BACKGROUND FOR BEGINNERS 2/10/01

What is the Homstrad protein structures database for?

HOMSTRAD (HOMologous STRucture Alignment Database) provides a collection of structure-based alignments of homologous protein sequences. The database is compiled from structures present in the PDB (Protein Data Bank - http://pdb.ccdc.cam.ac.uk/pdb/) and these have been grouped into just over 800 multi-member families at present.
This type of information is useful in two main areas:

- Comparative modeling – predicting a protein's structure by comparing it with sequence homologs of known structure.
- Detecting distantly related proteins by using structural as well as sequence comparison – the hypothesis being that structure is more strictly conserved than sequence.

We would also like to apply the database to genome wide analyses of protein families. Some initial work on microbial genomes has been carried out and the *Drosophila* genome will be our next project.

Determining the structure of a protein

Protein structures can be determined experimentally in a number of different ways including…

- X-ray crystallography (good if you can express it and get nice crystals)
- NMR (small proteins/domains)
- Electron microscopy (can be to 2D arrays eg used for proteins in membranes, 1D eg helical structures, or single particle).

These procedures are time consuming and can be problematic. It would be better if you could use bioinformatics approaches, firstly to support experimental approaches (by giving you an idea of secondary structure, domain organization, potential active sites or structurally important residues) and ultimately to reliably predict a protein's structure based solely on its sequence, something that is not possible at present. Advances are being made in structural prediction and a number of approaches are being taken (*ab initio*, threading – sequence to fold assignment, comparative modeling). Using sequence homology to known structures, i.e. comparative modeling, has been the most successful of these but depends on the availability of high quality sequence alignments, something which Homstrad provides.

Important databases

There are already a variety of publicly available protein databases on the web, each with a different focus. We use data from a number of these during the construction and maintenance of Homstrad, so a brief introduction to some of them is given here.

Biological databases can range from providing a basic searchable depository of data to more informative setups (often manually curated) that provide information on the relationships between the member entries and supply varying levels of annotation.

Type of database	Sequence database examples	Structure database examples
Primary depository - Often high level of redundancy	Genbank/EMBL/DDBJ http://www.ncbi.nlm.nih.gov/Genbank/ http://www.ebi.ac.uk/embl/ http://www.ddbj.nig.ac.jp/ TrEMBL	PDB http://pdb.ccdc.cam.ac.uk/pdb/
Secondary database - Some annotation added - Redundancy reduced	SwissProt http://www.ebi.ac.uk/swissprot/	-
Classification database - Hierarchical	ProtoMap http://www.protomap.cs.huji.ac.il/	SCOP http://scop.mrc-lmb.cam.ac.uk/scop/ CATH http://www.biochem.ucl.ac.uk/bsm/cath_new/
Family database - entries grouped and aligned according to sequence or structure	Pfam http://www.sanger.ac.uk/Software/Pfam/	Homstrad http://www-cryst.bioc.cam.ac.uk/~homstrad/

The PDB (http://pdb.ccdc.cam.ac.uk/pdb/) is the main depository for protein structures. It is the starting point for structural databases to build upon and it is updated weekly with the coordinates of every new model that is released. Each protein structure entry (which may contain details for more than one polypeptide chain) is assigned a four character ID starting with a number.

Note that the crystallographers and other experimentalists who submit structures to the PDB sometimes only work with domains and may modify the sequence to aid crystallization etc. This means that you need to map the PDB chain sequence back onto the original SwissProt entry to decide whether you are looking at a whole protein or just a domain and to find out whether it has been modified. Such information can be found in the DBREF and SEQADV fields of the PDB file and, failing that, there are other databases that can help eg 3Dseq at the EBI.

Both the PDB and its sequence equivalents (Genbank/EMBL/DDBJ) have to deal with large amounts of data that is donated by the scientific community. Because of this, there tends to be a lot of variation in the quality of annotation provided and a certain amount of redundancy.

