Mode:	F31
Delivery Address:	arielsf.infotrieve.com/140.163.217.217
Reply Via:	SMT
Reply Address:	conyersd@mskcc.org
Submitted:	2005/11/23 11:51:03
Received:	2005/11/23 11:51:03
Printed:	2005/11/23 15:08:45

Direct	Periodical	OPENURLOPAC	UNITED STATES

CAL	**p**	Last 15 yrs. c. 1.
IRBM Ser	**QP620 N96**	v. 21-26; 1993-1998. c. 1.
Internet	**QP620 N96**	v. 1, no. 1- Jan. 1974- c. 1.
MAIN Ser	**QP620 N96**	v. 1- 1974- c. 1.
MBH Ser	**QP620 N96**	v. 15-27, no. 24 1987-1999. c. 1.
PBIS MformSer	**QP620 N96**	v. 11-12; 1983-84. c. 1.
PBIS Ser	**QP620 N96**	v. 13-27; 1985-1999. c. 1.

Nucleic acids research
10052963

Title:	**NUCLEIC ACIDS RESEARCH**
DB Ref. No.:	IRN10052963
ISSN:	ISSN03051048
Vol./Issue:	22
Date:	3500
Pages:	3597-3599
Article Title:	THE HSSP DATABASE OF PROTEIN SEQUENCE-STRUCTURE ALIGNMENTS
Article Author:	SCHNEIDER, R
Report Number:	IRN10052963
Publisher:	INFORMATION RETRIEVAL LTD.,
Client Number:	DDS36580/GANGI-DINO

INSTRUCTIONS: PATRON REQUESTS COLOUR IF AVAILABLE. THANK YOU.

Estimated cost for this 3 page document: $10.2 document supply fee + $25.5 copyright = $35.7

Nucleic Acids Research, 1994, Vol. 22, No. 17 3597–3599

The HSSP database of protein structure–sequence alignments

Chris Sander and Reinhard Schneider*
Protein Design Group, European Molecular Biology Labratory, Heidelberg, Germany

ABSTRACT

HSSP (homology-derived structures of proteins) is a derived database merging structural (2-D and 3-D) and sequence information (1-D). For each protein of known 3D structure from the Protein Data Bank, the database has a file with all sequence homologues, properly aligned to the PDB protein. 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 is not only a database of sequence aligned sequence families, but it is also a database of implied secondary and tertiary structures.

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 basic fact 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 of all family members.

To exploit this principle, we align, for each protein of known 3-D structure in the Brookhaven Protein Data Bank (PDB) [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 directly carried over 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.

Relative to the experimentally derived structural information in PDB, HSSP increases the number of effectively known protein structures severalfold. 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 (like: 1PPT, 5PCY), there is a ASCII (text) file xxxx.HSSP which contains 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], and, as the key merged information, aligned sequences of a few or tens or hundreds of sequences from the SWISS-PROT database [3] deemed structurally homologous to this protein. In addition, at each position in the multiple sequence alignment, sequence variability is indicated using two different measures, as well as the number of sequences that span this position (occupancy). 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. Details of the methods are given elsewhere [4].

For example, the dataset 1PPT.HSSP (figure 1) contains 27 aligned sequences of pancreatic hormones, neuropetides Y and peptides YY from different species. Residue Y27 (Tyr) is in an alpha-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 about 30%–35% identical residues, so that proteins homologous to both hemoglobin and myoglobin appear in both files. To avoid redundant information sequence identical chains in the PDB entry are removed. The xxxx.hssp files only contain sequence-unique chains.

