226–230 Nucleic Acids Research, 1997, Vol. 25, No. 1 © 1997 Oxford University Press

The HSSP database of protein structure–sequence alignments

Reinhard Schneider*, Antoine de Daruvar and Chris Sander

Protein Design Group, European Molecular Biology Labratory, D-69012 Heidelberg, Germany

Received October 7, 1996; Accepted October 8, 1996

ABSTRACT

HSSP is a derived database merging structural (3-D) and sequence (1-D) information. For each protein of known 3-D structure from the Protein Data Bank (PDB), the database has a multiple sequence alignment of all available homologues and a sequence profile characteristic of the family. The list of homologues is the result of a database search in SwissProt using a position-weighted dynamic programming method for sequence profile alignment (MaxHom). The database is updated frequently. The listed homologues are very likely to have the same 3-D structure as the PDB protein to which they have been aligned. As a result, the database is not only a database of aligned sequence families, but also a database of implied secondary and tertiary structures covering 29% of all SwissProt-stored sequences.

INTRODUCTION

HSSP (homology-derived structures of proteins) is a derived database merging information from 3-D structures and 1-D 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 3-D).

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 (like: 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) at each position in the multiple sequence alignment, sequence variability, using two different measures, and (iv) 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. Some details of the methods are given elsewhere (4).

For example, the dataset 1PPT.HSSP (Fig. 1) contains 30 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. Sequence-identical chains in the PDB entry are removed so that 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

* To whom correspondence should be addressed. Tel: +49 6221 387305; Fax: +49 6221 387517; Email: schneider@embl-heidelberg.de

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 **HEADER**S 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.

```
(a)
HSSP        HOMOLOGY DERIVED SECONDARY STRUCTURE OF PROTEINS , VERSION 1.0 1991
PDBID       1ppt
DATE        file generated on 25-Jul-96
SEQBASE     RELEASE 33.0 OF EMBL/SWISS-PROT WITH  52205 SEQUENCES
PARAMETER   SMIN: -0.5  SMAX:  1.0
PARAMETER   gap-open:  3.0 gap-elongation:  0.1
PARAMETER   conservation weights: YES
PARAMETER   InDels in secondary structure allowed: YES
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
REFERENCE   structures. Proteins, 9:56-68 (1991).
CONTACT     e-mail (INTERNET) Schneider@EMBL-Heidelberg.DE
AVAILABLE   Free academic use. Commercial users must apply for license.
AVAILABLE   No inclusion in other databanks without permission.
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 data set
NALIGN       30
```