For protein sequence, there is a secondary database, **SwissProt** (http://www.ebi.ac.uk/swissprot/), which is non-redundant and more consistently annotated. This is the best place to look for native protein sequences and associated annotation. Failing this, you will need to look either at TrEMBL (automatically compiled database of protein

sequences not yet added to SwissProt) or at the primary Genbank/EMBL/DDBJ entries.

The next levels of information processing often involve manual inspection of the data and expert input. The way in which the data is organized and presented can very greatly.

Some databases have a hierarchical structure and can classify their entries in different ways eg…

SCOP (Structural classification of proteins - http://scop.mrc-lmb.cam.ac.uk/scop/)...
class of structure
types of fold
superfamilies (probable homology)
families (homologous)
protein domain
species

CATH (http://www.biochem.ucl.ac.uk/bsm/cath_new/)...
Class
Architecture
Topology
Homologous superfamily
sequence family

These two databases are, in part, manually compiled so it may take a little while to be updated after the structures are released to the PDB - there are fully automated databases of structural classification around (eg FSSP - http://www.ebi.ac.uk/dali/fssp/) updating regularly from new PDB structures but these are less informative.

Databases can also contain entries that have been grouped according to sequence or structure similarity.

PFAM (http://www.sanger.ac.uk/Software/Pfam/) clusters homologous regions of sequences into families using sensitive Hidden Markov Model based methods (HMM) to indicates common sequence domains. It does not pay much attention to structural information in generating the alignments (often no coordinates for a protein are available) and the sequence domains do not necessarily correlated with clear structural domains, but it is well maintained and comprehensive. Pfam keeps a separate family for each domain, therefore, some families include the whole protein length but many multi domain proteins have different regions included in different families. For the main (Pfam-A) part of the database, the main alignments (seed alignments) are manually checked and corrected, however, the full alignments are computer generated and are less reliable. Other motifs are held as Pfam-B families, these are automatically clustered sequences derived from the ProDom database, but tend to be less well annotated and less useful.

Pfam is now linked into a ‘hub’ database called **Interpro** (http://www.ebi.ac.uk/interpro/) that shares annotation between a number of databases (Pfam, PRINTS, PROSITE, ProDom,

SMART, SWISS_PROT + TrEMBL) and is a useful starting point when investigating proteins.

Homstrad (http://www-cryst.bioc.cam.ac.uk/~homstrad/) is the structural equivalent of Pfam. Like other higher order databases, it relies heavily on data from the other depositories mentioned here. Primary protein structures are taken from the PDB, candidate families are routinely identified by searching Pfam, SCOP structural domain definitions are used and information on the native proteins is collected from SwissProt, Pfam and Interpro.

What is Homstrad ?

Homstrad is a web accessible database that contains families of proteins of known structure that share sequence/structural similarity.

New PDB entries are automatically processed and added as 'single member families' on a weekly basis although it may take longer for them to be incorporated into a main family alignment since this requires significant manual input.

The main families are composed of representative members only -When the family is first made, all the potential new sequences are compared and if any are more than 90% similar at the sequence level they are grouped together and the structure with the best resolution is selected as a representative. These representatives are what you see on the web page although you can view the full list of PDB chains considered by clicking on 'show related pdbs'. Note that Homstrad does not include theoretical models (these may be inaccurate), or structures containing only Cα coordinates (no environment can be assigned) and X-ray structures are preferred over NMR structures.

Each family has its own web page. On each page, some basic information is given in the top fields, then details for the individual PDB chain entries are listed, followed by a list of links to Pfam and other databases (links related to individual PDB chains - SwissProt, SCOP, CATH, FSSP, RCBI and EBI etc. - are available by clicking on each PDB code) Other related information is provided below this, and finally, a line up of the sequences is displayed.

