Databases

Background

Genome Emerging Projects

↓

Massive Sequence Information

↓

Biological Information Inference

Sharp biological data growth

Not published in the traditional way

Compiled in databases and numbered

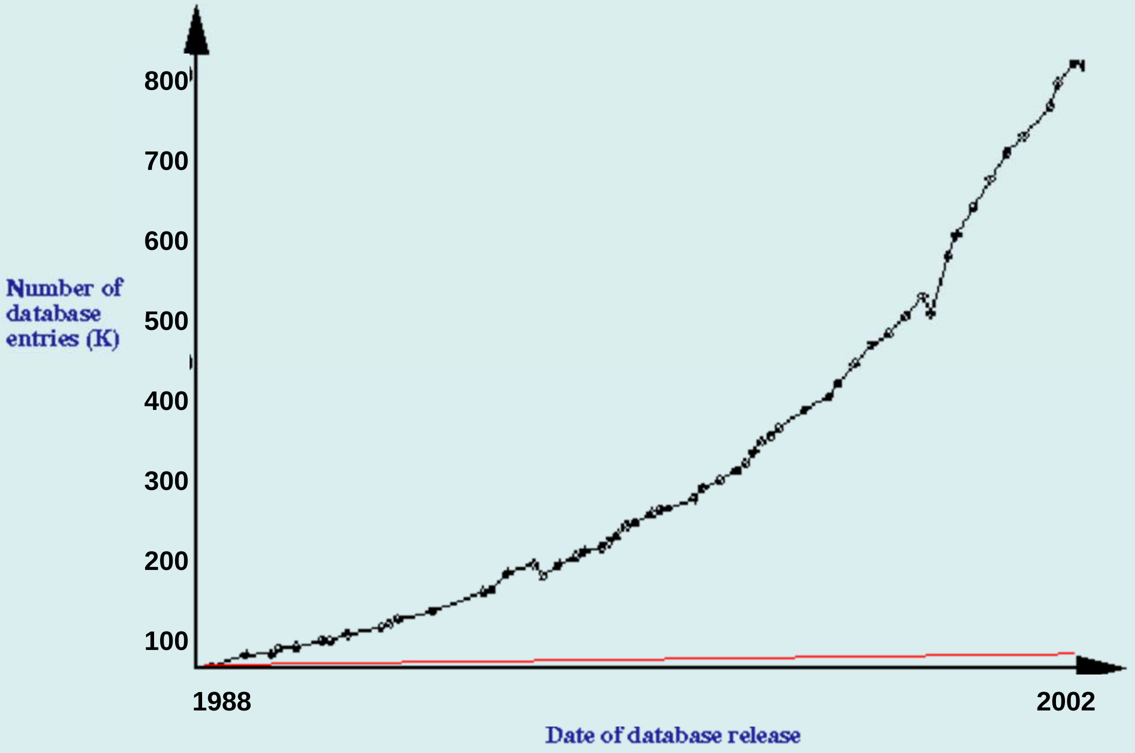
Identification allows localization, access and reference

Importance of sequence analysis

- >90,000,000 sequences available in public dbs
  - (including ESTs) in proprietary dbs
  - these #s will snowball with completion of more genomes
  - so what?
- Locked up in sequences is a huge amount of structural, functional & evolutionary info
  - they're a highly valuable resource
- By contrast, the # of unique protein structures is ~2000
  - a huge information deficit

The legacy of the genome projects

Sequence-structure deficit



Non-redundant growth of sequences during 1988-2002 (▬) & the corresponding growth in the number of structures (▬).

Challenges for bioinformatics

- Spurred on by the seq/structure deficit, the challenges
  - rationalise the mass of sequence data
  - derive more efficient means of data storage
  - design more incisive & reliable analysis tools
- The imperative - to convert sequence information into biochemical & biophysical knowledge
  - to decipher the structural, functional & evolutionary clues encoded in the language of biological sequences

The practical - BioActivity

- BioActivity – sequence analysis in action
  - begin with a fragment of a DNA sequence
  - try to find out what protein this codes for, the family to which it belongs, & whether its function & structure are known
- The practical is entirely Web-based
  - be mindful of traffic – don't waste time on slow links
- Most important of all
  - **read the instructions!**
- The Web is constantly evolving....
  - Beware of dead links!

Recommended Utilities

There is a great variety of utilities to face up to sequence analysis.

Most of them can be browsed from WEB sites and in most occasions they are grouped in important web pages as EBI, NCBI, EXPASY, etc.

It is highly recommendable not to concentrate on one particular and specific method, since it is necessary to explore different databases with different approaches and establish **"a general perspective"**

WHY?

- No database is already complete
- Contents of similar resources only overlap partially
- No data search or alignment is infallible
- Some databases and pattern recognition techniques are still being tested

Warnings

- **Do not believe all you can find in databases**: The information contained could be deceitful or wrong: estimates confer a 01-4% error in genome sequences, and a 5% error in proteins

- **Do not believe all what programmes show**: Results could be deceitful or wrong due to programmers' mistakes

- **Do not believe all you can find in Web servers**: The information contained could be deceitful or wrong due to Web programmers, including paramount bioinformatics centers

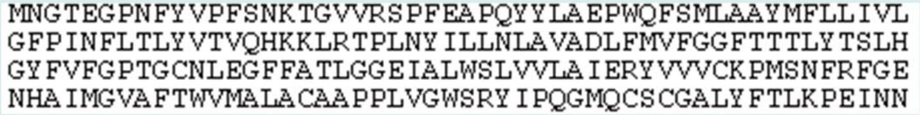
- **Do not believe all you read**: Errors are abundant in published literature, also in frequently quoted traditional papers

Recommendations:

Think. Question, Criticise
Get a global meaningful sense
Escape from small triumphs
Search for patterns away from "noise"

Do not be a gullible user

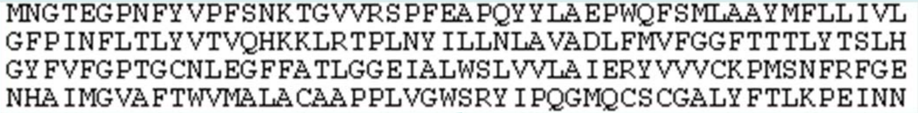
The Holy Grail of bioinformatics



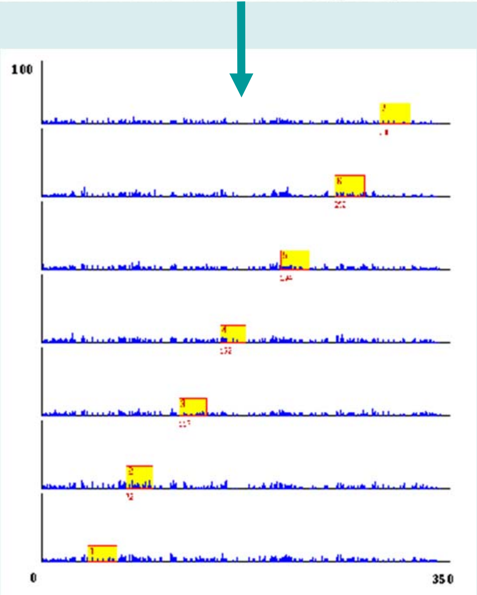
```
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIVL
GFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLH
GYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGE
NHAIMGVAFTWVMALACAAPPLVGWSRYIPQGMQCSCGALYFTLKPEINN
```

...to be able to understand the words in a sequence sentence that form a particular protein structure

The reality of sequence analysis



```
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIVL
GFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLH
GYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGE
NHAIMGVAFTWVMALACAAPPLVGWSRYIPQGMQCSCGALYFTLKPEINN
```



This means we can recognise words that form characteristic patterns, even if we don't know the precise syntax to build complete protein sentences

DNA Analysis: Why?

Comparisons among protein sequences are:

- More sensitive
- More simple (genetic code redundancy: 64 codons are reduced to 20 amino acids)

However, the loss of degeneration is followed by a loss of information related to the evolutive process.

Example: with sequence protein we can not follow silent mutations.

Features of the Sequence DNA Analysis

1) ORFs detection
2) Introns and exons detection
3) DNA sequence assembly
4) Expressed Sequence Tags Interpretation (EST)
   a) Insertions, deletions and ORFs changes
   b) Splicing Variants
   c) EST of non-coding regions
5) Gene hunting
   a) Positional cloning
   b) RNA transcritps analysis
   c) Hierarchy of genomic information
      - Cromosomic Genome (genome)
      - Expressed Genome (transcriptome)
      - Interacting proteins (proteome)

Pattern recognition & prediction

- In investigating the meaning of sequences, two distinct analytical approaches have emerged
  - **pattern recognition** is used to detect similarity between sequences & hence to infer related structures & functions
  - ***ab initio* prediction** is used to deduce structure, & to infer function, directly from sequence
- These methods are quite different!
  - pattern recognition methods demand that some characteristic has been seen before & housed in a db
  - prediction methods remove the need for template dbs, because deductions are made directly from sequence

Pattern recognition

It is the **main analytical focus** in Bioinformatics

It is based on the supposition of any **underlying characteristic** in the sequence or structure of a protein which serves to identify similar characters in related proteins.

Premise: If it is conserved, this is because it is important for the protein activity or for its correct folding.