(**a**) **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. 1 PPT), the SEQBASE-line specifies the source of the aligned sequences (e.g. EMBL/Swiss-Prot 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).

```
(b)
## 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  0.94  0.98    1   36   26   61   36    0    0    80  P01306 PANCREATIC HORMONE
  2 : paho_strca       0.94  0.98    1   36    1   36   36    0    0    36  P11967 PANCREATIC HORMONE
  3 : paho_larar       0.89  0.95    1   36    1   36   36    0    0    36  P41337 PANCREATIC HORMONE
  4 : paho_allmi       0.80  0.87    2   36    2   36   35    0    0    36  P06305 PANCREATIC HORMONE
  5 : paho_ansan       0.78  0.74    1   36    1   36   36    0    0    36  P06304 PANCREATIC HORMONE
  6 : neuy_sheep       0.60  0.75    2   36    2   36   35    0    0    36  P14765 NEUROPEPTIDE Y
  7 : neuy_pig         0.57  0.73    2   36    2   36   35    0    0    36  P01304 NEUROPEPTIDE
  8 : pyy_pig          0.54  0.72    2   36    2   36   35    0    0    36  P01305 PEPTIDE YY
  9 : neuy_rat         0.54  0.71    2   36   31   65   35    0    0    98  P07808 NEUROPEPTIDE Y PRECURSOR
 10 : pyy_human        0.54  0.71    2   36   30   64   35    0    0    97  P10082 PEPTIDE YY PRECURSOR
 11 : pyy_rat          0.54  0.72    2   36   30   64   35    0    0    98  P10631 PEPTIDE YY PRECURSOR
 12 : neuy_rabit       0.54  0.71    2   36    2   36   35    0    0    36  P09640 NEUROPEPTIDE Y
 13 : neuy_chick       0.54  0.72    2   36   30   64   35    0    0    97  P28673 NEUROPEPTIDE Y
 14 : neuy_human       0.54  0.71    2   36   30   64   35    0    0    97  P01303 NEUROPEPTIDE Y
 15 : neuy_ranri       0.51  0.70    2   36    2   36   35    0    0    36  P29949 NEUROPEPTIDE Y
 16 : neuy_torma       0.51  0.68    2   36   30   64   35    0    0    98  P28674 NEUROPEPTIDE Y PRECURSOR
 17 : pyy_amica        0.51  0.70    2   36    2   36   35    0    0    36  P28205 PEPTIDE YY-LIKE
 18 : neuy_xenla       0.51  0.70    2   36   30   64   35    0    0    97  P33689 NEUROPEPTIDE Y PRECURSOR
 19 : pyy_rajrh        0.49  0.69    2   36    2   36   35    0    0    36  P29206 PEPTIDE YY-LIKE
 20 : npy_lamfl        0.49  0.68    2   36   36   70   35    0    0   104  P49087 NEUROPEPTIDE Y PRECURSOR
 21 : pyy_lepsp        0.49  0.69    2   36    2   36   35    0    0    36  P09473 PEPTIDE YY-LIKE
 22 : neuy_oncmy       0.49  0.69    2   36    2   36   35    0    0    36  P29471 NEUROPEPTIDE Y
 23 : pyy_oncki        0.49  0.69    2   36    2   36   35    0    0    36  P09474 PEPTIDE YY-LIKE
 24 : neuy_carau       0.46  0.66    2   36   30   64   35    0    0    96  P28675 NEUROPEPTIDE Y PRECURSOR
 25 : neuy_gadmo       0.46  0.67    2   36    2   36   35    0    0    36  P81147 NEUROPEPTIDE Y
 26 : paho_didma       0.44  0.63    1   36    1   36   36    0    0    36  P16187 PANCREATIC HORMONE
 27 : paho_rante       0.44  0.60    1   36    1   36   36    0    0    36  P31289 PANCREATIC HORMONE
 28 : paho_canfa       0.44  0.64    1   36   30   65   36    0    0    93  P01295 PANCREATIC HORMONE
 29 : paho_rabit       0.44  0.64    1   36    1   36   36    0    0    36  P41336 PANCREATIC HORMONE
 30 : paho_pig         0.44  0.64    1   36    1   36   36    0    0    36  P01300 PANCREATIC HORMONE
```

