Info on biological databases

1. Molecular Biology Database Collections
   • The first issue of each year of *Nucleic Acids Research* is devoted to articles on biological database issue.
   http://nar.oupjournals.org/

2. Cambridge HealthTech Institute (CHI)
   • Databases directory (don’t click on database tab)
   http://www.genomicglossaries.com/content/lifesciences_databasesdirectory.asp
What is a database?

1. computerized storehouse of data (records)
2. allows user-defined queries
3. allows extraction of specified records
4. allows adding, changing, removing, and merging of records
5. uses standardized formats
How to search for database records?

1. Identifiers
   - unique string of letters and digits that often are interpretable in a meaningful way by humans
   - in GenBank: **LOCUS** (e.g. SCU32124)
   - in SwissProt: **ENTRY_NAME** (YNT2_YEAST)
   - can change !!!
   - *** OTHER IDENTIFIERS ***
   - **VERSION** = extension increases after every change
   - **GI** = may change completely when seq. is altered
GenBank Record

LOCUS SCU49845 5028 bp DNA PLN 21-JUN-1999
DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION U49845
VERSION U49845.1 GI:1293613
KEYWORDS .
SOURCE baker's yeast.
ORGANISM Saccharomyces cerevisiae Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE 1 (bases 1 to 5028)
AUTHORS Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
TITLE Cloning and sequence of REV7, a gene whose function is required for DNA damage-induced mutagenesis in Saccharomyces cerevisiae
How to search for database records?

2. **Accession numbers**
   - in GenBank: 1 letter + 5 digits (U12345)
     or 2 letters + 6 digits (AF123456)
   - in SwissProt: 1 letter + 5 digits (P04049)
   - stable, does not change
   - will get transferred to new record if records are merged

*** USE ACCESSION NO. FOR SEARCHES ***
GenBank Record

Identifier

LOCUS SCU49845 5028 bp DNA PLN 21-JUN-1999
DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds.

ACCESSION U49845
VERSION U49845.1 GI:1293613
KEYWORDS .
SOURCE baker's yeast.
ORGANISM Saccharomyces cerevisiae Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE 1 (bases 1 to 5028)
AUTHORS Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
TITLE Cloning and sequence of REV7, a gene whose function is required for DNA damage-induced mutagenesis in Saccharomyces cerevisiae
Why we need electronic databases?

1. Data explosion
2. Data distribution

GenBank Status in 2001
~ 15 Mill. Sequence entries,

Why are there so many databases?

Classification schemes:

• Database design (relational, object-oriented DB)
• Accessibility (public, academic, commercial)
• Data entry (curator, automated)
• Primary or derived databases
• Data type (DNA, RNA, ESTs, Glycans, Proteins)
Types of protein databases

1. Sequence sequence databases
2. Protein motif databases
3. Protein structure databases
Protein sequence databases

1. Manually annotated and curated
   PIR-PSD, SwissProt

2. Automatically annotated
   Genpept, TrEMBL
Manually curated protein seq. DBs

Advantages:
- highest quality available
- FASTA sequence format
- input from GenBank, EMBL, DDBJ, Literature, individual labs

Disadvantages:
- expensive to maintain
- don’t have all of the latest sequence info
PIR International– PSD

- probably the oldest protein seq. database
- current Release 71.04, March 01, 2002, contains **283153 Entries**
- public, **downloadable**
- collaboration between PIR, MIPS and DJPID
- offer numerous DB search tools (FASTA, BLAST)
- uses FASTA sequence format
SwissProt

- collaboration b/n the European Bioinformatics Institute (EBI) and the Swiss Institute for Bioinformatics (SIB)
- numerous attached protein analysis tools including homology modeling
- uses FASTA format
- SWISS-PROT Release 40.12 of 05-Mar-2002: 105967 entries
- public, downloadable

Source:
Taxonomic distribution of entries in Swiss-Prot

- Eukaryota (52%)
- Bacteria (35%)
- Archaea (6%)
- Viruses (8%)

Taken from: http://us.expasy.org/sprot/relnotes/relstat.html
TrEMBL

- TrEMBL = Translations of EMBL (contains translations of all coding sequences in EMBL)
- is a computer annotated supplement to SwissProt
- TrEMBL Release 19.10 of 01-Mar-2002: 594148 entries
- Entries in TrEMBL and SwissProt do **NOT** overlap
- TrEMBL + SwissProt  594,148 + 105,967 = 700,115
- SP-TrEMBL have an assigned a Swiss-Prot accession #
- REM-TrEMBL (REMaining TrEMBL)