Not only in **SEQUENCES** but also in **STRUCTURES**:

- Previous observation of any particular characteristic
- Storage in a reference database
- Identification of these features in new sequences/structures

Good results (40%), better in sequences than in structures, and only in professional hands

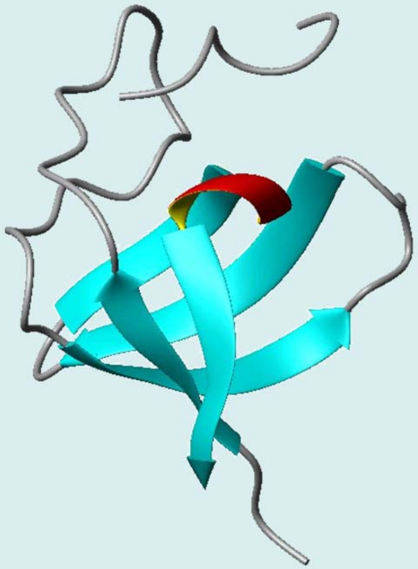
Pattern prediction

Direct deduction of a functional site characteristic or a whole structure without having been observed before.

Databases are not necessary, but design and calibration very sophisticated programmes are required

Folding recognition: Central topic in Molecular Biology

```
MVVDQLKNNIVSISSNIGTHIKDG
YHDIKQSVSGLKYITSQNHHLSKS
FDSWQATQEIPHVHPKEMQFLNES
MKKYLKKTLKLIKKRNMKRTEQDH
FNSLNKKCIQLLPNIVSFLDEFVE
KSYTEIIDDWESNSTATRLQIESL
VEKKHVVNTKDKVNGIFNDTLEID
KKDGPELPPRSNTAIVRHSVLVSP
SSISSMSSVSYSNHDSSEAKLRRI
LIQSFFEKLNINTDNKVLKTAKYD
WLIGVSRIDDESYRIGFVPNNYVE
```



Science fact & fiction

- Sequence pattern recognition is easier to achieve, & is much more reliable, than fold recognition
  - which is ~50% reliable even in expert hands
- Prediction is still not possible
  - & is unlikely to be so for decades to come (if ever)
- Structural genomics will yield representative structures for many (but not all) proteins in future
  - structures of new sequences will be determined by modelling
  - prediction will become an academic exercise
- But, to debunk a popular myth, knowing structure alone does not inherently tell us function

Tertiary Structure Prediction

IMPOSSIBLE: Protein folding rules are not understood

Secondary Structure Prediction

Only 50-60% reliability

Approximations are used:

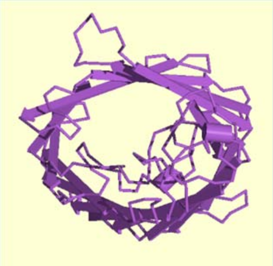
i) **Empirical statistical methods** derived from known 3D structures

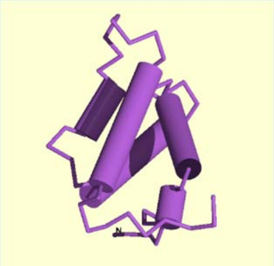
ii) Methods based on **physico-chemical criteria**: hydrophobicity, charge, etc.

iii) **Predictive algorithms**: they use known structures from homologous ones to assign secondary structures.

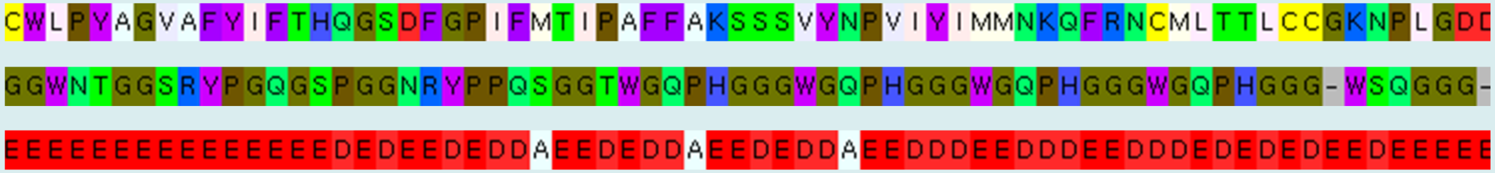
A reality check

- What is the function of this structure?



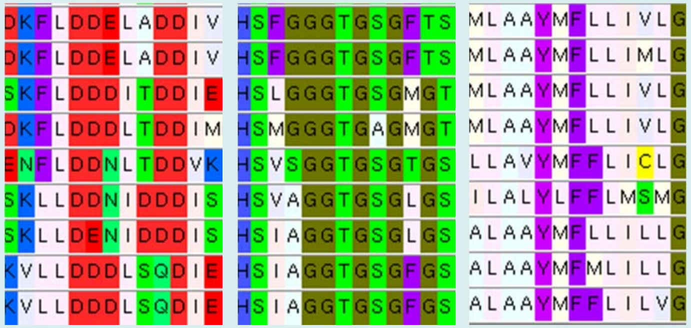


- What is the function of this sequence?



```
CWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKSSSVYNPVIYIMMNKQFRNCMLTTLCCGKNPLGDD

GGWNTGGSRYPGQGSPGGNRYPPQSGGTWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGG-WSQGGG-

EEEEEEEEEEEEEEEDEDEEDEDDAEEDEDDAEEDEDDAEEDDDEEDDDEEDDDEDEDEDEEDEEEEE
```

- What is the function of this motif?
  - the fold provides a scaffold, which can be decorated in different ways by different sequences to confer different functions – knowing the fold & function allows us to rationalise how the structure effects its function at the molecular level



```
DKFLDDELADDIV  HSFGGGTGSGFTS  MLAAYMFLLIVLG
DKFLDDELADDIV  HSFGGGTGSGFTS  MLAAYMFLLIMLG
SKFLDDDITDDIE  HSLGGGTGSGMGT  MLAAYMFLLIVLG
DKFLDDDLTDDIM  HSMGGGTGAGMGT  MLAAYMFLLIVLG
ENFLDDNLTDDVK  HSVSGGTGSGTGS  LLAVYMFFLICLG
SKLLDDNIDDDIS  HSVAGGTGSGLGS  ILALYLFFLMSMG
SKLLDENIDDDIS  HSIAGGTGSGLGS  ALAAYMFLLILLG
KVLLDDDLSQDIE  HSIAGGTGSGFGS  ALAAYMFMLILLG
KVLLDDDLSQDIE  HSIAGGTGSGFGS  ALAAYMFFLILVG
```

Sequence analysis

Information coded in primary structures **is not known**.

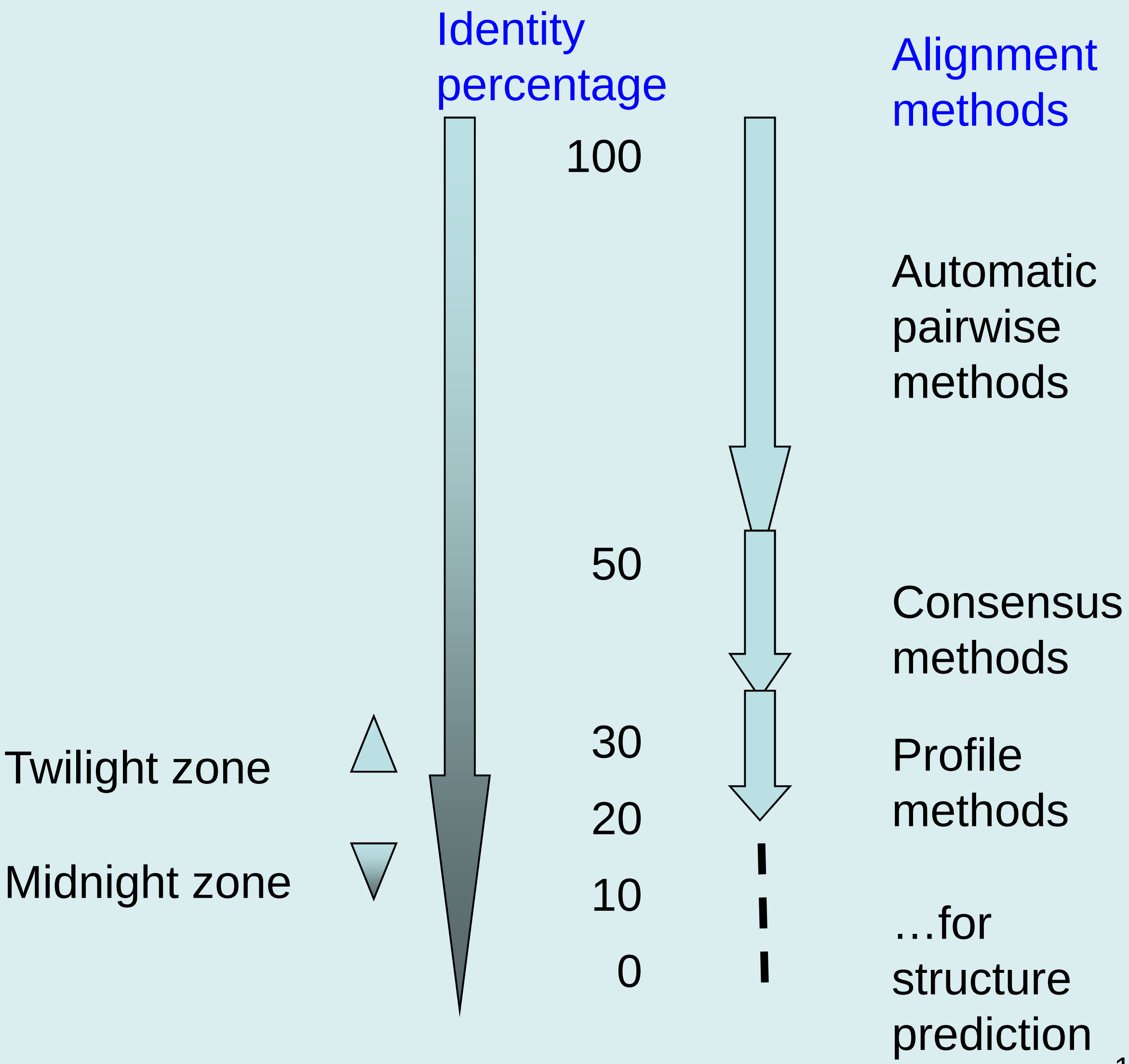
Folding rules **are not known**

Any scape?

