In order to obtain more information on a particular protein or domain, several homologous members of a family, may be aligned and the consensus features extracted and represented in a pattern. This pattern may be used to trawl the database for further examples of a particular region.

Further information may be used in the search, if the pattern represents a scoring matrix for the consensus region. The matrix records the occurrence of each residue in each position, and uses this information to mine the database. Thus additional sensitivity is given to the search and these matrices are known as profiles.

Hidden Markov models offer a further way of identifying distant homologues of a protein. They rely on heavy statistical calculation of the probability of the occurrence of each residue in a specific position. HMMs also calculate a statistical probability for the occurrence of an insertion - or deletion - and calculations are based, not only on the seed alignment, but also on the likelihood of certain residues occurring in a consecutive fashion.
Protein pattern profiling

A protein sequence can be obtained from gene predictions, papers, database matches or even peptide sequencing. On obtaining a protein sequence the next step is to identify it and find out something about its function. Bioinformatics methods can be used to support or refute any preliminary ideas about the physiological function. It is important to use and critically evaluate information from different sources (including the nucleotide analysis), all of which will aid the determination of possible function.

There may be several clues to the function of at least part of the protein in the form of distinctive patterns or domains associated with a specific protein function.

Protein sequence analysis is often more accurate than nucleotide sequence analysis because:

- It usually contains a higher signal to "junk" ratio.
- The 3-D structure of similar proteins may be known.
- Evolutionary relationships are sometimes more visible.
- Annotation of protein sequence and related databases is often comprehensive.

Protein Sequence Analysis

When using bioinformatics tools, every option of a program should be understood, together with information on what the program is really doing and what the output means. If a protein sequence is novel, the first sensible step would be to carry out a database similarity search, to identify any similar sequences. There may be sequences with a high incidence of identity to the novel protein, indicating possible homology. If these hits are well annotated, laboratory experiments should prove or disprove the relationship.

Should no definite identity result from a database search, or the matching hits are not well annotated, a multiple sequence alignment of these regions may then identify conserved residues that could characterise a specific functional domain. Alternatively, the consensus sequence may also indicate distinctive patterns that are known to relate to domains and motifs in characterised proteins.
If neither of these approaches produces conclusive evidence of the function of a particular protein, a more flexible search must be implemented.

A pattern is created from the consensus of a multiple sequence alignment using a specific "pattern language". This involves encasing various one-letter residue codes in a variety of parentheses to define the content of a pattern sequence. Thus, using the alignment produced using fruit.fasta:

```
CHE--RRI----ES
-P---ACH---ES
GRB--ENAPPLES
CLEMENTIN---
```

A pattern representing the occurrence of amino acid residues in specific positions may look like this:

```
E-X (3,6)-E-S
```

indicating the occurrence of a conserved Glutamate followed by three to six positions which may be occupied by any amino acid. The final two positions are occupied by conserved Glutamate and Serine residues. The residues in the first two positions of these sequences have been left out of the pattern, because there is no particular conservation.

This search is more powerful than a simple BLAST search, for instance, as the unconserved residue may cause the score of the hit to dip below the threshold required for an HSP. A pattern search may indicate other proteins with this particular sequence conservation and these may have been well characterised.

As with other aspects of bioinformatics, biological phenomena must be converted into mathematical problems for the computer to deal with effectively. Patterns represent a certain flexibility in the search for corresponding regions, but they do not contain all the information. The addition of scoring or weight matrices to the initial pattern confers more sensitivity on a search. This type of pattern is termed a Profile.

PRINTS, is a curated database and looks for a series of short motif patterns to determine the functionality of a protein, and thus it's family. Small motifs are identified across the whole of the protein, and a sequence of these motifs is known as a fingerprint. The fingerprint can then be used to identify other members of the same family, including any novel proteins.

---

1 The standard pattern language may be found with the Prosite Database (http://ca.expasy.org/prosite/).
2 Some software applications, such as GCG use a different pattern language.

---
It is very rare that a protein belonging to a particular group does not have all trademark motifs in the order suggested by the PRINTS database.

The Prosite database\textsuperscript{iii} contains over 1,500 protein families and domains, derived from multiple sequence alignments of protein family groups. Patterns are intentionally kept as short as possible in order to confer both high sensitivity and high specificity. A search of this database with an unknown sequence may reveal specific regions of the novel polypeptide to contain a known region of conservation. Annotation and selection of patterns is manual. These patterns are verified by using them to search the SwissProt database.

Short regions are known as signatures, and should be present in every variation of a domain with a specific function.

Prosite is free to all non-profit making organisations, although commercial use of the data requires a license, for which a fee is payable.

Over 300,000 domain families are held in the fully automated ProDom\textsuperscript{iv} database. They were built using PSI-BLAST using an initial seed alignment, from sequences held in the SwissProt and TrEMBL databases.