**Source:**
http://www.ebi.ac.uk/
Database connections

- DNA sequences
  - GenBank
    - Genome projects
      - Sequin
      - BankIt
    - Automatically translated CDS
      - GenPept
    - Protein seq. from ind.labs.
      - TrEMBL
        - Annotation / Curation
          - SwissProt
          - PIR-PSD
  - EMBL/EBI
    - DDBJ
## Errors in sequence databases

### Data entries:
- **Genome Projects:**
  - 1 in 10,000 nts
- **ESTs**
  - 1 in 100 nts

### Error rates
- **Frame shift errors**
- **insertion**
  - ATGCATCATCCCATT → ATGC\textcolor{red}{G}ATCATCCCATT
  - MetHisHisProIle → Met\textcolor{red}{Arg}SerPheHis
Protein sequence formats

• may differ from one database to another

• most software tools accept FASTA, GCG or GenBank formats

• some have hidden checksums
FASTA format

- begins with a description line indicated by a “>” sign
- followed by amino acid seq. in capital letters,
- no numbers, no blocks
- line length usually 80 characters

Example:

>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADIYDGFKTNCSNVVHVCTNLMNTTGGTLLNGSYSENRT
QIWQKHRTSNDALSILLNNKHYNLTVTCKRPGNKTVLPRVTIMAGLVFSQKYNLRQRAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWGDPETANLWFNCHEFFYCK
MDWFVLNYLNNLTVDADHNECKNTSGTKSNKRAPGPCVQRTYVACHIRSVIIIWLETISKK
TYAPPRGHLECTSTVTGMTEVNLNYIPKRNVTLSPIESIWAANLDRYKLVFETPIGF
APTEVRRYTGGHERQKRVPVXXXXXXX XXXXXXXXXX XXXXXXXXXX XXXXXXXXX
Q5SQRLLAGILQQQKNL
LAAVEAQQQMLKLTIWGVK
PIR format

• begins with keyword SEQUENCE
• next line are numbers indicating blocks
• lines containing sequence data start with running numbers
• line length usually 30 characters
• space between each letter

Example:

<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M</td>
<td>V</td>
<td>L</td>
<td>S</td>
<td>P</td>
<td>A</td>
</tr>
<tr>
<td>10</td>
<td>D</td>
<td>K</td>
<td>T</td>
<td>N</td>
<td>V</td>
<td>K</td>
</tr>
<tr>
<td>15</td>
<td>A</td>
<td>A</td>
<td>W</td>
<td>G</td>
<td>K</td>
<td>V</td>
</tr>
<tr>
<td>20</td>
<td>G</td>
<td>A</td>
<td>H</td>
<td>A</td>
<td>G</td>
<td>E</td>
</tr>
<tr>
<td>25</td>
<td>Y</td>
<td>G</td>
<td>A</td>
<td>E</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>E</td>
<td>R</td>
<td>M</td>
<td>F</td>
<td>L</td>
<td>S</td>
</tr>
<tr>
<td>61</td>
<td>K</td>
<td>K</td>
<td>V</td>
<td>A</td>
<td>D</td>
<td>A</td>
</tr>
</tbody>
</table>

Example continues...
GenBank format

- begins with qualifier `/translation=`
- amino acid sequence starts and ends with “”
- amino acid seq. in capital letters on same line as qualifier
- no numbers, no blocks

Example:
```
/translation="MNRWVEK WLRYLK CYINLILFYRNVYPQSFDTDYTTYSFNPQFVPINRHPALIDYLDELVLSDKLTHYRFSICIINKKNDLIEKYLDFSELQHVDKDDQIITEVDEFIRSLSLIMHLEKLPKVNDTITFEAVAINEIELELGHKLDRNR RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFPKIKLTSVGSDVGPLIIHQFSEKLISGDDKILNGYVEEESIFGSLF"
```
Geneppt Record

LOCUS       Z31371_1 [A7120FTSZ]
DEFINITION  Anabaena 7120 ftsZ and gsh-III genes. DATE 30-NOV-1995
ACCESSION  Z31371 NID ORGANISM Anabaena PCC7120 Bacteria;
            Cyanobacteria; Nostocales; Nostocaceae; Anabaena.
COMMENT CDS 385. .1671
          /product="FtsZ"
          /protein_id="CAA83241.1"
          /db_xref="GI:1100794"
          /db_xref="SWISS-PROT:P45482"
WEIGHT  44731
LENGTH  428
ORIGIN Translated using phase 1
Z31371_1 Length: 428 August 29, 2000 13:01 Type: P Check: 9372 ..

