```
#################################################
#
#   MSKCC Document Delivery Services
#
#   Thursday, November 17, 2005
#
##################################################
```

Request ID: DDS36278
User: Gangi-Dino, Rita
Location: MSK
Requested on: 11/17/2005
Needed by: 11/21/2005
Journal Title: Nucleic Acids Res
ISSN: 0305-1048
Article Author(s): Sander C
Article Title: The HSSP data base of protein structure-sequence alignments.
Year: 1993 Jul 1
Volume: 21
Issue: 13
Pages: 3105-9
PMID: 8332531

© 1993 Oxford University Press Nucleic Acids Research, 1993, Vol. 21, No. 13 3105–3109

The HSSP data base of protein structure–sequence alignments

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

INTRODUCTION

HSSP (homology-derived structures of proteins) is a derived data base merging information from three-dimensional structures and one-dimensional sequences of proteins. For each protein of known 3-D structure from the Brookhaven Protein Data Bank (PDB) [1] the data base has a file with all sequence homologues aligned to the PDB protein. Homologues are very likely to have the same 3-D structure as the PDB protein to which they have been aligned. As a result, HSSP is not only a data base of aligned sequence families, but also a data base 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 the analysis of residue conservation in structural context, for the definition of structurally meaningful sequence patterns, and for other questions of 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 proteins of known structure, the derived secondary structure and solvent accessibility from DSSP [2], as well a few or tens or hundreds of sequences deemed homologous to this protein in structure from the SWISS-PROT data base [3] In addition, two different measurements of sequence variability and occupancy at each residue position are given. Details about the methods used to derive the data base and the homology threshold used are given in reference 4.

For example, the dataset 1PPT.HSSP (figure 1) contains 17 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. The alignments could be used to build explicit 3-D models of each of the homologous sequences. If the 3-D-structure of an aligned sequence is known, a pointer to that structure in PDB is given in the column STRID.

As there is considerable redundancy in the PDB data bank, 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. Relative to xxxx.PDB, repeating sequence-identical chains are removed: xxxx.hssp files only contain sequence-unique chains.

DISTRIBUTION

CD-ROM

A subset of the HSSP data base, one file for each protein in a representative set of proteins from the Protein Data Bank (PDB) is distributed on CD-ROM by the EMBL Data Library. In this representative set of PDB proteins, sequence similarity between any two proteins does not exceed 25% identical residues (over a length of 80 or more residues). Detailed information on how the representative set was generated can be found in reference 5 and in documentation distributed with the data base. For enquiries regarding the distribution of HSSP on this medium contact:

EMBL Data Library
European Molecular Biology Laboratory
Postfach 10 2209, Meyerhofstrasse 1
6900 Heidelberg, Germany
Telephone: (+49 6221) 387 258
Telefax: (+49 6221) 387 519 or 387 306
Network: Datalib@embl-heidelberg.de

Network access

Data sets which are not included in this subset and source code of associated utility programs can be obtained from the EMBL file server (6). To get detailed instructions on how to use this service send the following message to the network address Netserv@embl-heidelberg.de:

```
HELP
HELP proteindata
```

If you have access to Internet you can get 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 structures not in the PDB 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 data base is permitted. No inclusion in other data bases, 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 data base.

CONTENT AND SIZE OF THE CURRENT RELEASE

The content and size of the HSSP data base is of course tightly coupled to the development of the PDB and SWISS-PROT data banks. An overview on the increase in size is given in Table 1.

The complete set of data files require around 70 Mb of disk storage. Updates of the data base are planned on a regular bases.

LIMITATIONS

Accuracy of reported alignments

In general, alignments may deviate from structural alignment in local detail (trailing ends differently aligned, shifted gaps etc.). 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 are often uncertain in loop regions.

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. The user may choose to use only positions with occupancy larger than, say, five proteins.