Further flexibility can be conferred on an initial pattern by using statistics to predict the likelihood of occupation of a particular position in a series by either a residue, or a gap. The sequence information produced here is called a Hidden Markov Model\textsuperscript{v} and can be used to search and group domain families.

Together with families derived from profiles, the Pfam\textsuperscript{vi} database, held at the Sanger Centre in Cambridge hold protein domain families derived from Hidden Markov Models.

\textbf{Motifs and Domains}

These conserved regions are generally known as motifs and domains, and contribute to the functionality of a polypeptide.

Sequence motifs and domains are derived from sequence alignments of related proteins. The distinction between a “motif” and a “domain” is not clear-cut, but the term “motif” is generally used to refer to a short sequence pattern and “domain” to a longer region of sequence similarity. Domains are almost always represented by statistical models (known as profiles) which describe the probability of finding each amino acid at each position in the domain.
Practical

Motifs

There is one main database of protein motifs: PROSITE. The sequence motifs in this database are described using a simple pattern description language. They include some very common, simple motifs, many only a few residues long, that indicate possible sites for post-translational modifications (e.g. glycosylation or phosphorylation).

Go to http://www.expasy.ch and select "PROSITE" from the "Databases" section. Select "ScanProsite" from the "Tools for PROSITE" section of the page, followed by "scan a sequence for the occurrence of PROSITE patterns". Enter pax6_human in the top entry field and tick the "exclude patterns with a high probability of occurrence". Use the "Start the scan" button to start the search. It should only take a few seconds, then examine the output.

The results should take only a few seconds to be returned, and you should see that only two patterns have been returned. The homeobox, and the paired box signatures both appear in the results section, together with the respective fragments of sequence found in your query. Following the link to the number beginning with "PS" will show you the prosite entry for each domain. The link starting with "PDOC" will display entries to give you more information about the domain, and proteins containing it.

Type of nuclear binding in Homeobox domain ......................................................

Homeobox domain consensus pattern .................................................................

Re-run the query, but this time, deselect the "exclude patterns with a high probability of occurrence" option. How many domains are you now presented with? Why?

3 This type of search will include such domains as glycosylation, or myristoylation sites, which, although biologically relevant to your protein, may not offer as many clues about the function of it.
A PROSITE search is also possible from within the EMBOSS suite of programs.

Select patmatmotifs from the Jemboss scroll menu, or the "Protein" and "motifs" menus. Drag and drop the protein sequence into the filename field and hit "Go".

This should give you exactly the same information as you obtained using the Prosit web page.

You should already know from the database entries that PAX6 is a DNA binding protein. Many (but by no means all) of these proteins bind DNA with a simple structural motif consisting of two alpha helices joined by a short loop (the so-called helix-turn-helix motif). A simple EMBOSS utility, helixturnhelix, can be used to look for the pattern of amino acid residues associated with this motif.

Select helixturnhelix from the Jemboss scroll menu, or from the "Protein/2D structure" menus. Drag your protein sequence into the sequence filename field, accept the defaults and press "Go".

Your results should display one fragment of the sequence indicated by the program to be a possible helix turn helix motif.

#==============================================
# # Sequence: PAX6_HUMAN from: 1 to: 422
# # HitCount: 1
# # Hits above +2.50 SD (972.73)
# #==============================================

Maximum_score_at at "*

(1) Score 1109.000 length 22 at residues 238->259

Sequence: FARERLAALKIDLPEDIQVWF
  238  259
Standard_deviations: 2.96

#==============================================
#==============================================

Have a look at the predicted helix turn helix motif. Does it correspond to the pattern you identified as a homeobox domain in prosite?

---

4 It should not be exactly the same, as the homeobox domain is obviously made up of much more than just one motif, but you should find the HTH pattern within this domain.
Domains

There are many protein domain databases available on the Web. One of the most widely used is Pfam, which is available at the Sanger Centre. Pfam is a collection of protein families and domains, and contains multiple protein alignments and profile-HMMs of these families. Over 50% of the entries in SwissProt have a match to one of the families in Pfam.

There are two types of protein family in Pfam:

**Pfam-A** families are curated; a seed alignment of representative sequences is inspected. A profile (see later) representing this seed alignment is used to search SwissProt and matching sequences are automatically aligned to the seed alignment to create the full alignment. There is usually extensive annotation for a Pfam-A family.

**Pfam-B** contains sequences that are not contained in Pfam-A. The alignments are produced completely automatically and may not be of such high quality as the Pfam-A alignments. The alignments used here come straight from the ProDom database.

Go to [http://www.sanger.ac.uk/Software/Pfam](http://www.sanger.ac.uk/Software/Pfam) and select “Protein Search”. Enter the SwissProt identifier `pax6_human` into the top field (the query box) and press the “Submit Query” button to start the search.

You should see the two domains you discovered using Prosite have now been picked out in Pfam.

Start and End of PAX domain

Start and End of homeobox domain

Are the regions, which these domains cover the same as you discovered in prosite?