    1 MTLDNQELT YRNSQSLGPQ GSFLAVNSSN PFNHSGLNFG QNNDSSKISV
    51 ENNRIGEIVP GVRAVINKI VG5GGGNAVN RMIESDVSVG EFSINTDAQ
   101 ALTLGAPSR LQIGQKLVTR LGAGGNAIG QKAIAESREDE IATALEGADL
   151 VFITAGMGV TGTGAAAPIA EGAAKELT VGVVTRPVVF EGGRRTSQAE
   201 QQIEGLKSRV DTLIIIPNNK LLEVIPEQTP VQEAFRYADD VLRQGVQGIS
   251 DIITIPGLVI VDFADVRVVM ADAGSALMGV GVSSGKSRAR EAAIAAISP
   301 LLECSIEGAR GVVFNITGGS DLTLHEVNAI AETIYEVVDP NANIIFGAVI
   351 DDRLLQGEVR TVIATGFTGE IQAAPPQQNAI NARVVSAPPK RTPTQTPLTN
   401 SPAPTPPEKE KGSLDIPDFL QRRPPLKN

//
Conversion of sequence formats

**READSEQ software**

- developed by D.G. Gilbert (Indiana University)
- converts 18 different file formats, including:
  - Fasta, GenBank, PIR, NBRF,GCG, Staden, EMBL

**Source:**
- BAYLOR COLLEGE OF MEDICINE
  
  [http://dot.imgen bcm.tmc.edu:9331/seq-util/readseq.html](http://dot.imgen bcm.tmc.edu:9331/seq-util/readseq.html)
Three important messages

1. Be cautious when you have a significant hit.

2. The absence of a match in the database does not mean that there is no homologous sequence in the database.

3. The presence of a similar sequence is no guarantee that the sequences are homologous.
Take a close look at your sequence

Sir - We have discovered a startling similarity between a dinosaur DNA sequence reported in the novel Jurassic Park\(^1\) and a partial human brain DNA sequence from the \textit{Yenter laboratory described in Nature}\(^2\) (see figure).

The dinosaur sequence (DINO1) consists of duplication, with 117 base pairs from the first member of the repeat aligning with the human sequence, \textit{HUMX01451}, at the 95 per cent level of identity with only two gaps. The extraordinary degree of nucleotide sequence conservation between organisms as distantly related as dinosaur and human suggests strongly conserved function. Expression of \textit{HUMX01451} in human brain raises the possibility that the dinosaurs were smarter than has been supposed, arguing against the hypothesis that their extinction resulted from lack of intelligence.

Our discovery also seems to raise the interesting legal question as to whether the copyright on \textit{Jurassic Park} takes precedence over the pending patent on the human sequence. However, it appears that neither group is entitled to legal protection for its sequence, because both sequences also align with cloning vector pBR322, raising the possibility that both groups inadvertently sequenced vector DNA.

\textbf{Alan C. Christensen, Dept of Biochemistry and Molecular Biology, Thomas Jefferson University, Philadelphia, Pennsylvania, 19107 USA.}

\textbf{Steven Henikoff, Howard Hughes Medical Institute and Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle Washington 98104 USA.}

1 Crichton, M. Jurassic Park, 102 (Ballantine, New York 1990).

\begin{verbatim}
HUMX  317  GCCTGCTGCGGCTTTTCCATAAGGCTCGGCCGCCATAAGCATGACAAATCTGACCTCA
                    **************************************************
DINO1  1  GCCTGCTGCGGCTTTTCCATAAGGCTCGGCCGCCATAAGCATGACAAATCTGACCTCA--
                    **************************************************
DINO1  676  GCCTGCTGCGGCTTTTCCATAAGGCTCGGCCGCCATAAGCATGACAAATCTGACCTCA--
\end{verbatim}
What to do when your search produced no hits?