Sequence analysis

Find similarities

Infer function



Identity percentage

Alignment methods

100

Automatic pairwise methods

50

Consensus methods

Twilight zone

30

Profile methods

20

Midnight zone

10

0

…for structure prediction

The Twilight Zone

- Prediction methods don’t work because we don’t fully understand the Folding Problem
  - we can’t read the language sequences use to create their folds
- But, with sequence analysis techniques, we can try to find similarities between new sequences & those in dbs
  - whose structures & functions we hope have been elucidated
- This is straightforward at high levels of identity, but below 50% it is difficult to establish relationships reliably
- Analyses can be pursued with decreasing certainty towards the Twilight Zone
  - ~20% identity, where results may look plausible to the eye, but are no longer statistically significant

Beyond the Twilight Zone

- To penetrate deeper into the Twilight Zone is the aim of most analytical methods
  - whether using single sequences, motifs, complex weighting schemes or raw amino acid frequencies
- Each offers a different perspective, depending on the type of information used in the search
  - none gives the **right** answer
- It is good practice to devise an analysis protocol that uses a variety of methods
  - but don’t expect the impossible – no method is infallible!

Homology & Analogy

- The term homology is confounded & abused!
  - sequences are homologous if they are related by divergence from a common ancestor
  - analogy relates to the acquisition of common features (similar foldings, or proteins with the same cathalytic residues with almost exact geommetry) from unrelated ancestors via convergent evolution (they do not have a proven sequence similitude)
    - *e.g.*, β-barrels occur in soluble & membrane proteins; enzymes chymotrypsin & subtilisin share groups of catalytic residues, with near identical spatial geometries, but no other similarities
- It is not a measure of similarity & is not quantifiable
  - it is an absolute statement that sequences have a divergent rather than a convergent relationship
  - the phrases "the level of homology is high" or "the sequences show 50% homology", or any like them, are strictly meaningless!

Orthology & Paralogy

Orthology

Among homologue sequences, proteins which carry out the same function in different species (divergence in one gene in different species)

Paralogy

They develop different functions, but related, in the same living organism (by duplication of genes which follow different evolutive routes)

The comparison of orthologue sequences opens a way for the study of molecular paleonthology:

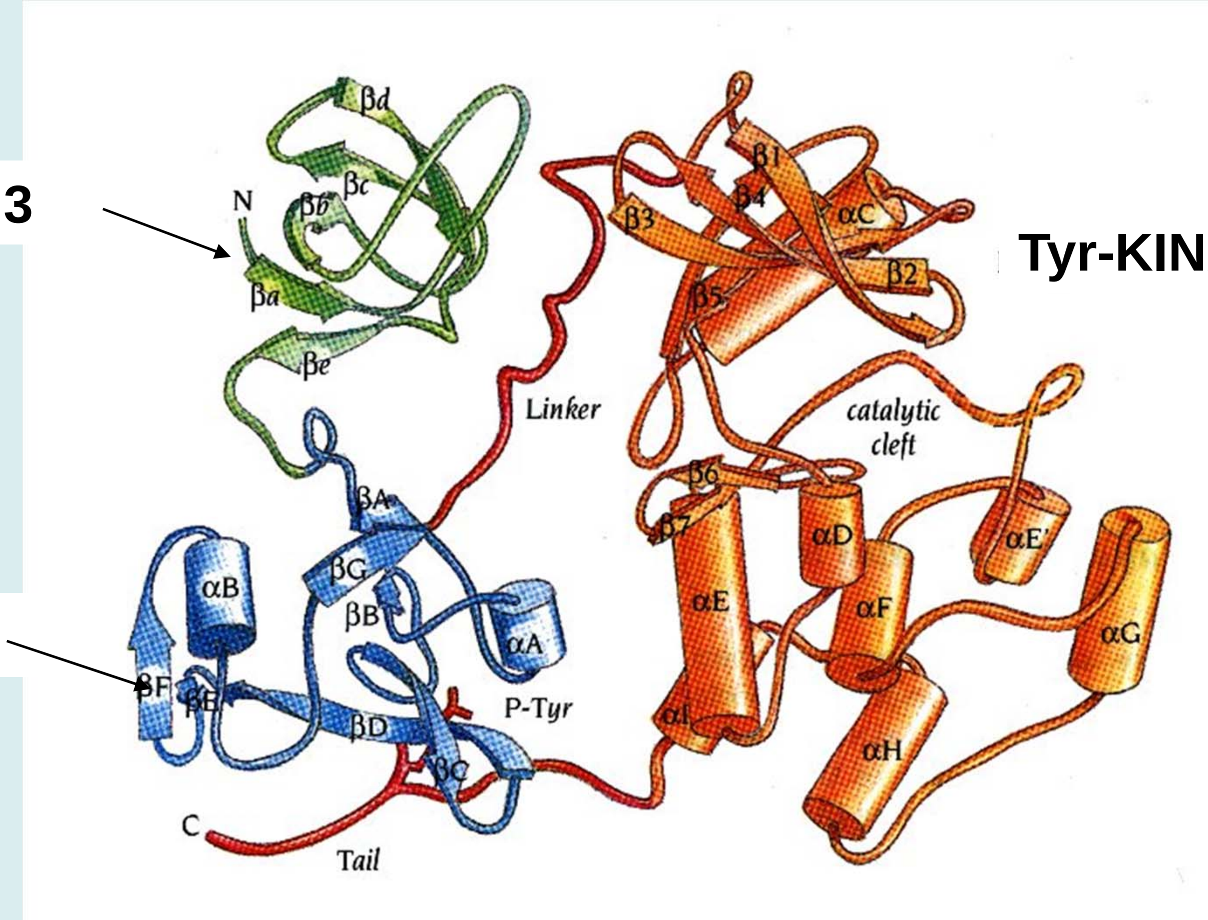
Phylogeny

The construction of phylogenetic trees has related bacterial, fungi and mammal proteins, and also among animals, insects and plants.

Devil is in the details

- Analytical complexity for functional assignment in an unknown sequence.
- Partial alignment is restricted to certain parts of the sequence.
- Much of the challenge is in getting the biology right
  - this is complicated by the problem of orthology *vs* paralogy
- Following a search, how much functional annotation can be legitimately inherited by a query?
  - source of numerous annotation errors in databases
  - error propagation could lead to an error catastrophe
- Further complications arise due to modular nature of proteins
  - modules are autonomous folding units (protein building blocks)
  - confer variety of functions on a parent protein, by multiple combinations of the same module, or different modules to form mosaics
- Automatic analysis systems don’t distinguish orthologues from paralogues & don’t consider the modular nature of proteins