Examine each domain entry by clicking on the domain icon to find out more about it. Click on the “View Graphic” button in the “Domain organisation” box. This will give you a graphical list of all other SwissProt proteins containing these domains.
Another database that defines functional protein families in a similar way is PRINTS. In this database, each domain is identified by a number of short, particularly well-conserved sequences. A full match to one of these "fingerprints" will match all the relevant short sequences in the correct order. A partial match is recorded if some are missing or if they occur in an incorrect order.

The PRINTS database can be searched using the pscan program which is available within EMBOSS.

Select pscan from the Jemboss scroll menu, or from the "Protein" and "Motifs" menus. Type sw:pax6_human into the sequence filename box and click on the "LOAD SEQUENCE ATTRIBUTES" button to display the default values for this protein. Accept these values and press "Go".

CLASS 1
Fingerprints with all elements in order

Fingerprint PAIREDBOX Elements 4
   Accession number PR00027
   Paired box signature
   Element 1 Threshold 79% Score 100%
       Start position 8 Length 16
   Element 2 Threshold 60% Score 98%
       Start position 26 Length 19
   Element 3 Threshold 71% Score 96%
       Start position 46 Length 18
   Element 4 Threshold 61% Score 97%
       Start position 64 Length 18

CLASS 2
All elements match but not all in the correct order

Fingerprint HTHREPRESSR Elements 2
   Accession number PR00031
   Lambda and other repressor helix-turn-helix signature
   Element 1 Threshold 50% Score 52%
       Start position 239 Length 10
   Element 2 Threshold 32% Score 34%
       Start position 395 Length 17
   Element 2 Threshold 32% Score 61%
       Start position 248 Length 17

CLASS 3
Not all elements match but those that do are in order

Fingerprint HIVTATDOMAIN Elements 3
   Accession number PR00055
   HIV TAT domain signature
   Element 2 Threshold 55% Score 55%
       Start position 55 Length 8
   Element 3 Threshold 40% Score 45%
       Start position 270 Length 17
CLASS 4
Remaining partial matches

Fingerprint EUTPISMRASEI Elements 6
Accession number PR00416
Eukaryotic DNA topoisomerase I signature
Element 6 Threshold 34% Score 37%
   Start position 366 Length 12
Element 6 Threshold 34% Score 38%
   Start position 72 Length 12

These are the results you should have obtained. Motifs are separated into separate classes, depending on how well they match the query sequence. Class one consists of the most significant matches as it displays only those motifs which contain all elements in the same order as the query sequence.

How do the results of this scan compare with those from Prosite and Pfam? Are there any false positive matches?

Interpro

Interpro* is a tool that searches many pattern databases, to give you a consensus result from PROSITE, PRINTS, Pfam, SMART, BLOCKS, TIGRFAMs and ProDom to produce a powerful diagnostic tool for protein signatures.:

Go to http://www.ebi.ac.uk/interpro/ and follow the link to “Sequence Search”. Copy the peptide sequence of your pax6_human protein into the paste field. Select “interactive run” from the Query Mode option, and enter your email address. Then hit the “Submit” button.

The URL of your results will be emailed to you if you cannot wait for the application to run. Otherwise they will be available in a few minutes, and should give a graphical display of all functional regions defined in various databases. Each region found is represented by a block, indicating the length of the region relevant to the sequence entered into the query field.

The results indicate that two domains have been found - a homeobox and a paired box domain. Each domain has a list of entries from the various databases that Interpro scans. The databases are defined by the first two letters of the accession number.

<table>
<thead>
<tr>
<th>PR</th>
<th>PRints</th>
<th>PD</th>
<th>ProDom</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF</td>
<td>PFam</td>
<td>PS</td>
<td>ProSite</td>
</tr>
<tr>
<td>SM</td>
<td>SMart</td>
<td>IPR</td>
<td>InterPRo</td>
</tr>
</tbody>
</table>
Size of each Homeobox domain

Size of each Paired Box domain

Why do you think there may be discrepancies in the region which each database expects a particular domain to cover?

There are many more pattern and motif databases; a full analysis of these resources is outside the scope of this course. Each uses a different approach and contains a different collection of protein families. Some are particularly useful for a certain type of protein; for example, the SMART database specialises in domains that occur in signalling proteins. Therefore, it is always good practice to search a range of databases with your sequence.

You should have found that the databases agree that PAX6_HUMAN contains two distinct domains. The C-terminal domain is a homeobox (Prosite finds both characteristic homeobox motifs); the N-terminal one is a PAX domain, which contains the longer "paired box" motif.
References

1 Attwood, T.K. and Beck, M.E. (1994) Protein Engineering, 7 (7), 841-848
PRINTS - A protein motif fingerprint database

PRINTS and PRINTS-S shed light on protein ancestry.

The PROSITE database, its status in 1999


Maximum discrimination hidden Markov models of sequence consensus.

The Pfam Protein Families Database

The InterPro database, an integrated documentation resource for protein families, domains and functional sites.