Standard solutions:

1. Check your sequence format

2. Raise the Expect (E) -value

3. Turn-off filtering of low complexity
Expect value (E) for a Score (S)

- In a database similarity search, the probability that an alignment score as good as the one found between a query sequence and a database sequence would be found by random chance.

  **Example:**
  
<table>
<thead>
<tr>
<th>Score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>108</td>
<td>$10^{-2}$</td>
</tr>
</tbody>
</table>
  
  =$\geq 1$ in 100 will have the same score

- The E-value decreases exponentially with the score
- The E-value changes with the size of the database.
BLASTp results from a SwissProt database search

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>sp P07807</td>
<td>DRY_YEAST Dihydrofolate reductase (EC 1.5.1.3) [DFR1]</td>
<td>434</td>
<td>e-121</td>
</tr>
<tr>
<td>sp P22906</td>
<td>DRY_CANAL Dihydrofolate reductase (EC 1.5.1.3) [DFR1]</td>
<td>146</td>
<td>1e-34</td>
</tr>
<tr>
<td>tr Q39687</td>
<td>DIHYDROFOLATE REDUCTASE-THYMIDYLATE SYNTHETASE PRECURS</td>
<td>111</td>
<td>5e-24</td>
</tr>
<tr>
<td>sp P45350</td>
<td>DRTS_DAUCA Bifunctional dihydrofolate reductase-thymid...</td>
<td>111</td>
<td>5e-24</td>
</tr>
<tr>
<td>sp O81395</td>
<td>DRTS_MAIZE Bifunctional dihydrofolate reductase-thymid...</td>
<td>108</td>
<td>2e-23</td>
</tr>
<tr>
<td>sp P51820</td>
<td>DRTS_SOYBN Bifunctional dihydrofolate reductase-thymid...</td>
<td>108</td>
<td>3e-23</td>
</tr>
<tr>
<td>sp Q05763</td>
<td>DRT2_ARATH Bifunctional dihydrofolate reductase-thymid...</td>
<td>106</td>
<td>1e-22</td>
</tr>
<tr>
<td>sp Q05762</td>
<td>DRT1_ARATH Bifunctional dihydrofolate reductase-thymid...</td>
<td>105</td>
<td>3e-22</td>
</tr>
</tbody>
</table>

View these as pairs

ProteinID Accession Locus name (ProteinName_Organism)
What is a significant E-value?

How many false positives to expect?
E-value: $10^{-4} = 1$ in 10,000 with score $y$

<table>
<thead>
<tr>
<th>Database</th>
<th>No. of Entries</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>SwissProt</td>
<td>105,967</td>
<td></td>
</tr>
<tr>
<td>PIR-PSD</td>
<td>283,153</td>
<td></td>
</tr>
<tr>
<td>TrEMBL</td>
<td>594,148</td>
<td></td>
</tr>
</tbody>
</table>
What is a significant E-value?

How many false positives to expect?
E-value: $10^{-4} = 1$ in $10,000$

<table>
<thead>
<tr>
<th>Database</th>
<th>No. of Entries</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>SwissProt</td>
<td>105,967</td>
<td>10.6</td>
</tr>
<tr>
<td>PIR-PSD</td>
<td>283,153</td>
<td></td>
</tr>
<tr>
<td>TrEMBL</td>
<td>594,148</td>
<td></td>
</tr>
</tbody>
</table>
What is a significant E-value?

How many false positives to expect?
E-value: $10^{-4}$ = 1 in 10,000

<table>
<thead>
<tr>
<th>Database</th>
<th>No. of Entries</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>SwissProt</td>
<td>105,967</td>
<td>10.6</td>
</tr>
<tr>
<td>PIR-PSD</td>
<td>283,153</td>
<td>28.3</td>
</tr>
<tr>
<td>TrEMBL</td>
<td>594,148</td>
<td>59.4</td>
</tr>
</tbody>
</table>
What is low complexity sequence?

- sequences that don’t require a specific residue but a specific type of amino acid residue in specific positions
  (e.g. coiled-coil and transmembrane regions).

**Coiled coils pattern:** a and d are hydrophobic

```
abcdefgabcdefg
```

Protein 1  `LSEIMEKLRKAMRKL`
Protein 2  `LAGAILHLMEMETEML`
NCBI notes on BLAST

• **Warning:** For short sequences 20 – 40, 50% of alignment happens by chance

• **Associative homologies:** If seq. A and seq. B are homologous and seq. B and seq. C are homologous then seq. A and seq. C are homologous even if you can’t see it.
  