Protein Domains

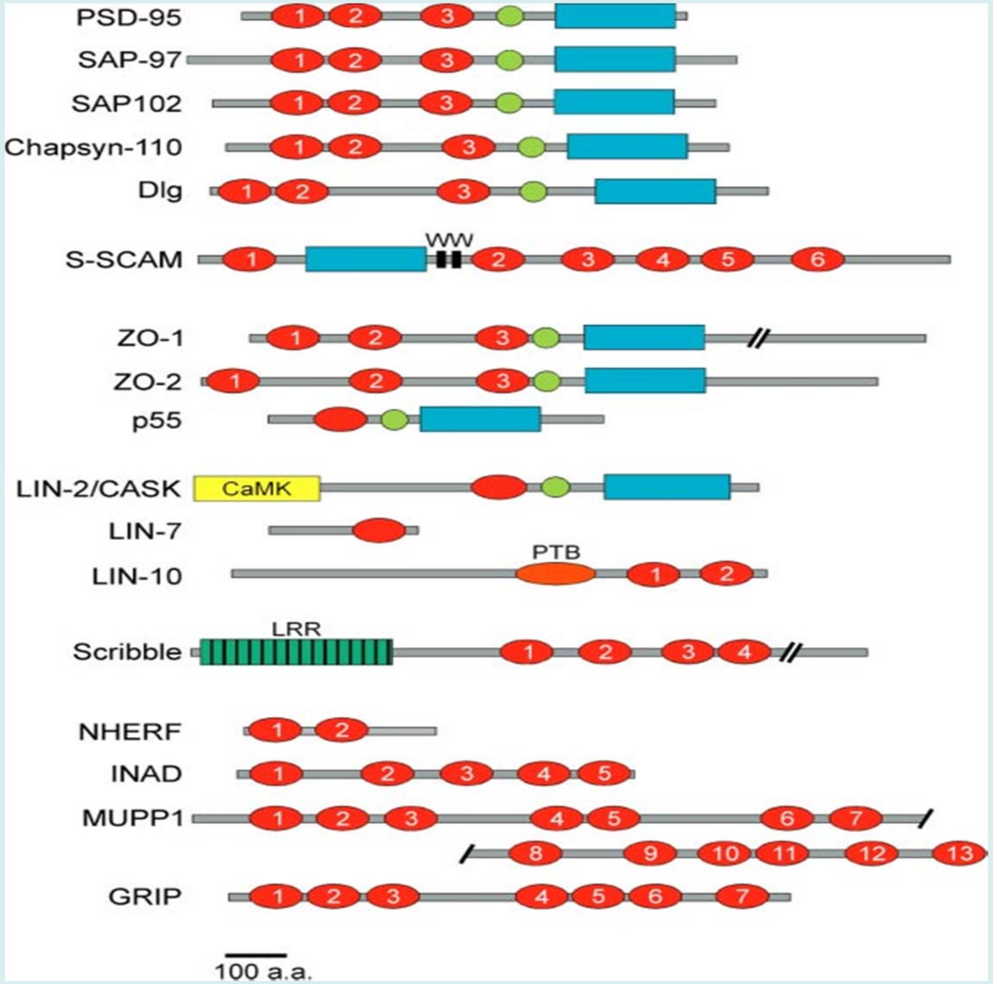


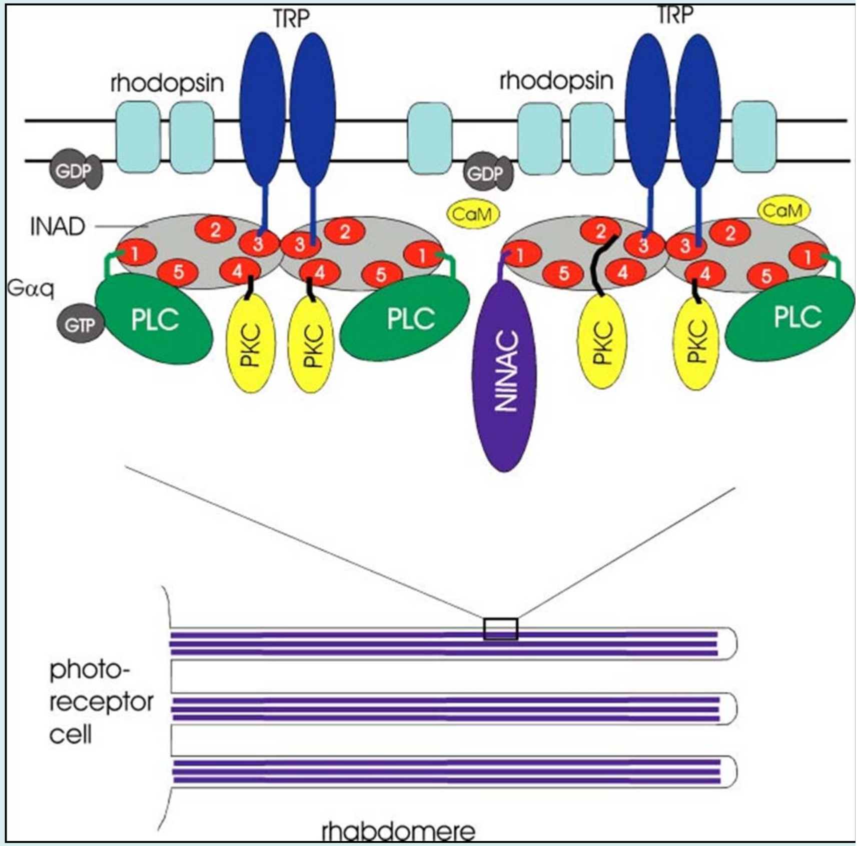
SH3

SH2

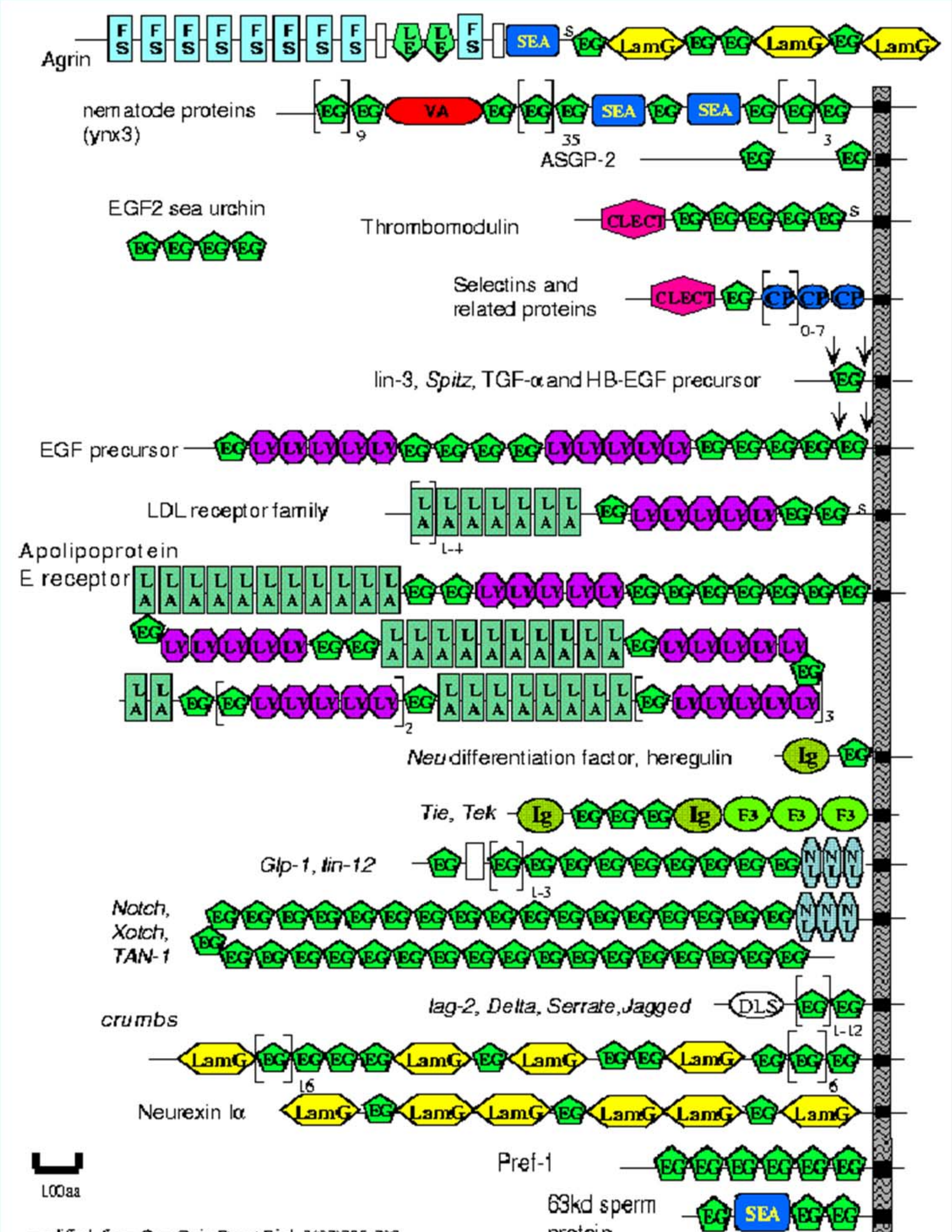
Tyr-KINASE

Mosaic Proteins





Mosaic Proteins



modified from Curr.Opin.Struct.Biol. 3(93)385-392

MOLECULAR BIOLOGY DATABASES

1. BIBLIOGRAPHIC
2. TAXONOMIC
3. NUCLEOTIDE
4. GENOMIC
5. PROTEIN
6. MICROARRAY

Molecular biology databases description for research purposes

Database are available from WWW sites and are highly interlinked

1. BIBLIOGRAPHIC DATABASES

MEDLINE is accessible through EBI's SRS.

PUBMED is accessible through NCBI's ENTREZ.

EMBASE is a commercial product for the medical literature.

BIOSIS, the inheritor of the old Biological Abstracts, covers a broad biological field; the Zoological Record indexes the zoological literature.

CAB International maintains abstract databases in the fields of agriculture and parasitic diseases.

AGRICOLA is for the agricultural field what MEDLINE is for the medical field . The bibliographical databases are with the exception of MEDLINE/PUBMED only available through commercial database vendors.

2. TAXONOMIC DATABASES

1. NEWT
2. The Tree of Life project
3. Species 2000
4. International Organization for Plant Information
5. Integrated Taxonomic Information System

Taxonomic databases can be considered to be rather controversial, due to the differing views within the taxonomic community

3. NUCLEOTIDE DATABASES

This collaboration is a joint operation by EMBL-Bank at the European Bioinformatics Institute (EBI), the DNA Data Bank of Japan (DDBJ) at the Center for Information Biology (CIB) and Genebank at the National Center for Biotechnology Information (NCBI).

DDBJ, GenBank and EMBL-Bank exchange new and updated data on a daily basis to achieve optimal synchronisation. The result is that they contain exactly the same information, except for sequences that have been added in the last 24 hours.

In Europe, the vast majority of the nucleotide sequence data produced is collected, organised and distributed by the EMBL Nucleotide Sequence Database located at the EBI in Cambridge UK, an Outstation of the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany.