(**b**) **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/Swiss-Prot 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/Swiss-Prot 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, Swiss-Prot accession number; PROTEIN, one-line description of aligned protein.

```
(c)
## ALIGNMENTS    1 -   30
SeqNo PDBNo AA STRUCTURE BP1 BP2 ACC NOCC VAR  ....:....1....:....2....:....3
    1    1  G              0   0 101   10  23  GGG G                    AAAAA
    2    2  P              -   0   0  60   31   0  PPPPPPPPPPPPPPPPPPPPPPPPPPPPP
    3    3  S              -   0   0 106   31  44  SAVLSSSASIASSSSSPSPNPVPTIQSLPL
    4    4  Q              -   0   0 139   31  31  QQQQQKKKKKKKKKKKKKKKKKKKEEEEE
    5    5  P              -   0   0  26   31   0  PPPPPPPPPPPPPPPPPPPPPPPPPPPPP
    6    6  T              -   0   0 126   31  43  TTTKTDDEDEEDDDDDEDEDEDEEEDEVHVVV
    7    7  Y              -   0   0 121   31  47  YYYYYNNANAANSNNNNNNSNNNNNYHYYY
    8    8  P              -   0   0  55   31   0  PPPPPPPPPPPPPPPPPPPPPPPPPPPPP
    9    9  G    >         -   0   0  27   31   0  GGGGGGGGGGGGGGGGGGGGGGGGGGGGG
   10   10  D  T 3   S+     0   0 128   31  14  DDDDNDEEEEEEEEEEEEDEEEEEEDDDDD
   11   11  D  T 3   S+     0   0 165   31   8  DDDGDDDDDDDDDDDDGDDDDDDDGDDQDDD
   12   12  A  S <   S-     0   0  17   31   0  AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
   13   13  P     >>  -     0   0  73   31  25  PPPPPPPSPSSPPPPPPPPAPPPPPPTTTTT
   14   14  V  H 3> S+     0   0 108   31  38  VVVVVAAPAPPAAAAAPAPAPTPAAPQPPP
   15   15  E  H 3> S+     0   0 111   31   3  EEEEEEEEEEEEEEEEEEEEEEEEDEDEEE
   16   16  D  H <> S+     0   0  58   31  18  DDDDDDDEDEEDDDDDEDEDEEEEEQQQQQ
   17   17  L  H  X S+     0   0  62   31   4  LLLLLLLLMLLMMMMLLMLLLLLLLMLMMM
   18   18  I  H  X S+     0   0  91   31  33  IVVIRAASANSAAAAAAAAAAAAAAAAAAA
   19   19  R  H  X S+     0   0 142   31  30  RRRQFRRRRRRRRKKKRKKKKKKKKKQQEQ
   20   20  F  H  X S+     0   0  39   31   1  FFFFYYYYYYYYYYYYYYYYYYYYYYYYYY
   21   21  Y  H  X S+     0   0 143   31  23  YYYYYYYYYYYYYYYYYYYYLYYYYYAYAVA
   22   22  D  H  X S+     0   0  79   31  37  NDNDDSSASAASSSSSSTSSSSTTSSASAAA
   23   23  N  H  X S+     0   0  97   31  37  DNDDNAASASSAAAAAAAAAAAAAAEDEDE
   24   24  L  H  X S+     0   0  58   31   2  LLLLLLLLLLLLLLLLLLLLVLLLLLLLLL
   25   25  Q  H  X S+     0   0  90   31  26  QQQQQRRRRRRRRRRRRRRRRRRRRYRRR
   26   26  Q  H  X S+     0   0 112   31  25  QQQQQHHHHHHHHHHHHHHHHHHHHRQRRR
   27   27  Y  H  X S+     0   0  56   31   0  YYYYYYYYYYYYYYYYYYYYYYYYYYYYYY
   28   28  L  H  X S+     0   0  90   31  19  LLLLRIILILLIIIIIIIIIIIIIIIIIII
   29   29  N  H  <>S+     0   0  33   31  11  NNNNLNNNNNNNNNNNNNNNNNNNNNTNNN
   30   30  V  H ><5S+     0   0  21   31  24  VVVVNLLLLLLLLLLLLLLLLLLLRFMMM
   31   31  V  H 3<5S+     0   0  85   31  15  VVVVVIIVIVVIIIIIIIIIIIIIILVLLL
   32   32  T  T 3<5S-     0   0  91   31   6  TTTFTTTTTTTTTTTTTTTTTTTTTTTTTT
   33   33  R  T < 5S+     0   0 215   31   0  RRRRRRRRRRRRRRRRRRRRRRRRRRRRR
   34   34  H      < +     0   0 101   31  29  HHHPHQQQQQQQQQQQQQQQQQQPPPPP
```