• if $A \rightarrow B$ and $B \rightarrow C$, then $A \rightarrow C$
What to do when your BLAST search produced no hits?

Solutions we learned so far:

1. Switch databases
2. Use Dynamic Programming algorithms  
   (more rigorous, more sensitive)
3. Change the substitution matrix  
   (e.g. BLOSUM40 instead of BLOSUM62)
4. Decrease the Gap penalty
Change the Gap-Penalty (G=g+l n)

Match Score = 100, Cutoff = 89

1. THESESENTENSESALIGN--NICELY
   ||||| | | | ||||| ||||||
2. THESEQUENCE----ALIGNEDNICELY

1. THESESENTENSESALIGN--NICELY
   ||||| || |  ||||| |||||
2. THESE-Q--ENCE--ALIGNEDNICELY

1. THESESENTENSESALIGN--NICELY
   |||||  || |  ||||| |||||
2. THE--SEQ--ENCE--ALIGNEDNICLEY

1. THESESENTENSESALIGN--NICELY
   |||  ||  || |  ||||| |||||
2. THE--SEQ--ENCE--ALIGNEDNICLEY

g = 3  g = 1
Change the Gap-Penalty \((G=g*l\) n\)

Match Score = 100, Cutoff = 89

1. THESESENTENSESALIGN---NICELY
   \[\boxed{100} \]

2. THESEQENCE----ALIGNEDNICELY
   \[\boxed{-12} \]

1. THESESENTENSESALIGN---NICELY
   \[\boxed{88} \]

2. THESE-Q--ENCE--ALIGNEDNICLEY
Change the Gap-Penalty (G=g*l n)

Match Score = 100, Cutoff = 89

1. THESESENTENSESALIGN--NICELY 100
   11111 11 1 11111 11111 11111 11111 11111
   \underline{-12}

2. THESEQENCE----ALIGNEDNICELY 88

1. THESESENTENSESALIGN--NICELY 100
   11111 11 1 11111 11111 11111 11111 11111
   \underline{-18}

2. THESE-Q--ENCE-ALIGNEDNICLEY 82

1. THESESENTENSESALIGN--NICELY 100
   111 11 11 11111 11111 11111 11111
   \underline{-18}

2. THE--SEQ-ENCE-ALIGNEDNICLEY 82
Change the Gap-Penalty (G=g*1 n)

Match Score = 100, Cutoff = 89

1. THESESENTENSESALIGN--NICELY
   ||||| || | |||||  |||||
   100 100
2. THESEQENCE----ALIGNEDNICLEY
   |||||    || | |||||  |||||
   88 92

1. THESESENTENSESALIGN--NICELY
   ||||| || | |||||  |||||
   100
2. THESE-Q--ENCE--ALIGNEDNICLEY
   |||||    || | |||||  |||||
   82

1. THESESENTENSESALIGN--NICELY
   ||||| || | |||||  |||||
   100
2. THE--SEQ--ENCE--ALIGNEDNICLEY
   |||||    || | |||||  |||||
   82
Change the Gap-Penalty ($G = g \cdot l \cdot n$)

Match Score = 100, Cutoff = 89

<table>
<thead>
<tr>
<th></th>
<th>THESESENTENSESALIGN---NICELY</th>
<th>THESEQUENCE----ALIGNEDNICELY</th>
</tr>
</thead>
<tbody>
<tr>
<td>g = 3</td>
<td>100 100</td>
<td>88 92</td>
</tr>
<tr>
<td>g = 1</td>
<td>-10 -8</td>
<td>-18 -10</td>
</tr>
<tr>
<td>1</td>
<td>THESESENTENSESALIGN---NICELY</td>
<td>THESEQUENCE----ALIGNEDNICELY</td>
</tr>
<tr>
<td></td>
<td>100 100</td>
<td>82 90</td>
</tr>
<tr>
<td>2</td>
<td>THESESENTENSESALIGN---NICELY</td>
<td>THESEQUENCE----ALIGNEDNICELY</td>
</tr>
<tr>
<td></td>
<td>100 100</td>
<td>82 90</td>
</tr>
</tbody>
</table>
Automatic-BLAST M