Example: Genbank functional division

GB division	*Nucleotides*
Organism	2,300,497,789
EST	2,451,695,768
HTG	4,402,496,751
GSS	1,051,117,888
PAT	72,022,274
STS	51,227,345

•PAT -Patent

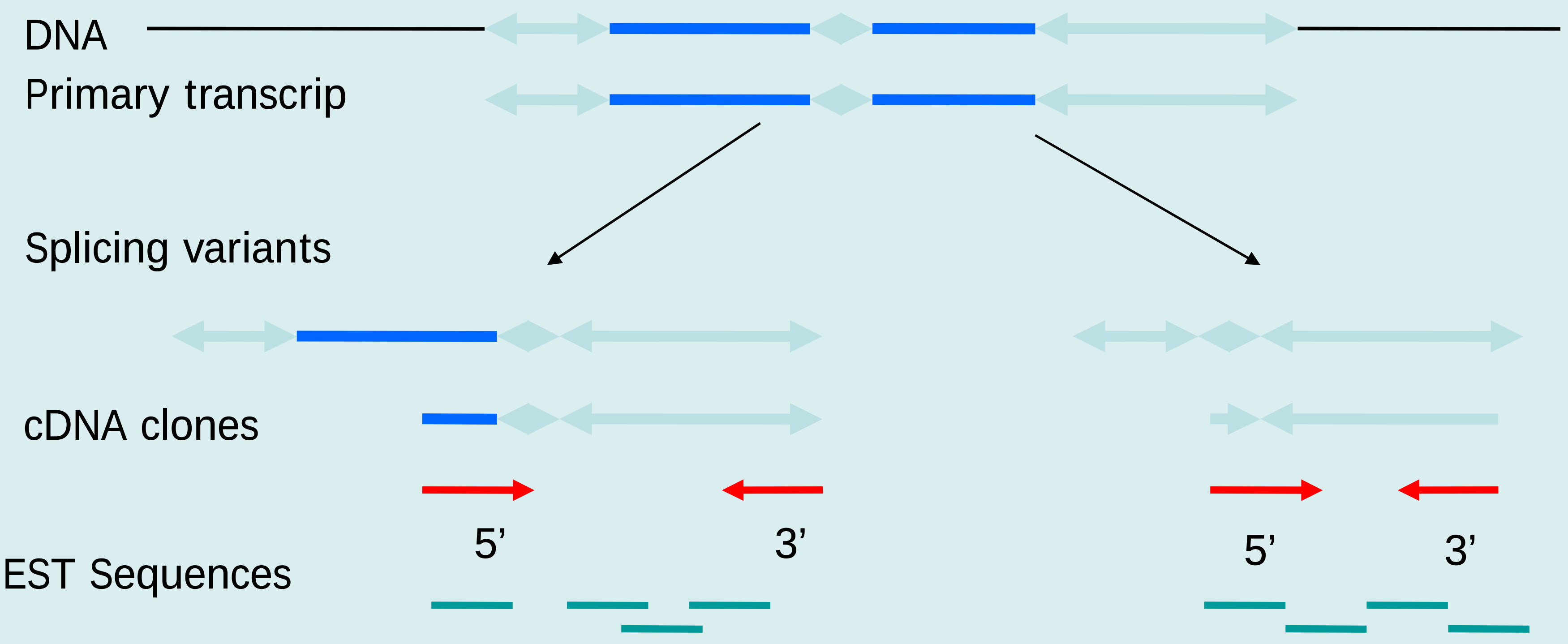
•EST –Expressed Sequence Tags are short (300-500 bp) single reads from mRNA (cDNA) which are produced in large numbers. They represent a snapshot of what is expressed in a given tissue, and developmental stage.

•STS –Sequence Tagged Sites are operationally unique sequence that identifies the combination of primer pairs used in a PCR assay that generate a mapping reagent which maps to a single position within the genome.

•GSS -Genome Survey Sequences are similar in nature to the ESTs, except that its sequences are genomic in origin, rather than cDNA (mRNA).

•HTG -High ThroughputGenome

EST Sequences



Tools available for EST analysis:

-Similarity search

-Sequence assembly

-Sequence grouping

OTHER NUCLEOTIDE SEQUENCE DATABASES

Genomes Server - this gives access to a large number of complete genomes.

UniGene - a sequence-cluster database which address the redundancy problem by coalescing sequences that are sufficiently similar that one may reasonably infer that they are derived from the same gene.

STACK - the 'Sequence Tag Alignment and Consensus Knowledgebase' another sequence-cluster database which address the same problem as UniGene.

EMBL-SVA - the 'EMBL Sequence Version Archive' server is a repository of all entries that have been made public since release 1 of the EMBL database. It comprises more than 100 million entries and includes entries pre-dating the first electronic release of the database in 1982.

SPECIALISED NUCLEOTIDE DATABASES

RDP - the 'Ribosomal Database Project' provides ribosome related data services to the scientific community, including online data analysis, rRNA derived phylogenetic trees, and aligned and annotated rRNA sequences.

HIV-SD - the 'HIV Sequence Database' collects, curates and annotates HIV and SIV sequence data and provides various tools for analysing this data.

IMGT - the 'ImMunoGeneTics database' is a database specialising in Immunoglobulins, T cell receptors and the Major Histocompatibility Complex (MHC) of all vertebrate species.

TRANSFAC - contains sequence information on transcription factors and transcription factor binding sites.

EPD - the 'Eukaryotic Promoter Database' is an annotated non-redundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally.

REBASE - for restriction enzymes and restriction enzyme sites.

GOBASE - is a specialised database of organelle genomes.

4. GENOMIC DATABASES

Genomes Server - this server gives access to a hundreds of **complete genome sequences**, including those from archaea, bacteria, eukaryota, organelles, phages, plasmids, viroids and viruses.
Proteome Analysis - the Proteome Analysis database has been set up to provide comprehensive statistical and comparative analyses of the predicted proteomes of fully sequenced organisms.
Ensembl - this is a joint project between the EBI and the Wellcome Trust Sanger Institute that aims at developing a system that maintains automatic annotation of large eukaryotic genomes.
Karyn's Genomes - contains general information about organisms whose genomes are completely sequenced.
WormBase – specific information for *C. elegans* (and some other nematodes).
FlyBase - database for *Drosophila melanogaster*.
MGD – “Mouse Genome Database”. Database for mice.
RGD – “'Rat Genome Database “. Database for rats.
SGD - the “Saccharomyces Genome Database” is another major yeast database.
HIV-SD - the 'HIV Sequence Database' collects, curates and annotates HIV and SIV sequence data and provides various tools for analysing this data.

OTHER GENOMIC DATABASES

E.coli databases – 'E. coli Genetic Stock Center' (CGSC), maintains a database of E.coli genetic information.

Plant databases –
MaizeDB is the database for genetic data on maize.
MENDEL is a plant-wide database for plant genes.

Fish databases –
ZFIN - a database for the zebrafish Brachydanio rerio.

Genetic databases of economic importance to humans –
PIGBASE
BovBASE
SheepBASE
ChickBASE.

OTHER GENOMIC DATABASES

Human databases-
OMIM - is a catalogue of human genes and genetic disorders.

GENATLAS also provides a database of human genes, with links to diseases and maps.
GeneCards - integrates information about human genes from a variety of databases, including OMIM, UniProt/Swiss-Prot.

Parasite databases-
Parasite genome - this database is supported by the World Health Organisation (WHO) at the EBI. It covers the five ‘targets' of its Tropical Diseases Research programme: Leishmania, Trypanosoma cruzi, African Trypanosomes, Schistosoma and Filariasis.

AnoDB -for some vectors of parasitic diseases are also available, such as for Anopheles.
AaeDB for Aedes aegypti.

5. PROTEIN DATABASES

The protein databases are the most comprehensive source of information on proteins. It is necessary to distinguish :

UNIVERSAL DATABASES: covering proteins from all species
1) Simple archives of sequence data : Only a sequence is included
2) Annotated databases: additional information has been added to the sequence record

A. Primary protein sequence databases (i.e. UniProt/Swiss-Prot)
B. Secondary protein databases (i.e.: InterPro)
C. Structure databases (PDB)

SPECIALIZED: data collections storing information about specific families or groups of proteins, or about the proteins of a specific organism.

Specialized protein sequence databases (i.e. GOA)
Specialized protein databases (i.e. ENZYME)

PRIMARY PROTEIN SEQUENCE DATABASES

- UniProt/Swiss-Prot

UniProt (Universal Protein Resource) is the world's most comprehensive catalogue of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

UniProt is comprised of three components :

•UniProt Knowledgebase (UniProt)) is the central access point for extensive curated protein information, including function, classification, and cross-reference.

•UniProt Non-redundant Reference (UniRef) databases combine closely related sequences into a single record to speed searches.

•UniProt Archive (UniParc) is a comprehensive repository, reflecting the history of all protein sequences.

UniProt/Swiss-Prot

UniProt contains data that originates from a wide variety of organisms from more than 6,000 different species. Half of the entries come from about 20 organisms, which are the target of many biological studies

1. Homo sapiens
2. Saccharomyces cerevisiae
3. Escherichia coli
4. Mus musculus
5. Rattus norvegicus
6. Bacillus subtilis
7. Caenorhabditis elegans
8. Haemophilus influenzae
9. Schizosaccharomyces pombe
10. Methanococcus jannaschii
11. Bos taurus
12. Drosophila melanogaster
13. Mycobacterium tuberculosis
14. Gallus gallus
15. Arabidopsis thaliana
16. Salmonella typhimurium
17. Xenopus laevis
18. Synechocystis sp. (strain PCC 6803)
19. Sus scrofa
20. Oryctolagus cuniculus

Supplements for UniProt/Swiss-Prot: UniProt/TrEMBL

UniProt/TrEMBL (Translation of EMBL Nucleotide Sequence Database), consists of computer-annotated entries derived from the translation of all coding sequences (CDS) in the EMBL Nucleotide Sequence Database, except for those already included in UniProt/Swiss-Prot.