The line-ups and associated annotation are a distinguishing feature of the database. These are based, not only on sequence similarity and type of amino acid, but also on the environment of each amino acid - its solvent accessibility, what type of secondary structure etc.
Annotation is provided using a program called Joy, which highlights a number of features including…

- What sort of secondary structure an amino acid is in
- Relative side chain accessibility (7% cut off, below this the amino acid is considered buried)
- If the amino acid is hydrogen bonded to the main chain amide or carbonyl (hydrogen bonding to other sidechains is also indicated, but only in the postscript output)
- Disulphide bonds
- Positive phi torsion angle

The protein structures themselves can be viewed, singly or with the whole family superimposed on one another, using Rasmol - see the information pages on the web for notes on setting this up. Once you have configured your web browser, click on the symbols next to each PDB code to see the individual structures or click on 'Rasmol' for the superposition.

In a number of cases, the Homstrad family page provides additional information (these families are part of what is known as Homstradplus)…

- Other sequences without structures but with homology may be viewed lined up with the main alignment (these additional members are not annotated with Joy). Currently, the additional sequences are taken from the seed alignment of a Pfam entry although another approach, using PSI-Blast is being implemented. Click on "show homologous sequences" to get this view. This feature was added by Paul de Bakker and is being integrated into the update procedures at the moment.
- Prosite motifs are sometimes indicated in this full alignment with each conserved amino acid coloured according to how well conserved/functionally important it is, red = very, through green, blue (very low probability) and then black is 'n ot considered' . See the key on Homstrad pages with PS (Prosite) links for more details. This is Ricardo Nunez Miguel's work and also needs integrating into our procedures.
- Another additional feature, added by Jiye Shi, shows the evolutionary trace analysis. This allows you to predict functionally important residues that have been conserved – click on 'show evolutionary trace'. Again, it needs to be integrated fully into Homstrad.

Searching Homstrad

Hosmstrad families can be searched with a keyword or with an amino acid sequence or multiple alignment.

Keyword searches can be done from the home page or the Homstrad search page. If you are interested in, for example, EGF proteins, type 'EGF' (make sure there are no spaces on either side) and a list of families containing that term in their description line will be returned. If you have a particular structure in mind and you know the PDB code, type that in and you will get a list of families that include either any of the chains detailed in that PDB entry or representatives of them (ie another PDB chain that shares more than 90% sequence identity). Searching for a PDB chain in this way may also give you access to single member families - these family names start with 'hs' or 'hsd' followed by the PDB code and in some instances a chain code letter (a,b,c etc) and perhaps a number to label the domain. These may not have been incorporated into a main family yet, either because the entry has not been fully processed by us, or because it shows no sequence or structure similarity to any other PDB chains already held in Homstrad. Such entries may or may not have been manually inspected and annotated but following links to Pfam etc can provide you with additional information.

You can also browse the families arranged alphabetically or structurally (click on the browse botton on the home page). The structural classes are broadly based on SCOP's structural classes (although there are some differences) and protein families are assigned to these groups

manually.
The classes are…
Small
Small disulphide
All alpha
Membrane bound all alpha
All beta
Membrane bound all beta
Alpha plus beta (beta sheets are mainly antiparallel)
Alpha beta (beta sheets are mainly parallel)
Alpha beta barrel

If you want to compare an amino acid sequence to existing protein families in Homstrad, one option is to do a quick blast search by either pasting your sequence (Fasta format) into the search form provided, or by specifying a PDB code and chain identifier in the form.
So the results for the PDB chain, 5grt, which has the sequence…

>5grt_ mol:protein length:461 Glutathione Reductase
VASYDYLVIGGGSGGLESAWRAAELGARAAVVESHKLGGTCVNVGCVPKKVMWN
TAVHSEFMHDHADYGFPSCEGKFNWRVIKEKRDAYVSRLNAIYQNNLTKSHIEIIRG
HAAFTSDPKPTIEVSGKKYTAPHILIATGGMPSTPHESQIPGASLGITSDGFFQLEELPG
RSVIVGAGYIAVEMAGILSALGSKTSLMIRHDKVLRSFDSMISTNCTEELENAGVEVL
KFSQVKEVKKTLSGLEVSMVTAVPGRLPVMTMIPDVDCLLWAIGRVPNTKDLSLNK
LGIQTDDKGHIIVDEFQNTNVKGIYAVGDVCGKALLTPVAIAAGRKLAHRLFEYKED
SKLDYNNIPTVVFSHPPIGTVGLTEDEAIHKYGIENVKTYSTSFTPMYHAVTKRKTKC
VMKMVCANKEEKVVGIHMQGLGCDEMLQGFAVAVKMGATKADFDNTVAIHPTSS
EELVTLR