- developed by Brian Osborne

- automatically sends sequences by email to the BLAST server at ncbi.nlm.nih.gov at user-specified times, daily or weekly

Source:
Automatic-BLAST M

Requirements:
1. Eudora version 1.4 or higher
2. Must have Applescript installed
3. You must have these files:
   • "System Folder:Extensions:Scripting Additions:Read/Write Commands"
   • "System Folder:Extensions:Scripting Additions:Current Date"
Automatic-BLAST M

Disadvantages:
• accepts limited number of sequence formats
• can’t place multiple sequences in 1 file
• doesn’t handle sequences > 4000 letters
  (increase memory in “Get infobox”)

More details at:

http://bioinformatics.weizmann.ac.il/software/mac/automatic-blast.readme
When to use what BLAST?

- BLASTN = nucleic acids against nucleic acids
- BLASTP = protein query against protein database
- BLASTX = translated nucleic acids against protein database
- TBLAST = protein query against translated nucleic acid database
- TBLASTX = translated NA against translated NA
What to do when your search returns homologs with weak similarities?

- Run PSI-BLAST
- Look for sequence motifs
PSI-BLAST

Description:

• PSI-BLAST: Position-Specific Iterated BLAST

• used to identify more divergent homologs

• may help identify distant members of protein families
How PSI BLAST Works

• Step 1. Standard Blast with your protein

• Step 2. Generate a consensus sequence

• Step 3. Use consensus sequence for a second search
<table>
<thead>
<tr>
<th>Sequences</th>
<th>Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family</td>
<td>Family</td>
</tr>
<tr>
<td>Domain</td>
<td>Domain</td>
</tr>
<tr>
<td>Motif</td>
<td>Motif</td>
</tr>
<tr>
<td>Blocks</td>
<td>Fold /Foldon</td>
</tr>
<tr>
<td>Pattern</td>
<td>Class</td>
</tr>
<tr>
<td></td>
<td>Active site</td>
</tr>
<tr>
<td></td>
<td>Architecture</td>
</tr>
</tbody>
</table>
Protein Family Definition

1) Structural context
- two structures that have significant levels of structural similarity but not necessarily sequence similarity
Protein Family Definition

2) Sequence context (Dayhoff 1978)
   group of proteins of similar function that are more than 50% identical when aligned
Pfam

Description:

• **Pfam** = Protein families
• database of multiple alignments of protein domains or conserved protein regions.
• **Pfam-A** are accurate human crafted multiple alignments
• **Pfam-B** is an automatic clustering of the rest of SWISS-PROT and TrEMBL proteins using the program **Domainer**
Pfam

Sources:

• SWISS-PROT and TrEMBL
  http://pfam.wustl.edu/

• EBI, UK
  http://www.sanger.ac.uk/Software/Pfam

• Sweden
  http://www.cgr.ki.se/Pfam/
Pfam Results

Accession number: PF00096

Zinc finger, C2H2 type

The C2H2 zinc finger is the classical zinc finger domain. The two conserved cysteines and histidines co-ordinate a zinc ion. The following pattern describes the zinc finger. #-X-C-X(1-5)-C-X3-#-X5-#-X2-H-X(3-6)-[H/C] Where X can be any amino acid, and numbers in brackets indicate the number of residues. The positions marked # are those that are important for the stable fold of the zinc finger. The final position can be either his or cys. The C2H2 zinc finger is composed of two short beta strands followed by an alpha helix. The amino terminal part of the helix binds the major groove in DNA binding zinc fingers.
Zinc finger domains [MEDLINE:00000816], [MEDLINE:00005329] are nucleic acid-binding protein structures first identified in the *Xenopus* transcription factor TFIIIA. These domains have since been found in numerous nucleic acid-binding proteins. A zinc finger domain is composed of 25 to 30 amino-acid residues including 2 conserved Cys and 2 conserved His residues in a C-2-C-12-H-3-H type motif. The 12 residues separating the second Cys and the first His are mainly polar and basic, implicating this region in particular in nucleic acid binding. The zinc finger motif is an unusually small, self-folding domain in which Zn is a crucial component of its tertiary structure. All bind 1 atom of Zn in a tetrahedral array to yield a finger-like projection, which interacts with nucleotides in the major groove of the nucleic acid. The Zn binds to the conserved Cys and His residues. Fingers have been found to bind to about 5 base pairs of nucleic acid containing short runs of guanine residues. They have the ability to bind to both RNA and DNA, a versatility not demonstrated by the helix-turn-helix motif. The zinc finger may thus represent the original nucleic acid binding protein. It has also been suggested that a Zn-centred domain could be used in a protein interaction, e.g. in protein kinase C. Many classes of zinc fingers are characterized according to the number and positions of the histidine and cysteine residues involved in the zinc atom coordination. In the first class to be characterized, called C2H2, the first pair of zinc coordinating residues are cysteines, while the second pair are histidines.
Pfam links to other databases
2. Sequence MOTIF databases
Motif definition