TrEMBL is split in two main sections :
SP-TrEMBL: contains the entries, which should be eventually incorporated into UniProt/Swiss-Prot

REM-TrEMBL (REMaining TrEMBL): contains the entries that will NOT get included in UniProt/Swiss-Prot. It contains sequences that are either synthetic, truncated, pseudogenes, patented, small fragments or immunoglobulins and T-cell receptors which UniProt/Swiss-Prot are not interested in annotating.

Supplements for UniProt/Swiss-Prot: UniProt/PIR

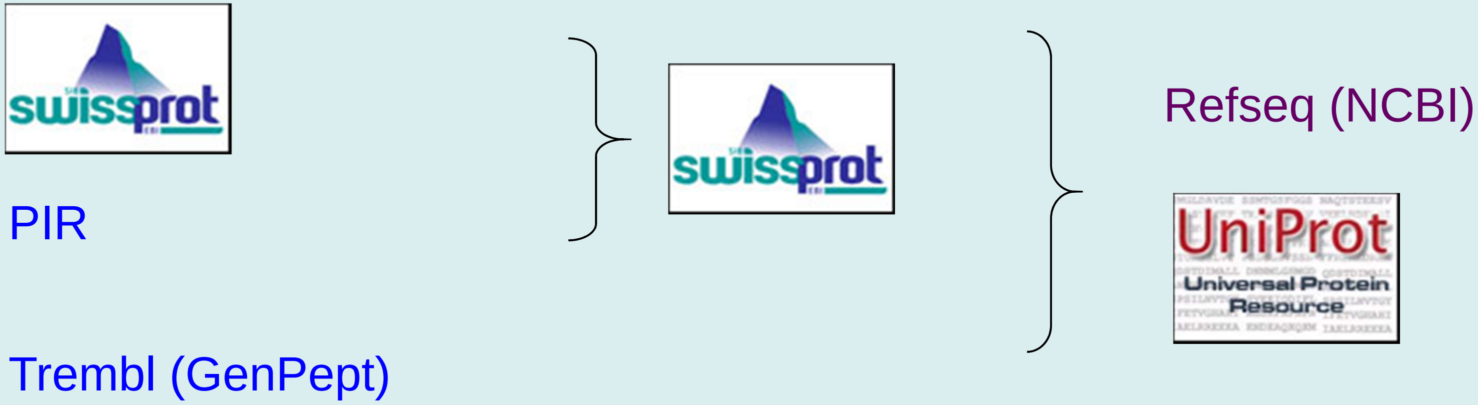
The database is partitioned into four sections :

PIR1: are fully classified by superfamily assignment, fully annotated and fully merged with respect to other entries in PIR1.

PIR2: The annotation content as well as the level of redundancy reduction varies in PIR2 entries. Many entries in PIR2 are merged, classified, and annotated.

PIR3: Entries in PIR3 are not classified, merged or annotated. PIR3 serves as a temporary buffer for new entries.

PIR4: PIR4 was created to include sequences identified as not naturally occurring or expressed, such as known pseudogenes, unexpressed ORFs, synthetic sequences, and non-naturally occurring fusion, crossover or frameshift mutations.



SPECIALISED PROTEIN DATABASES

Proteome Analysis– The Proteome Analysis database has been set up to provide comprehensive statistical and comparative analyses of the predicted proteomes of fully sequenced organisms.

MEROPS – This database provides a catalogue and structure-based classification of peptidases. Families that are thought to have common evolutionary origins and are known or expected to have similar tertiary folds are grouped into clans. The MEROPS database provides sets of files called FamCards and ClanCards describing the individual families and clans. Each FamCard document provides links to other databases.

GPCRDb– This is a database of sequences and other data relevant to the biology of G-protein coupled receptors, components of many different signalling systems in animals.

YPD – *S. Cerevisiae* proteins. Information related to more than 6000 proteins.

ENZYME – Annotated extension of the Enzyme Commission's publication.

LIGAND – linked to the metabolic pathways .

2 – DIMENSIONAL GEL ELECTROPHORESIS DATA –

MASS SPECTROMETRY PROTEIN DATA –

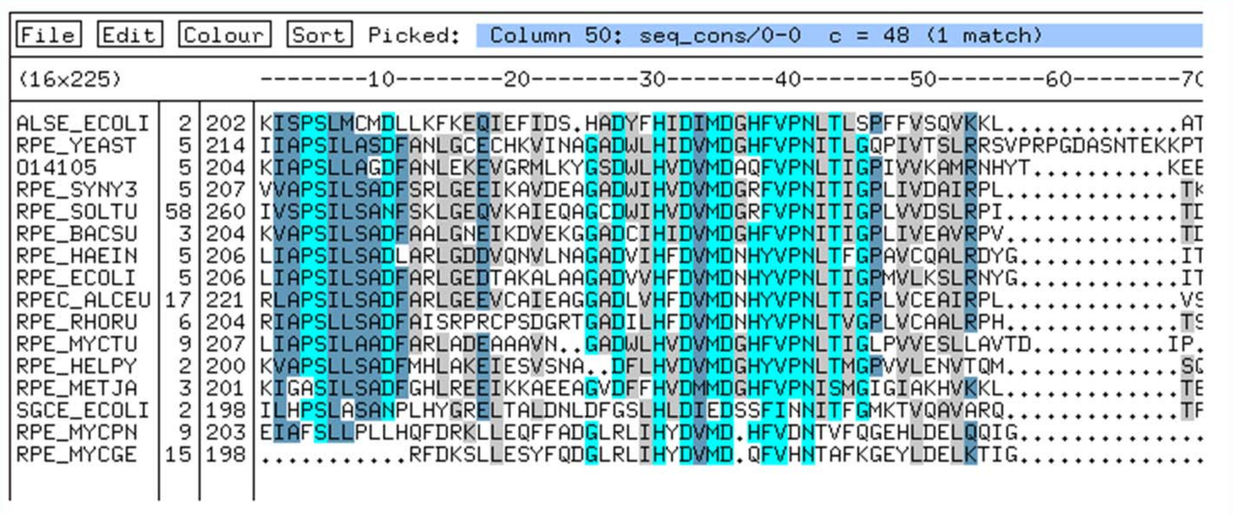
SECONDARY PROTEIN DATABASES: Why?

They contain the study of a wide range of analyses

MSA of homologue sequences present regions of low or no variation: MOTIFS (or blocks, segments or properties)

Motif: Consecutive series of amino acids which are repeated or conserved in the same position of a multiple alignment.