…are as follows (not all of the alignments are shown to save space) - note that hits to both the main Homstrad families and to single-member families are detailed:

BLAST interface to HOMSTRAD

5grt_
Searching families with alignment
Searching single-member families

Searching families with alignment

```
BLASTP 2.2.1 [Apr-13-2001]
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= 5grt_  mol:protein length:461   Glutathione Reductase (461 letters)
Database: homstrad 3003 sequences; 627,733 total letters
Searching.......done
                                                              Score     E
Sequences producing significant alignments:                   (bits)  Value
3grs (grs)                                                       922  0.0
1gera (grs)                                                      460  e-131
```

```
2tpra (grs)                                                    265  2e-72
1ndaa (grs)                                                    261  4e-71
3lada (grs)                                                    178  2e-46
1ebda (grs)                                                    175  2e-45
1lpfa (grs)                                                    171  3e-44
1lvl (grs)                                                     141  4e-35
1ojt (grs)                                                     134  5e-33
1trb (grs)                                                      46  2e-06
1npx (grs)                                                      40  9e-05
1chua (FAD_binding_2)                                           28  0.50
1chua (FAD_binding_2)                                           28  0.50

>3grs_grs
          Length = 461
 Score =  922 bits (2382), Expect = 0.0
 Identities = 459/461 (99%), Positives = 459/461 (99%)

Query: 1   VASYDYLVIGGGSGGLESAWRAAELGARAAVVESHKLGGTCVNVGCVPKKVMWNTAVHSE 60
           VASYDYLVIGGGSGGL SA RAAELGARAAVVESHKLGGTCVNVGCVPKKVMWNTAVHSE
Sbjct: 1   VASYDYLVIGGGSGGLASARRAAELGARAAVVESHKLGGTCVNVGCVPKKVMWNTAVHSE 60

Query: 61  FMHDHADYGFPSCEGKFNWRVIKEKRDAYVSRLNAIYQNNLTKSHIEIIRGHAAFTSDPK 120
           FMHDHADYGFPSCEGKFNWRVIKEKRDAYVSRLNAIYQNNLTKSHIEIIRGHAAFTSDPK
Sbjct: 61  FMHDHADYGFPSCEGKFNWRVIKEKRDAYVSRLNAIYQNNLTKSHIEIIRGHAAFTSDPK 120
.
.
.
.
>1gera_grs
          Length = 448
 Score =  460 bits (1184), Expect = e-131
 Identities = 252/462 (54%), Positives = 318/462 (68%), Gaps = 20/462 (4%)

Query: 4   YDYLVIGGGSGGLESAWRAAELGARAAVVESHKLGGTCVNVGCVPKKVMWNTAVHSEFMH 63
           YDY+ IGGGSGG+ S  RAA  G + A++E+ +LGGTCVNVGCVPKKVMW+ A   E +H
Sbjct: 3   YDYIAIGGGSGGIASINRAAMYGQKCALIEAKELGGTCVNVGCVPKKVMWHAAQIREAIH 62

Query: 64  DHA-DYGFPSCEGKFNWRVIKEKRDAYVSRLNAIYQNNLTKSHIEIIRGHAAFTSDPKPT 122
            +  DYGF +   KFNW  +   R AY+ R++  Y+N L K+++++I+G A F  D K T
Sbjct: 63  MYGPDYGFDTTINKFNWETLIASRTAYIDRIHTSYENVLGKNNVDVIKGFARFV-DAK-T 120
.
.
.
.
>1chua_FAD_binding_2
          Length = 478
 Score = 28.1 bits (61), Expect = 0.50
 Identities = 10/22 (45%), Positives = 18/22 (81%)

Query: 294 GHIIVDEFQNTNVKGIYAVGDV 315
           G ++VD+   T+V+G+YA+G+V
Sbjct: 303 GGVMVDDHGRTDVEGLYAIGEV 324

  Database: homstrad
    Posted date:  Sep 21, 2001  5:31 PM
  Number of letters in database: 627,733
  Number of sequences in database:  3003
```

```
Lambda     K      H
   0.318    0.134    0.395
Lambda     K      H
   0.267   0.0410    0.140
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 596,375
Number of Sequences: 3003
Number of extensions: 25079
Number of successful extensions: 151
Number of sequences better than  1.0: 13
Number of HSP's better than  1.0 without gapping: 11
Number of HSP's successfully gapped in prelim test: 2
Number of HSP's that attempted gapping in prelim test: 86
Number of HSP's gapped (non-prelim): 17
length of query: 461
length of database: 627,733
effective HSP length: 81
effective length of query: 380
effective length of database: 384,490
effective search space: 146106200
effective search space used: 146106200
T: 11
A: 40
X1: 16 ( 7.3 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (21.7 bits)
S2: 59 (27.3 bits)
```