RELATED DATA BANKS AND PROGRAMS

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

DSSP, a data base 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 data base of protein structure families with similar folding motifs, based on 3-D alignments of protein structures. 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 for academics users that provides a secondary structure prediction for any protein sequence with homologues in SwissProt. Rated at 70.8% sustained 3-state accuracy. [6].

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 by Liisa Holm and Chris Sander [9].

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. Stoehr P.J., Omond R.A., Nucleic Acid Res. 17:6763–6764 (1989).
7. Rost B., Schneider R., Sander C., TIBS, in press.
8. Vriend G., J. Mol. Graphics, 8:52–56, (1990).
9. Holm L., Sander C., J. Mol. Biol., 218: 183–194, (1991).

Table 1.

HSSP Release	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%)

* (at least 30% identical to a PDB protein over a length of 80 or more residues)

```
HSSP       HOMOLOGY DERIVED SECONDARY STRUCTURE OF PROTEINS , VERSION 1.0 1991
PDBID      1ppt
DATE       file generated on 31-Aug-92
SEQBASE    RELEASE 22.0 OF EMBL/SWISS-PROT WITH  25044 SEQUENCES
PARAMETER  SMIN: -0.5  SMAX:  1.0
PARAMETER  gap-open:  3.0 gap-elongation:  0.1
PARAMETER  conservation weights
PARAMETER  no insertions/deletions in secondary structure allowed
PARAMETER  alignments sorted according to: DISTANCE
THRESHOLD  according to t(L)=(290.15 * L ** -0.562) +  5
REFERENCE  Sander C., Schneider R. : Database of homology-derived protein structures. Proteins, Proteins, 9:56-68 (1991).
CONTACT    e-mail (INTERNET) Schneider@EMBL-Heidelberg.DE or Sander@EMBL-Heidelberg.DE / phone +49-6221-387361 / fax +49-6221-387306
AVAILABLE  Free academic use. Commercial users must apply for license.
HEADER     PANCREATIC HORMONE
COMPND     AVIAN PANCREATIC POLYPEPTIDE
SOURCE     TURKEY (MELEAGRIS GALLOPAVO) PANCREAS
AUTHOR     T.L.BLUNDELL,J.E.PITTS,I.J.TICKLE,S.P.WOOD
SEQLENGTH    36
NCHAIN        1 chain(s) in 1ppt.DSSP data set
NALIGN       17
NOTATION : ID: EMBL/SWISSPROT identifier of the aligned (homologous) protein
NOTATION : STRID: if the 3-D structure of the aligned protein is known, then STRID is the Protein Data Bank identifier as taken
NOTATION : from the database reference or DR-line of the EMBL/SWISSPROT entry
NOTATION : %IDE: percentage of residue identity of the alignment
NOTATION : %SIM (%WSIM):  (weighted) similarity of the alignment
NOTATION : IFIR/ILAS: first and last residue of the alignment in the test sequence
NOTATION : JFIR/JLAS: first and last residue of the alignment in the alignend protein
NOTATION : LALI: length of the alignment excluding insertions and deletions
NOTATION : NGAP: number of insertions and deletions in the alignment
NOTATION : LGAP: total length of all insertions and deletions
NOTATION : LSEQ2: length of the entire sequence of the aligned protein
NOTATION : ACCNUM: SwissProt accession number
NOTATION : PROTEIN: one-line description of aligned protein
NOTATION : SeqNo,PDBNo,AA,STRUCTURE,BP1,BP2,ACC: sequential and PDB residue numbers, amino acid (lower case = Cys), secondary
NOTATION : structure, bridge partners, solvent exposure as in DSSP (Kabsch and Sander, Biopolymers 22, 2577-2637(1983)
NOTATION : VAR: sequence variability on a scale of 0-100 as derived from the NALIGN alignments