Motifs usually reflect a vital biological role for the structure or the function

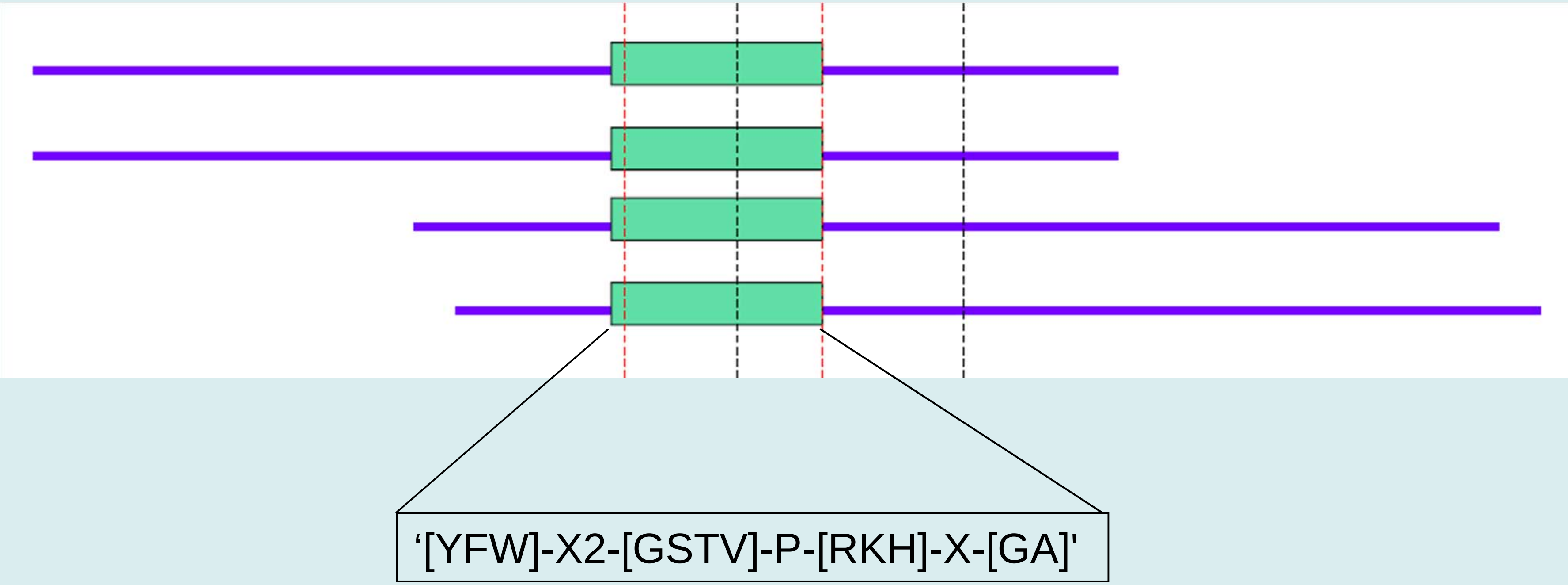


```
Picked:   Column 50: seq_cons/0-0   c = 48 (1 match)

(16x225)              --------10--------20--------30--------40--------50--------60--------70

ALSE_ECOLI   2 202   KISPSLMCMDLLKFKEQIEFIDS.HADYFHIDIMDGHFVPNLTLSPFFVSQVKKL.............AT
RPE_YEAST    5 214   IIAPSILASDFANLGCECHKVINAGADWLHIDVMDGHFVPNITLGQPIVTSLRRSVPRPGDASNTEKKPT
O14105       5 204   KIAPSLLAGDFANLEKEVGRMLKYGSDWLHVDVMDAQFVPNLTIGPIVVKAMRNHYT..........KEE
RPE_SYNY3    5 207   VVAPSILSADFSRLGEEIKAVDEAGADWIHVDVMDGRFVPNITIGPLIVDAIRPL.............TK
RPE_SOLTU   58 260   IVSPSILSANFSKLGEQVKAIEQAGCDWIHVDVMDGRFVPNITIGPLVVDSLRPI.............TI
RPE_BACSU    3 204   KVAPSILSADFAALGNEIKDVEKGGADCIHIDVMDGHFVPNITIGPLIVEAVRPV.............TI
RPE_HAEIN    5 206   LIAPSILSADLARLGDDVQNVLNAGADVIHFDVMDNHYVPNLTFGPAVCQALRDYG............IT
RPE_ECOLI    5 206   LIAPSILSADFARLGEDTAKALAAGADVVHFDVMDNHYVPNLTIGPMVLKSLRNYG............IT
RPEC_ALCEU  17 221   RLAPSILSADFARLGEEVCAIEAGGADLVHFDVMDNHYVPNLTIGPLVCEAIRPL.............VS
RPE_RHORU    6 204   RIAPSLLSADFAISRPRCPSDGRTGADILHFDVMDNHYVPNLTVGPLVCAALRPH.............TS
RPE_MYCTU    9 207   LIAPSILAADFARLADEAAAVN..GADWLHVDVMDGHFVPNLTIGLPVVESLLAVTD..........IP.
RPE_HELPY    2 200   KVAPSLLSADFMHLAKEIESVSNA..DFLHVDVMDGHYVPNLTMGPVVLENVTQM.............SC
RPE_METJA    3 201   KIGASILSADFGHLREEIKKAEEAGVDFFHVDMMDGHFVPNISMGIGIAKHVKKL.............TE
SGCE_ECOLI   2 198   ILHPSLASANPLHYGRELTALDNLDFGSLHLDIEDSSFINNITFGMKTVQAVARQ.............TF
RPE_MYCPN    9 203   EIAFSLLPLLHQFDRKLLEQFFADGLRLIHYDVMD.HFVDNTVFQGEHLDELQQIG..............
RPE_MYCGE   15 198   ...........RFDKSLLESYFQDGLRLIHYDVMD.QFVHNTAFKGEYLDELKTIG..............
```

PATTERNS

Using motifs, databases have been compiled on PATTERNS (or regular expressions or rules)

Pattern: simple consensus expression of a conserved region in a multiple sequence alignment

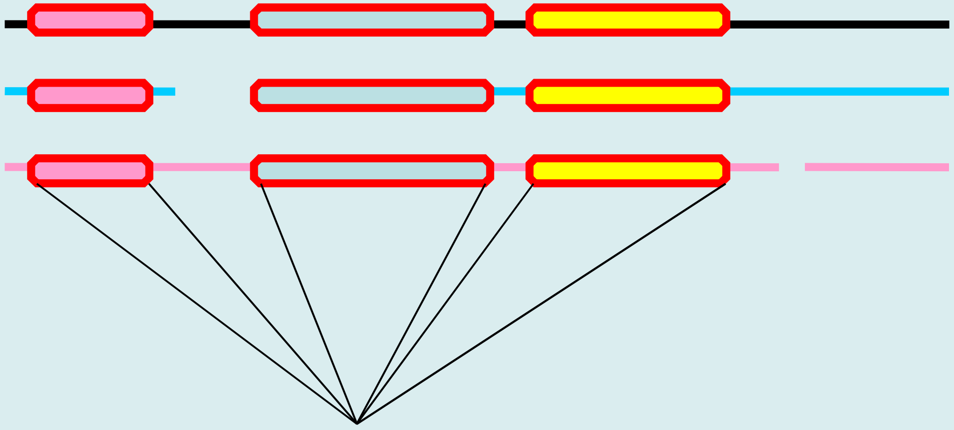


‘[YFW]-X2-[GSTV]-P-[RKH]-X-[GA]'

FINGERPRINTS

From sequence alignment without gaps, databases can be compiled on FINGERPRINTS

FINGERPRINTS: Motifs used to construct a print belonging to a protein family.

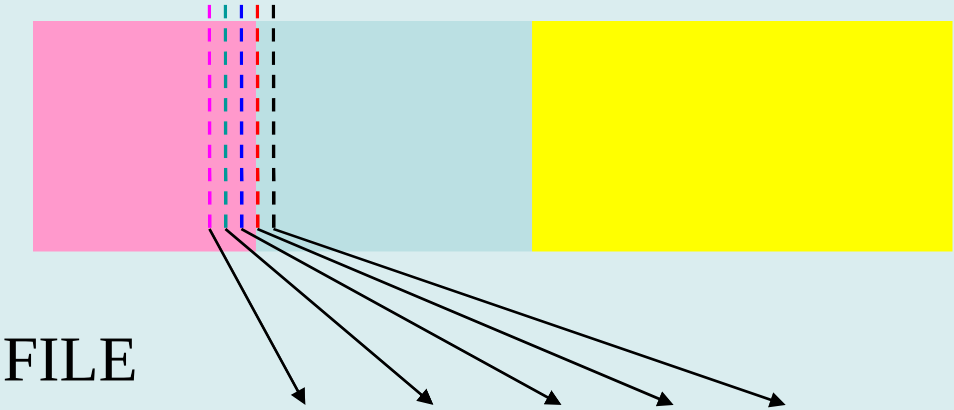


Fingerprint

PROFILES

Profiles are based on variable regions between conserved motifs. They also contain important information on sequences.

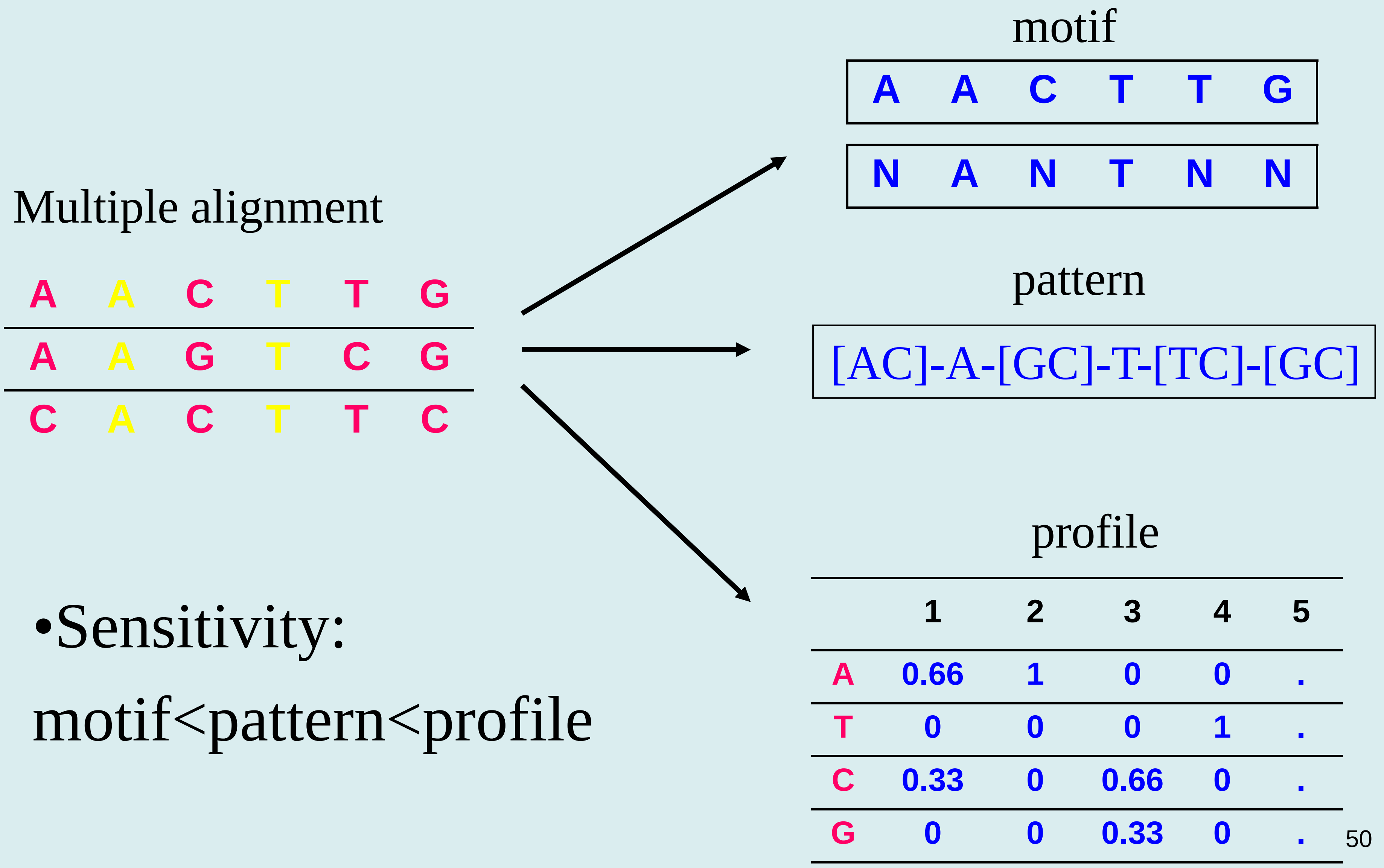
PROFILES: Position specific scoring matrices (PSSM) including information on multiple alignments. Allowed residues in fixed, conserved and degenerated positions are defined.