Searching single-member families

```
BLASTP 2.2.1 [Apr-13-2001]
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= 5grt_  mol:protein length:461  Glutathione Reductase (461 letters)
Database: HOMS 2341 sequences; 410,612 total letters

Searching.....done
                                                                  Score     E
Sequences producing significant alignments:                       (bits)  Value

hs1d7ya crystal structure of nadh-dependent ferredoxin reductase..32  0.015
hs1hyua crystal structure of intact ahpf                          29  0.15

>hs1d7ya crystal structure of nadh-dependent ferredoxin reductase,
           bpha4
          Length = 401
 Score = 32.3 bits (72), Expect = 0.015
 Identities = 53/223 (23%), Positives = 90/223 (39%), Gaps = 35/223 (15%)

Query: 106 IEIIRGHAAFTSDPKPTIEVSGKKYTAPH--ILIATGGMPS---TPHESQIPGASLGITS 160
           +E + G  A + DP+          T P+  +++ATG  P    T   + +P  +L
Sbjct: 70  VEWLLGVTAQSFDPQAHTVALSDGRTLPYGTLVLATGAAPRALPTLQGATMPVHTLRTLE 129
.
```

```
.
.
.
Database: HOMS
Posted date:  Sep 30, 2001 11:30 AM
Number of letters in database: 410,612
Number of sequences in database:  2341
Lambda     K      H
   0.318    0.134    0.395
Lambda     K      H
   0.267   0.0410    0.140
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 378,534
Number of Sequences: 2341
Number of extensions: 15326
Number of successful extensions: 51
Number of sequences better than  1.0: 4
Number of HSP's better than  1.0 without gapping: 2
Number of HSP's successfully gapped in prelim test: 2
Number of HSP's that attempted gapping in prelim test: 45
Number of HSP's gapped (non-prelim): 8
length of query: 461
length of database: 410,612
effective HSP length: 77
effective length of query: 384
effective length of database: 230,355
effective search space: 88456320
effective search space used: 88456320
T: 11
A: 40
X1: 16 ( 7.3 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (21.7 bits)
S2: 57 (26.6 bits)
```

This tells you that a representative of 5grt (3grs – it looks identical) is in the family grs along with several other homologous PDB chains and you can click on the family to take you to its home page. Here you find that the family description is ‘pyridine nucleotide-disulphide oxidoreductases class-I’. 11 structures are part of the family that share an average sequence identity of 30%, there are 2 links to Pfam and Homstradplus information is available. The weak hit to the FAD_binding_2 family turns out to be over only a 22 amino acid motif and may be worth following up as a functionally important region. Two single-member families are detected. Despite their lower similarity to 5grt than other members of the grs family (look at the E value – a smaller E value means higher similarity), both share a Pfam link with the grs family so they may at some point be added to this family if the structures are not too dissimilar. To get to their home pages either click on the hs1d7ya/hs1hyua links or type 1d7y or 1hyu (the PDB codes) in the Homstrad keyword search box.