NOTATION : pair of lower case characters (AvaK) in the alignend sequence bracket a point of insertion in this sequence
NOTATION : dots (....) in the alignend sequence indicate points of deletion in this sequence
NOTATION : SEQUENCE PROFILE: relative frequency of an amino acid type at each position. Asx and Glx are in their
NOTATION : acid/amide form in proportion to their database frequencies
NOTATION : NOCC: number of aligned sequences spanning this position (including the test sequence)
NOTATION : NDEL: number of sequences with a deletion in the test protein at this position
NOTATION : NINS: number of sequences with an insertion in the test protein at this position
NOTATION : ENTROPY: entropy measure of sequence variability at this position
NOTATION : RELENT: relative entropy, i.e.  entropy normalized to the range 0-100
NOTATION : WEIGHT: conservation weight


## PROTEINS : EMBL/SWISSPROT identifier and alignment statistics
  NR.    ID         STRID   %IDE %WSIM IFIR ILAS JFIR JLAS LALI NGAP LGAP LSEQ2 ACCNUM     PROTEIN
    1 : paho_chick  1PPT    1.00  1.00    1   36    1   36   36    0    0   36  P01306     PANCREATIC HORMONE.
    2 : paho_strca          0.94  0.97    1   36    1   36   36    0    0   36  P11967     PANCREATIC HORMONE.
    3 : paho_allmi          0.80  0.85    2   36    2   36   35    0    0   36  P06305     PANCREATIC HORMONE.
    4 : paho_ansan          0.78  0.76    1   36    1   36   36    0    0   36  P06304     PANCREATIC HORMONE.
    5 : neuy_sheep          0.60  0.75    2   36    2   36   35    0    0   36  P14765     NEUROPEPTIDE Y (NPY).
    6 : neuy_pig            0.57  0.73    2   36    2   36   35    0    0   36  P01304     NEUROPEPTIDE Y (NPY).
    7 : pyy_human           0.54  0.70    2   36    2   36   35    0    0   36  P10082     PEPTIDE YY (PYY).
    8 : neuy_rat            0.54  0.72    2   36   31   65   35    0    0   98  P07808     NEUROPEPTIDE Y PRECURSOR (NPY).
    9 : neuy_human          0.54  0.72    2   36   30   64   35    0    0   97  P01303     NEUROPEPTIDE Y PRECURSOR.
   10 : neuy_rabit          0.54  0.72    2   36    2   36   35    0    0   36  P09640     NEUROPEPTIDE Y (NPY).
   11 : pyy_rat             0.54  0.71    2   36   30   64   35    0    0   98  P10631     PEPTIDE YY PRECURSOR (PYY).
   12 : pyy_pig             0.54  0.71    2   36    2   36   35    0    0   36  P01305     PEPTIDE YY (PYY).
   13 : pp_lepsp            0.49  0.68    2   36    2   36   35    0    0   36  P09473     PANCREATIC POLYPEPTIDE (PP) (NEUROPEPTIDE
   14 : pp_oncki            0.49  0.68    2   36    2   36   35    0    0   36  P09474     PANCREATIC POLYPEPTIDE (PP).
   15 : paho_canfa          0.46  0.63    2   36   31   65   35    0    0   93  P01299     PANCREATIC HORMONE PRECURSOR.
   16 : paho_pig            0.46  0.63    2   36    2   36   35    0    0   36  P01300     PANCREATIC HORMONE.
   17 : paho_didma          0.46  0.63    2   36    2   36   35    0    0   36  P18107     PANCREATIC HORMONE.
```

```
## ALIGNMENTS    1 -   17
 SeqNo  PDBNo AA STRUCTURE BP1 BP2  ACC NOCC  VAR  ....:....1....:....2....:....3....:....4....:....5....:....6....:....7
     1     1  G              0   0  101   `4    0  GG G
     2     2  P          -   0   0   60   18    0  PPPPPPPPPPPPPPPPP
     3     3  S          -   0   0  106   18   45  SALSSSISSSAAPPLLQ
     4     4  Q          -   0   0  139   18   32  QQQQKKKKKKKKKKEEE
     5     5  P          -   0   0   26   18    0  PPPPPPPPPPPPPPPPP