(**c**) **ALIGNMENT**S 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 × l0); 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 of up to 70 database proteins in one line. If the number of alignments (NALIGN) is >70, the alignments block is repeated (1..70, 71–140 etc.) until the total number of alignments is reached.

```
(d)
## 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  50  50   0   0   0   0   0   0   0   0   0   0   0   10    0    0   0.693     30   0.87
    2    2  0   0   0   0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0   31    0    0   0.000      0   1.31
    3    3  6  10   6   0   0   0   0   0  10  16  42   3   0   0   0   0   3   0   3   0   31    0    0   1.797     60   0.43
    4    4  0   0   0   0   0   0   0   0   0   0   0   0   0   0  65  19  16   0   0   0   31    0    0   0.895     30   0.78
    5    5  0   0   0   0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0   31    0    0   0.000      0   1.31
    6    6 13   0   0   0   0   0   0   0   0   0  16   0   3   0   3   0   3   0  29   0  35   31    0    0   1.507     50   0.51
    7    7  0   0   0   0   0   0  32   0  10   0   6   0   0   3   0   0   0   0  48   0   31    0    0   1.230     41   0.61
    8    8  0   0   0   0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0   31    0    0   0.000      0   1.31
    9    9  0   0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0   0   0   31    0    0   0.000      0   1.31
   10   10  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  58   3  39   31    0    0   0.794     26   1.10
   11   11  0   0   0   0   0   0   0  10   0   0   0   0   0   0   0   0   3   0   0  87   31    0    0   0.457     15   1.13
   12   12  0   0   0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0   0   31    0    0   0.000      0   1.31
   13   13  0   0   0   0   0   0   0   0   3  71  10  16   0   0   0   0   0   0   0   0   31    0    0   0.874     29   0.73
   14   14 19   0   0   0   0   0   0   0  39  35   0   3   0   0   0   0   3   0   0   0   31    0    0   1.274     43   0.64
   15   15  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  94   0   6   31    0    0   0.239      8   1.25
   16   16  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  16  32   0  52   31    0    0   1.001     33   0.92
   17   17  0  68   0  32   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   31    0    0   0.629     21   1.24
   18   18  6   0  10   0   0   0   0   0  71   0   6   0   0   0   3   0   0   0   3   0   31    0    0   1.045     35   0.82
   19   19  0   0   0   0   3   0   0   0   0   0   0   0   0   0  48  32  13   3   0   0   31    0    0   1.202     40   0.73
   20   20  0   0   0   0  16   0  84   0   0   0   0   0   0   0   0   0   0   0   0   0   31    0    0   0.442     15   1.09
   21   21  3   3   0   0   0   0  81   0  10   0   0   0   0   0   0   0   0   0   0   0   31    0    0   0.595     20   0.76
   22   22  0   0   0   0   0   0   0   0  23   0  48  10   0   0   0   0   0   0   6  13   31    0    0   1.354     45   0.65
   23   23  0   0   0   0   0   0   0   0  55   0  10   0   0   0   0   0   0  10  10  16   31    0    0   1.302     43   0.66
   24   24  3  97   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   31    0    0   0.143      5   1.26
   25   25  0   0   0   0   0   0   3   0   0   0   0   0   0   0  77   0  19   0   0   0   31    0    0   0.627     21   0.90
   26   26  0   0   0   0   0   0   0   0   0   0   0   0   0  65  13   0  23   0   0   0   31    0    0   0.883     29   0.86
   27   27  0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0   0   0   0   31    0    0   0.000      0   1.31
   28   28  0  26  71   0   0   0   0   0   0   0   0   0   0   0   3   0   0   0   0   0   31    0    0   0.704     23   1.05
   29   29  0   3   0   0   0   0   0   0   0   0   0   3   0   0   0   0   0   0  94   0   31    0    0   0.284      9   1.15
   30   30 16  65   0  10   3   0   0   0   0   0   0   0   0   0   3   0   0   0   3   0   31    0    0   1.135     38   0.93
   31   31 32  13  55   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   31    0    0   0.959     32   1.01
   32   32  0   0   0   0   3   0   0   0   0   0   0  97   0   0   0   0   0   0   0   0   31    0    0   0.143      5   1.27
   33   33  0   0   0   0   0   0   0   0   0   0   0   0   0   0 100   0   0   0   0   0   31    0    0   0.000      0   1.31
//
```

(**d**) **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/Swiss-Prot 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, ~1.0; lower for less conserved positions; higher for more conserved positions.

Table 1.