Fugue can be used to generate more accurate sequence/structure based alignments between your sequence and the families identified in the Blast search. To do this, click ‘align’ in the top left hand corner of the family page and submit your query sequence.

If you want to search Homstrad using Fugue (and not just align sequences) for detecting more distant homologs, click on 'Fugue' on the Homstrad home page, which takes you to the Fugue search page. Fugue operates using environment-specific substitution tables and structure-dependent gap penalties. Environment-specific substitution tables have been derived from selected Homstrad families dictating the degree of conservation for each type of amino acid, i.e. if it is buried and in an alpha helix it is more likely to be conserved that if it is in a surface loop. Then, for each Homstrad family, a scoring matrix (profile) has been calculated using these substitution tables. Fugue lines the new sequence up to the existing family according to this scoring matrix and allows you to…
a) identify distant homologies that may not be identified in other ways.
b) give some indication as to the shape the new protein may adopt, functional sites etc. as a guide for bench experimental work.
To run the search, either upload a sequence file or paste your amino acid sequence (in FASTA or pure amino acid format) into the box provided. By default, other homologous sequences are collected using PSI-Blast and lined up with the corresponding Homstrad family and your query sequence. If you wish, you can submit your own multiple sequence alignment as the input (FASTA / NBRF / CLUSTAL / MSF format). Results are sent out by e-mail.
So for the 5grt example you get the following e-mail response….

```
##########################################################################
#  FUGUE  v1.s.16 (JAN 2001)
# Search sequence(s) against fold library using environment-specific
# substitution tables and structure-dependent gap penalties.
#
#Fold library and substitution tables are based on the HOMSTRAD database.
#http://www/cryst.bioc.cam.ac.uk/~homstrad/
#
#  FUGUE server is available at:
#         http://www-cryst.bioc.cam.ac.uk/~fugue/
#         http://www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html
#  Citation:   J. Shi, T. L. Blundell and K. Mizuguchi
#              (manuscript in preparation)
#  Size of fold library: 3157
#  Probe sequence ID   : 5grt
#  Probe sequence len  : 461
#  Probe divergence    : 0.657
#  Recommended cutoff  : ZSCORE >=    6.0  (CERTAIN   99% confidence)
#  Other cutoff        : ZSCORE >=    5.0  (LIKELY    95% confidence)
#  Other cutoff        : ZSCORE >=    4.7  (MARGINAL  90% confidence)
#  Other cutoff        : ZSCORE >=    3.5  (GUESS     50% confidence)
#  Other cutoff        : ZSCORE <     3.5  (UNCERTAIN)
#
#  PLEN   : Profile length
#  RAWS   : Raw alignment score
#  RVN    : (Raw score)-(Raw score for NULL model)
#  ZSCORE : Z-score normalized by sequence divergence
#  PVZ    : P-value based on Z-score jumbling        (Currently Disabled)
#  ZORI   : Original Z-score (before normalization)
#  EVP    : E-value based on profile calibration     (Currently Disabled)
#  EVF    : E-value based on library search          (Currently Disabled)
#  AL     : Alignment algorithm used for Zscore/Alignment calculation
#           0 -- Global, 2 -- GloLocSeq (No sequence termini gap penalty)
#           3 -- GloLocPrf (No profile termini gap penalty)
```

```
#-------------------------------------------------------------------------
#  Profile        PLEN   RAWS RVN   ZSCORE    PVZ      ZORI     EVP      EVF    AL
#-------------------------------------------------------------------------
 grs              558    610 1282    65.75 1.0E+03    67.06 1.0E+03 1.0E+03 00
 hs1d7ya          401     20 404     15.47 1.0E+03    16.79 1.0E+03 1.0E+03 00
 hsd2tmda3        233     -5 146     10.36 1.0E+03    11.67 1.0E+03 1.0E+03 22
 FAD_binding_2    590   -477 226      8.72 1.0E+03    10.04 1.0E+03 1.0E+03 00
 cox              661   -737 124      7.48 1.0E+03     8.80 1.0E+03 1.0E+03 00
 hs1b3ma          385   -261 224      7.31 1.0E+03     8.63 1.0E+03 1.0E+03 00
 AlaDh_PNT        397   -303 108      4.83 1.0E+03     6.14 1.0E+03 1.0E+03 00
 hsd1cjca2        230    -79  79      4.63 1.0E+03     5.94 1.0E+03 1.0E+03 22
 hsd1fcda1        186   -213  79      4.56 1.0E+03     5.88 1.0E+03 1.0E+03 02
 hs1qlaa          655   -593  95      4.32 1.0E+03     5.64 1.0E+03 1.0E+03 00
```