DISTRIBUTION

CD-ROM

A subset of the HSSP database, one file for each protein in a representative set of proteins, is distributed on CD-ROM by the EMBL Data Library. In this representative set of proteins selected from PDB, sequence similarity between any two proteins does not exceed 25% identical residues (over a length of 80 or more residues). For detailed information on how the representative set was generated see [5] and the documentation distributed with the

*To whom correspondence should be addressed

database. For enquiries regarding the distribution of HSSP on this medium contact:

EMBL Data Library
European Molecular Biology Laboratory
EBI Outstation
Hinxton, CB10 1RQ, UK
Telephone: +44 (1223) 494400
Telefax: +44 (1223) 494468
Network: Datalib@ebi.ac.uk

HEADERS block: the first four bytes in the file, 'HSSP', can be used for file type detection. The first line also has the version number of the HSSP software (program MaxHom). The PDBID (protein data bank identifier) line identifies the test protein of known structure (e.g. 1PPT), the SEQBASE-line specifies the source of the aligned sequences (e.g. EMBL/Swissprot or PIR/NBRF). The PARAMETER line specifies alignment parameters used in the alignment program. The THRESHOLD line refers to the homology threshold curve used. Information about the test protein as copied from PDB (name, source, author) and as derived (length of the sequence SEQLENGTH, number of distinct chains NCHAIN, and the number of aligned sequences NALIGN).

PROTEINS block: pair alignment data for each of the proteins deemed structurally homologous to the test protein, where the word pair alignment refers to the alignment of the test protein with the single homologous protein

ID	EMBL/SWISSPROT identifier of the aligned (homologous) protein
STRID	if the 3-D structure of this protein is known, then STRID (structure ID) is the Protein Data Bank identifier as taken from the database reference line or DR-line (latest date) of the EMBL/SWISSPROT entry
%IDE	percentage of residue identity of the alignment.
IFIR/ILAS	first and last residue position of the alignment in the test protein
JFIR/JLAS	first and last residue position of the alignment in the aligned protein.
LALI	length of the alignment excluding insertions and deletions.
NGAP	number of insertions and deletions in the alignment.
LGAP	total length of all insertions and deletions
LSEQ2	length of the entire sequence of the aligned protein
ACCNUM	SwissProt accession number.
PROTEIN	one-line description of aligned protein.

ALIGNMENTS block: residue-by-residue details of the family alignment. From left to right in one line: sequence and structure information for one position in the test protein taken from the corresponding DSSP file [2]; sequence variability for this position followed by the aligned sequences in the same order as in the PROTEINS-block; equivalent (aligned) residue in each of the homologous database proteins. The sequences of the test protein and the aligned database proteins run vertically.

SeqNo	sequential residue number of test protein as in DSSP file.
PDBNo	residue number/name as in PDB file.
AA	amino acid type in one letter code
STRUCTURE	secondary structure summary, hydrogen bonding patterns for turns and helices, geometrical bend, chirality, one character name of β-ladder and of β-sheet
BP1, BP2	β-bridge partners.
ACC	solvated residue surface area in $Å^2$ (number of contacting water molecules *10)
NOCC	number of aligned sequences spanning this position (including the test sequence).
VAR	sequence variability (see text) as derived from the NALIGN alignments
....:....1	ruler to identify alignments by their number in the PROTEINS block.

NOTE that lower case characters in the sequence of the test protein (AA-column) indicate cysteines in SS-bridges. Insertions and deletions in either sequence are indicated by special characters in the sequence of the aligned protein;

dots (...)	indicate a deletion in the aligned sequence
lower case characters	bracket an insertion point in the aligned sequence, e.g AkeV means AK[insertion]EV

There are residues from up to 70 database proteins in one line. If the number of alignments (NALIGN) is greater than 70, the alignments block is repeated (1..70, 71−140 etc) until the total number of alignments is reached.

SEQUENCE PROFILE block: relative frequency for each of the 20 amino acid residue in a given sequence position, from counting the residue at that position in each of the aligned sequences including the test sequence. A value of 100 means that at this position only one type of amino acid is found. Asx and Glx are counted in their acid/amide form in proportion to their database frequencies (Asx to Asp: 0.521, Asx to Asn: 0.439, Glx to Glu: 0.623, Glx to Gln: 0.410 as in EMBL/Swissprot release 12, November 1989). For each line, corresponding to a particular sequence position:

NOCC	number of aligned sequences spanning this position (including the test sequence).
NDEL	number of sequences with a deletion in the test protein at this position
NINS	number of sequences with an insertion in the test protein at this position
ENTROPY	entropy measure of sequence variability at this position
RELENT	relative entropy, i.e. entropy normalized to the range 0−100
WEIGHT	conservation weight, around 1.0, lower for less conserved positions, higher for more conserved positions.