Sequence context: (Prosite catalog)
- conserved pattern of amino acids found in 2 or more proteins (e.g. 3’ –5’ exonuclease motifs)
- often near the active site
- found in a group of proteins with similar functions
Motif definition

Structural context:
- a group of several secondary structural elements produced by folding of adjacent sections of the protein into a specific 3D configuration (e.g. helix-loop-helix motif)
- also called supersecondary structures or folds
Prosite Database

• allows retrieval of documents by “author names, citation, accession number, or as full text searches”
• allows motif scans
• Example:

AMLKSSGRGRP\k
T–LRETPRGRPR
AM–KALGRGRPR

[AT]-x(1,2)-[RK](2)-[GP]-R-G-R-P-[RK]-x

Source:
http://www.expasy.ch/prosite/
Search in PROSITE for: DNA binding

(Release 17.5, of 07-Mar-2002 )
Please choose one of the following entries:

• **PDOC00031** Nuclear hormones receptors DNA-binding region signature
• **PDOC00037** Myb DNA-binding domain repeat signatures and profile
• **PDOC00381** HSF-type DNA-binding domain signature
• **PDOC00044** Bacterial histone-like DNA-binding proteins signature
• **PDOC00306** HMG-I and HMG-Y DNA-binding domain (A+T-hook)
# Prosite results

<table>
<thead>
<tr>
<th>PROSITE cross-reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS00354: HMGI_Y</td>
</tr>
</tbody>
</table>

## Documentation

High mobility group (HMG) proteins are a family of relatively low molecular weight non-histone components in chromatin. HMG-I and HMG-Y are proteins of about 100 amino acid residues which are produced by the alternative splicing of a single gene (HMG-Y differs from HMG-I by the internal deletion of 11 amino acids). HMG-I/Y bind preferentially to the minor groove of A+T-rich regions in double-stranded DNA. It is suggested that these proteins could function in nucleosome phasing and in the 3’ end processing of mRNA transcripts. They are also involved in the transcription regulation of genes containing, or in close proximity to, A+T-rich regions.

DNA-binding of these proteins is effected by a 11 residues domain, called the A+T-hook [1]. This domain is repeated three times in the sequence of HMG-I/Y.
**Prosite results**

<table>
<thead>
<tr>
<th>Description of pattern(s) and/or profile(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Consensus pattern</strong></td>
</tr>
<tr>
<td><strong>Sequences known to belong to this class detected by the pattern</strong></td>
</tr>
<tr>
<td><strong>Other sequence(s) detected in SWISS-PROT</strong></td>
</tr>
</tbody>
</table>
3. Structural databases
Structural completeness

Definition:
• at least one coordinate value for each and every atom in the chemical graph is present.

The Myth:
• all data in the public structure database are of textbook quality
• only terminal parts of a structure are missing
Structural completeness

Fact:
- structural completeness is quite rare in structure database records
- most X-ray structures lack coordinates for hydrogen atoms because the locations of hydrogens in space are not resolved by the experimental methods currently available.

Solution:
Some modeling software can be used to predict the locations of the hydrogen atoms and reconstruct a structure record with the “modeled” hydrogens added
SCOP

Description:
- **SCOP** = Structural Classification Of Proteins
- classification by **class**
- additionally classification by **family, superfamily** and **fold**
- reflects both structural and evolutionary relationships

Sources:
http://scop.mrc-lmb.cam.ac.uk/scop/
CATH

Description:
• CATH = Classification by class, Architecture, Topology, and Homology
• Similar to SCOP
• \(\alpha/\beta\) and \(\alpha + \beta\) proteins are in one class

Sources:
http://www.biochem.ucl.ac.uk/bsm/cath/