If you go to the results website you get more information and access to the alignments….

FUGUE v1.s.16 (JAN 2001)

```
Search sequence(s) against fold library using environment-specific
substitution tables and structure-dependent gap penalties.
 Fold library and substitution tables are based on the HOMSTRAD database.
        http://www-cryst.bioc.cam.ac.uk/~homstrad/
 FUGUE server is available at:
        http://www-cryst.bioc.cam.ac.uk/~fugue/
        http://www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html
 Citation:     J. Shi, T. L. Blundell and K. Mizuguchi
            (manuscript in preparation)

 Size of fold library: 3157
 Probe sequence ID   : 5grt
 Probe sequence len  : 461
 Probe divergence    : 0.657
 Recommended cutoff  : ZSCORE >=    6.0  (CERTAIN   99% confidence)
 Other cutoff        : ZSCORE >=    5.0  (LIKELY    95% confidence)
 Other cutoff        : ZSCORE >=    4.7  (MARGINAL  90% confidence)
 Other cutoff        : ZSCORE >=    3.5  (GUESS     50% confidence)
 Other cutoff        : ZSCORE <     3.5  (UNCERTAIN)

 PLEN   : Profile length
 RAWS   : Raw alignment score
 RVN    : (Raw score)-(Raw score for NULL model)
 ZSCORE : Z-score normalized by sequence divergence
 ZORI   : Original Z-score (before normalization)
 AL     : Alignment algorithm used for Zscore/Alignment calculation
          0 -- Global, 2 -- GloLocSeq (No sequence termini gap penalty)
          3 -- GloLocPrf (No profile termini gap penalty)
 The sequence(s) you submitted is HERE (in original format).
 The sequence(s) actually used by FUGUE is HERE (in PIR format).
 Download all the results in compressed format HERE.  new!
```

View Ranking (Click on a profile hit will bring you to the corresponding HOMSTRAD family)

Profile Hit	PLEN	RAWS	RVN	ZSCORE	ZORI	AL		
grs	558	610	1282	65.75	67.06	00	CERTAIN	Alignment
hs1d7ya	401	20	404	15.47	16.79	00	CERTAIN	Alignment
hsd2tmda3	233	-5	146	10.36	11.67	22	CERTAIN	Alignment
FAD_binding_2	590	-477	226	8.72	10.04	00	CERTAIN	Alignment

cox	661	-737	124	7.48	8.80	00	CERTAIN	Alignment
hs1b3ma	385	-261	224	7.31	8.63	00	CERTAIN	Alignment
AlaDh_PNT	397	-303	108	4.83	6.14	00	MARGINAL	Alignment
hsd1cjca2	230	-79	79	4.63	5.94	22	GUESS	Alignment
hsd1fcda1	186	-213	79	4.56	5.88	02	GUESS	Alignment
hs1qlaa	655	-593	95	4.32	5.64	00	GUESS	Alignment

View Alignments (Keys [aa,ma,mh,hh])

Hint: check 'ma' first if your query is a single sequence, otherwise start with 'aa'.