     6     6  T          -   0   0  126   18   46  TTKTDDEDDDEEEEVVV
     7     7  Y          -   0   0  121   18   47  YYYYNNANNNAANNYYY
     8     8  P          -   0   0   55   18    0  PPPPPPPPPPPPPPPPP
     9     9  G      >   -   0   0   27   18    0  GGGGGGGGGGGGGGGGG
    10    10  D  T 3   S+   0   0  128   18   15  DDDNDEEEEEEEEEDDD
    11    11  D  T 3   S+   0   0  165   18    4  DDGDDDDDDDDDDDDDD
    12    12  A  S <   S-   0   0   17   18    0  AAAAAAAAAAAAAAAAA
    13    13  P     >>   -   0   0   73   18   27  PPPPPPSPPPSSPPTTT
    14    14  V  H 3> S+   0   0  108   18   38  VVVVAAPAAAPPPPPPP
    15    15  E  H 3> S+   0   0  111   18    0  EEEEEEEEEEEEEEEEE
    16    16  D  H <> S+   0   0   58   18   18  DDDDDDEDDDEEEEQQQ
    17    17  L  H  X S+   0   0   62   18    4  LLLLLLLMMLLLLMMM
    18    18  I  H  X S+   0   0   91   18   42  IVIRAANAAASSAAAAA
    19    19  R  H  X S+   0   0  142   18   30  RRQFRRRRRRRRKKQQK
    20    20  F  H  X S+   0   0   39   18    2  FFFYYYYYYYYYYYYYY
    21    21  Y  H  X S+   0   0  143   18   26  YYYYYYYYYYYYYYAAA
    22    22  D  H  X S+   0   0   79   18   39  DDDDSSASSSAASTAAA
    23    23  N  H  X S+   0   0   97   18   41  NNDNAASAAASSAAEEE
    24    24  L  H  X S+   0   0   58   18    0  LLLLLLLLLLLLLLLLL
    25    25  Q  H  X S+   0   0   90   18   23  QQQQRRRRRRRRRRRRR
    26    26  Q  H  X S+   0   0  112   18   27  QQQQHHHHHHHHHHRRR
    27    27  Y  H  X S+   0   0   56   18    0  YYYYYYYYYYYYYYYYY
    28    28  L  H  X S+   0   0   90   18   24  LLLRIILIIILLIIIII
    29    29  N  H  <>S+   0   0   33   18   10  NNNLNNNNNNNNNNNNN
    30    30  V  H ><5S+   0   0   21   18   30  VVVNLLLLLLLLLLMMR
    31    31  V  H 3<5S+   0   0   85   18   16  VVVVIIVIIIVVIILLL
    32    32  T  T 3<5S-   0   0   91   18    9  TTTFTTTTTTTTTTTTT
    33    33  R  T < 5S+   0   0  215   18    0  RRRRRRRRRRRRRRRRR
    34    34  H      < +   0   0  101   18   31  HHPHQQQQQQQQQQPPP
    35    35  R              0   0  180   18    0  RRRRRRRRRRRRRRRRR
    36    36  Y              0   0  224   18    1  YYFYYYYYYYYYYYYYY
```

```
## SEQUENCE PROFILE AND ENTROPY
 SeqNo PDBNo   V   L   I   M   F   W   Y   G   A   P   S   T   C   H   R   K   Q   E   N   D  NOCC NDEL NINS ENTROPY RELENT WEIGHT
    1     1    0   0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0   0   0     4    0    0   0.000      0   1.03
    2     2    0   0   0   0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0    18    0    0   0.000      0   1.27
    3     3    0  17   6   0   0   0   0   0  17  11  44   0   0   0   0   0   6   0   0   0    18    0    0   1.523     53   0.75
    4     4    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  56  28  17   0   0    18    0    0   0.981     34   0.74
    5     5    0   0   0   0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0    18    0    0   0.000      0   1.25
    6     6   17   0   0   0   0   0   0   0   0   0   0  22   0   0   0   6   0  28   0  28    18    0    0   1.505     52   0.60
    7     7    0   0   0   0   0   0  44   0  17   0   0   0   0   0   0   0   0   0  39   0    18    0    0   1.026     36   0.60
    8     8    0   0   0   0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0    18    0    0   0.000      0   1.17
    9     9    0   0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0   0   0    18    0    0   0.000      0   1.28
   10    10    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  50   6  44    18    0    0   0.868     30   1.19