HSSP Release (month / year)	number of HSSP data sets	number of SWISS-PROT entries	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	37715	3065 (15.3%)
02/92	621	22654	43266	3498 (15.4%)
04/92	652	23742	45140	4556 (19.2%)
09/92	736	25044	49784	4825 (19.2%)
02/93	694	28154	54043	5370 (19.1%)
07/93	1361	29955	104837	7197 (24.0%)
10/93	1532	31808	123810	7642 (24.0%)
04/94	1959	36000	148175	9554 (26.5%)
08/94	2158	38303	154590	10136 (26.5%)
08/95	3158	43470	241518	11762 (27.0%)
09/96	4189	52205	317231	15140 (29.0%)

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 ref. (5) and the documentation distributed with the database. For enquiries regarding the distribution of HSSP on this medium contact: EMBL Data Library, European Bioinformatics Institute, Hinxton Hall, Hinxton, Cambridge CB10 1RQ, UK. Tel: +44 (0)1223 494401; Fax: +44 (0)1223 494468; Network: datalib@embl-ebi.ac.uk; WWW: http://www.embl-ebi.ac.uk

Anonymous FTP

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

World Wide Web

The HSSP database and HSSP-related information and data are accessible via the generic URL:
http://www.sander.embl-heidelberg.de/

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 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 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 3-D structures (PDB) and sequences (e.g. SWISS-PROT). An overview of the increase in size is given in Table 1. Interestingly, >15 000 of 52 205 known sequences (SwissProt release 33) are homologues of known structures and therefore have an implied know 3-D structure.

The complete set of data files currently requires ~450 Mb of disk storage; the selected subset (480 datasets), ~50 Mb. 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 DATA BANKS AND INFORMATION SERVICES

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

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-heidelberg.de/dssp/
personal email: sander@embl-ebi.ac.uk

FSSP, a database of protein structure families with similar folding motifs, based on 3-D alignments of protein structures. http://www.sander.embl-heidelberg.de/fssp/

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/pdbselect/
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 SwissProt. Rated at 72% sustained three-state accuracy (6,7).
http://www.embl-heidelberg.de/predictprotein/
personal email: predict-help@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 [e.g., *Haemophilus influenzae* (9), yeast].
http://www.sander.embl-heidelberg.de/genequiz/
personal email: genequiz@embl-heidelberg.de

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

Dali, an electronic mail server that performs a 3-D similarity search in the Protein Data Bank, given the atomic coordinates of a 3-D protein model as input (10).
http://www.sander.embl-heidelberg.de/dali/
personal email: holm@embl-heidelberg.de

Special software is available to construct 3-D models by homology based on the information in HSSP files, such as *WHATIF* by Gert Vriend (11) or *MaxSprout/Torso* by Liisa Holm and Chris Sander (12,13).

Report any problems with the HSSP database to the authors by electronic mail: schneider@embl-heidelberg.de or sander@embl-heidelberg.de

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. (1977) *J. Mol. Biol.* **112**, 535–542.
2 Kabsch W., Sander C. (1983) *Biopolymers* **22**, 2577–2637.
3 Bairoch A., Boeckmann B. (1992) *Nucleic Acid Res.* **20**, 2019–2022.
4 Sander C., Schneider R. (19991) *Proteins* **9**, 56–68.
5 Hobohm U., Scharf M., Schneider R., Sander C. (1992) *Protein Sci.* **3**, 409–417.
6 Rost B., Schneider R., Sander C. (1993) *Trends Biol. Sci* **18**, 120–123.
7 Rost B., Schneider R., Sander C. (1994) *Comput. Appl. Biosci.* **10**, 53–60.
8 Hobohm U., Sander C. (1995) *J. Mol. Biol.* **251**, 390–399.
9 Casari G., Andrade M.A., Bork P., Boyle J., Daruvar A., Ouzounis C., Schneider R., Tamames J., Valencia A., Sander C. (1995) *Nature* **376**, 647–648.
10 Holm L, Sander C. (1993) *J. Mol. Biol.* **233**, 123–138.
11 Vriend G. (1990) *J. Mol. Graphics* **8**, 52–56.
12 Holm L., Sander C. (1991) *J. Mol. Biol.* **218**, 183–194.
13 Holm L., Sander C. (1992) *Proteins* **14**, 213–223.