Anonymous FTP

If you have access to Internet you can obtain HSSP by anonymous ftp (File Transfer Protocol) from ftp.embl-heidelberg.de in directory:

/pub/databases/protein_extras/hssp.

The program that generates the alignments is currently not available for distribution. Request for alignments based on

Figure 1. Description of HSSP files: One HSSP file contains a structural protein family: one test protein of known structure and all its structurally homologous (as judged by our homology threshold [4]) relatives from the database of known sequences. The file is divided into four blocks, HEADERS, PROTEINS, ALIGNMENTS and SEQUENCE PROFILE. The HEADERS block is mandatory. The other three blocks are present only if at least one homologous alignment is found; each of the additional blocks begins with the string '# #'. File organization is line-oriented. Lines have a maximum length of 132 bytes. Some of the line types are self-explanatory.

structures not in the Protein Data Bank may be sent to R. Schneider 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 maintained. No inclusion in other databases, 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 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, almost 10000 out of 36000 known sequence are homologues of known structures and therefore have an implied know 3-D structure. The complete set of data files currently requires around 200 Mb of disk storage; the selected subset, about 35 Mb. Updates of the database are planned 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 DATA BANKS AND PROGRAMS

The following databases and data services are also available from the Protein Design Group at EMBL, with network access provided by the same mechanisms as for HSSP (see above).

DSSP, a database of secondary structure, solvent accessibility and other information derived from 3-D structures in the Protein Data Bank [2].

personal email: sander@embl-heidelberg.de

FSSP, a database of protein structure families with similar folding motifs, based on 3-D alignments of protein structures [11].
personal email: holm@embl-heidelberg.de

PDB_SELECT, a representative subset of sequence-unique proteins of known 3-D structure selected from the Protein Data Bank [5].
personal email: hobohm@embl-heidelberg.de

PredictProtein, an electronic mail server that provides a predicted secondary structure for any protein sequence with homologues in SwissProt. Rated at 72% sustained 3-state accuracy. [6,7]. Special software is available to construct 3-D models by homology based on the information in HSSP files, such as *WHATIF* by Gert Vriend [8] or *MaxSprout /Torso* by Liisa Holm and Chris Sander [9,10] .

Report any problems to the authors by electronic mail.

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., Tasumi M., J. Mol. Biol. 112:535–542 (1977).
2. Kabsch W., Sander C., Biopolymers 22:2577–2637 (1983).
3. Bairoch A., Boeckmann B., Nucleic Acid Res., 20:2019–2022 (1992).
4. Sander C., Schneider R., PROTEINS 9:56–68 (1991).
5. Hobohm U., Scharf M., Schneider R., Sander C., Protein Science 3:409–417 (1992).
6. Rost B., Schneider R., Sander C., TIBS, 18:120–123 (1993).
7. Rost B., Schneider R., Sander C., CABIOS 10, 53–60
8. Vriend G., J. Mol. Graphics, 8:52–56, (1990).
9. Holm L., Sander C., J. Mol. Biol., 218: 183–194, (1991).
10. Holm L., Sander C., Proteins 14: 213–223. (1992)
11. Holm L., Sander C., NAR, this issue

Table 1.

HSSP Release (month/year)	number of HSSP data sets	number of SWISS-PROT entries (release number)	total number of alignments in the HSSP database	number of unique alignments and fraction of SWISS-PROT in the HSSP database
05/91	488	20024 (17.0)	37715	3065 (15.3%)
02/92	621	22654 (20.0)	43266	3498 (15.4%)
04/92	652	23742 (21.0)	45140	4556 (19.2%)
09/92	736	25044 (22.0)	49784	4825 (19.2%)
02/93	694	28154 (24.0)	54043	5370 (19.1%)
07/93	1361	29955 (25.0)	104837	7197 (24.0%)
10/93	1532	31808 (26.0)	123810	7642 (24.0%)
04/94	1959	36000 (28.0)	148175	9554 (26.5%)