   11    11    0   0   0   0   0   0   0   6   0   0   0   0   0   0   0   0   0   0   0  94    18    0    0   0.215      7   1.26
   12    12    0   0   0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0   0    18    0    0   0.000      0   1.28
   13    13    0   0   0   0   0   0   0   0   0  67  17  17   0   0   0   0   0   0   0   0    18    0    0   0.868     30   0.98
   14    14   28   0   0   0   0   0   0   0  28  44   0   0   0   0   0   0   0   0   0   0    18    0    0   1.072     37   0.60
   15    15    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 100   0   0    18    0    0   0.000      0   1.31
   16    16    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  17  28   0  56    18    0    0   0.981     34   1.00
   17    17    0  67   0  33   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0    18    0    0   0.637     22   1.13
   18    18    6   0  17   0   0   0   0   0  56   0  11   0   0   0   6   0   0   0   6   0    18    0    0   1.351     47   0.60
   19    19    0   0   0   0   6   0   0   0   0   0   0   0   0   0  61  17  17   0   0   0    18    0    0   1.059     37   0.99
   20    20    0   0   0   0  22   0  78   0   0   0   0   0   0   0   0   0   0   0   0   0    18    0    0   0.530     18   1.44
   21    21    0   0   0   0   0   0  83   0  17   0   0   0   0   0   0   0   0   0   0   0    18    0    0   0.451     16   1.00
   22    22    0   0   0   0   0   0   0   0  33   0  33   6   0   0   0   0   0   0   0  28    18    0    0   1.249     43   0.66
   23    23    0   0   0   0   0   0   0   0  39   0  17   0   0   0   0   0   0  17  22   6    18    0    0   1.459     50   0.61
   24    24    0 100   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0    18    0    0   0.000      0   1.14
   25    25    0   0   0   0   0   0   0   0   0   0   0   0   0   0  72   0  28   0   0   0    18    0    0.  0.591     20   0.87
   26    26    0   0   0   0   0   0   0   0   0   0   0   0   0  56  17   0  28   0   0   0    18    0    0   0.981     34   0.80
   27    27    0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0   0   0   0    18    0    0   0.000      0   1.12
   28    28    0  39  56   0   0   0   0   0   0   0   0   0   0   0   6   0   0   0   0   0    18    0    0   0.854     30   0.80
   29    29    0   6   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  94   0    18    0    0   0.215      7   1.08
   30    30   22  56   0  11   0   0   0   0   0   0   0   0   0   0   6   0   0   0   6   0    18    0    0   1.226     42   0.82
   31    31   44  17  39   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0    18    0    0   1.026     36   1.01
   32    32    0   0   0   0   6   0   0   0   0   0   0  94   0   0   0   0   0   0   0   0    18    0    0   0.215      7   1.13
   33    33    0   0   0   0   0   0   0   0   0   0   0   0   0   0 100   0   0   0   0   0    18    0    0   0.000      0   1.22
   34    34    0   0   0   0   0   0   0   0   0  22   0   0   0  22   0   0  56   0   0   0    18    0    0   0.995     34   0.84
   35    35    0   0   0   0   0   0   0   0   0   0   0   0   0   0 100   0   0   0   0   0    18    0    0   0.000      0   1.14
   36    36    0   0   0   0   6   0  94   0   0   0   0   0   0   0   0   0   0   0   0   0    18    0    0   0.215      7   1.33
//
```

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.

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.