Profile Hit	HTML				POSTSCRIPT				TEXT(PIR FORMAT)						
grs	aa	ma	mh	hh	aa	ma	mh	hh	aa	ma	mh	hh	CERTAIN	Model	65.75
hs1d7ya	aa	ma	mh	hh	aa	ma	mh	hh	aa	ma	mh	hh	CERTAIN	Model	15.47
hsd2tmda3	aa	ma	mh	hh	aa	ma	mh	hh	aa	ma	mh	hh	CERTAIN	Model	10.36
FAD_binding_2	aa	ma	mh	hh	aa	ma	mh	hh	aa	ma	mh	hh	CERTAIN	Model	8.72
cox	aa	ma	mh	hh	aa	ma	mh	hh	aa	ma	mh	hh	CERTAIN	Model	7.48
hs1b3ma	aa	ma	mh	hh	aa	ma	mh	hh	aa	ma	mh	hh	CERTAIN	Model	7.31
AlaDh_PNT	aa	ma	mh	hh	aa	ma	mh	hh	aa	ma	mh	hh	MARGINAL	Model	4.83
hsd1cjca2	aa	ma	mh	hh	aa	ma	mh	hh	aa	ma	mh	hh	GUESS	Model	4.63
hsd1fcda1	aa	ma	mh	hh	aa	ma	mh	hh	aa	ma	mh	hh	GUESS	Model	4.56
hs1qlaa	aa	ma	mh	hh	aa	ma	mh	hh	aa	ma	mh	hh	GUESS	Model	4.32

Keys
aa: query sequences (including PSI-BLAST homologues) aligned against all the representative structures from a HOMSTRAD family
ma: master sequence aligned against all the representative structures from a HOMSTRAD family
mh: master sequence aligned against a single structure of highest sequence identity from a HOMSTRAD family
hh: single sequence/structure pair with highest sequence identity in 'aa'
Note: If your query is a single sequence, master sequence is equivelent to your query and all the other sequences (if any) are collected by PSI-BLAST.
If your query is a sequence alignment, master sequence is set to the first sequence in the alignment.
JOY Keys are described here

In this instance, Fugue has identified a number of hits not found using the simple Blast search and, additionally, the confidence level of these hits is given in simple terms – a translation of the Z score. Generally it is best to consider just the ‘CERTAIN’ hits although ‘LIKELY’ hits (none have been classified this in the current example) may be relevant. To view the multiple alignments, in the ‘view alignments’ section, first click on the HTML ‘ma’ buttons to see your sequence lined up with each of the Homstrad family alignments. Then click the ‘aa’ buttons, which gives you the lineup including all the additional homologs collected by PSI-Blast. The ‘aa’ alignments are additionally annotated to highlight different types of amino acid using ‘Taylor’ notation. Fugue also creates a rough model for your query sequence based on the backbone coordinates of the most significant hit, although in this instance the query sequence’s structure has already been determined.

Examples…

There are a number of published examples of how Homstrad has been used including…

Nunez Miguel et al. (2001) 'Protein Fold Recognition and Comparative Modelling using Homstrad, Joy and Fugue' (This book chapter is in Press – [local link]).

Shirai et al. (2001) A novel superfamily of enzymes that catalyze the modification of guanidine groups. TIBS 26 (8) 465-468

Parker et al. (2001) A family of proteins related to Spätzle, the toll receptor ligand, are encoded in the *Drosophila* genome. Proteins. 45 (1) 71-80.

A final word…

As with all higher order protein databases there are different ways to organize the data and different places to specify cut-offs.

At present, the cut off point at which we define a protein as being homologous to another protein is variable. Sometimes, whether a protein is added to a family depends as much on whether the structures look similar as on sequence similarity (although there must be some implied evolutionary relationship between the two proteins to justify the term 'homology').

We occasionally have problems deciding whether a protein should belong to one family or another if a continuous chain of homology has developed between them. Sometimes we decide that the families should be merged into one and in other cases we leave them separate. At present what we do tends to be guided by PFAM and SCOP but the decision is ultimately arbitrary and is in no way the final word on protein grouping.

Finally, you will notice that membrane spanning protein domains are significantly under-represented in Homstrad and other protein structure databases, despite their great importance to researchers (as key regulators of signaling pathways, drug targets etc). This we can do nothing about, but the situation should improve in the near future as experimental techniques improve.