PROFILE

	1	2	3	4	5
A	0.66	1	0	0	.
T	0	0	0	1	.
C	0.33	0	0.66	0	.
G	0	0	0.33	0	.

Motif-Pattern-Profile Sensitivity



Multiple alignment

A	A	C	T	T	G
A	A	G	T	C	G
C	A	C	T	T	C

motif

A	A	C	T	T	G
N	A	N	T	N	N

pattern

[AC]-A-[GC]-T-[TC]-[GC]

profile

	1	2	3	4	5
A	0.66	1	0	0	.
T	0	0	0	1	.
C	0.33	0	0.66	0	.
G	0	0	0.33	0	.

- Sensitivity:

motif<pattern<profile

SECONDARY PROTEIN DATABASES

PROSITE – contains biologically significant sites and patterns formulated in such a way that with appropriate computational tools it can rapidly and reliably identify to which family of proteins the new sequence belongs.
Generalised profiles are remarkably similar to the specific type of Hidden Markov Models.

PRINTS – A different approach to pattern recognition, termed "fingerprinting" is used by this database. Within a sequence alignment, it is usual to find not one, but several motifs that characterise the aligned family. Diagnostically, it makes sense to use many, or all, of the conserved regions to build a family signature. In a database search, there is then a greater chance of identifying a distant relative, whether or not all parts of the signature are matched. The ability to tolerate mismatches, both at the level of residues within individual motifs, and at the level of motifs within the fingerprint as a whole, renders fingerprinting a powerful diagnostic technique.

Pfam –The methodology used by Pfam to create protein family or domain signatures is **Hidden Markov Models** (HMMs). HMMs are closely related to profiles, but are based on probability theory methods.
One feature that distinguishes HMMs and profiles from regular expressions and fingerprints is that the formers allow the full extent of a domain to be identified in a sequence. The biggest drawback of Pfam is its lack of biological information (annotation) of the protein families.

BLOCKS - Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. The blocks for the Blocks Database are made automatically by looking for the most highly conserved regions in groups of proteins documented in InterPro.

WHEN AND HOW TO APPLY SECONDARY PROTEIN DATABASES

Diagnostically, the most commonly used secondary protein databases, PROSITE, PRINTS and Pfam have different areas of optimum application owing to the different strengths and weaknesses of their underlying analysis methods .

Examples:

- regular expressions are likely to be unreliable in the identification of members of highly divergent super-families whereas profiles and HMMs excel.
- fingerprints perform relatively poorly in the diagnosis of very short motifs whereas regular expressions do well.
- profiles and HMMs are less likely to give specific sub-family diagnoses whereas fingerprints excel.

Special databases

Pfam→ divergent domains
PROSITE → functional sites
PRINTS → families, specialising in hierarchical definitions from super-family down to sub-family levels in order to pin-point specific functions
ProDom/SMART → domain identification

COMPLEX SECONDARY DATABASES- InterPro

WARNING: Unfortunately, these secondary databases do not share the same formats and nomenclature as each other, which makes the use of all of them in an automated way difficult.

To solve this problem: InterPro is an integration of:

PROSITE
PRINTS
Pfam
ProDom

Uses of InterPro:

1. computational functional classification of newly determined sequences that lack biochemical characterisation.
2. identifying those families and domains for which the existing discriminators are not optimal and could therefore be usefully supplemented with an alternative pattern. Example: where a regular expression identifies large numbers of false matches it could be useful to develop an HMM, or where a Pfam entry covers a vast super-family it could be beneficial to develop discrete family fingerprints, and so on.
3. InterPro highlight key areas where none of the databases has yet made a contribution.

STRUCTURE DATABASES

SCOP - aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.

- Family: Identities>30%
- Superfamily: similar functional and structural characteristics
- Foldings: Identical folding whether homologue or not.

CATH – Class, Architecture, Topology, Homology. Hierarchical classification of protein domain structures

- Class: secondary structure content and package
- Architecture: secondary structure disposition without connectivity
- Topology: global shape and connectivity of the secondary structure
- Homology: Domains with more than35% identity
- Sequence: High order grouping within homology

PDBsum – provides an at-a-glance overview of every macromolecular structure deposited in the Protein Data Bank (PDB), giving schematic diagrams of the molecules in each structure and of the interactions between them.

STRUCTURE DATABASES (cont.)

Protein Data Bank (PDB). The number of known protein structures is increasing very rapidly and these are available through the Protein Data Bank (PDB).

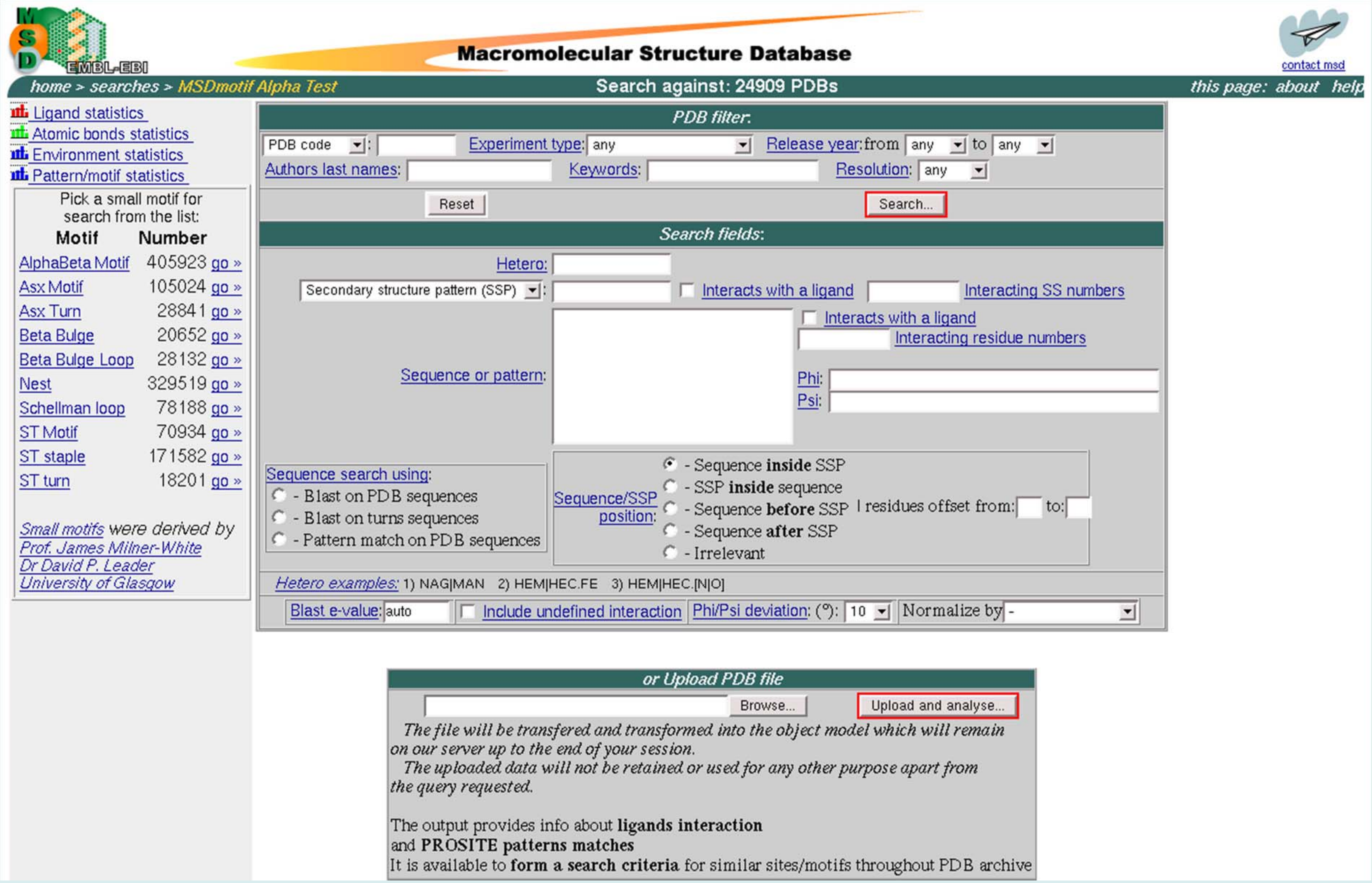
Nucleic Acid Database (NDB) is the database for structural information about nucleic acid molecules.

SRS3D is an integrated environment that allows the end-user to quickly and easily retrieve/visualise sequence structure and also feature data from primary, secondary and tertiary protein databases.

MSD is the European Project for the management and distribution of data on macromolecular structures. they have close ties with the Research Collaboratory for Structural Bioinformatics (RCSB) who in collaboration with MSD maintain and administer the PDB.

CCDC provides a database of structures of ‘small molecules', of interest to biologists concerned with protein-ligand interactions.

Example of Structure Database: MSD


MSD EMBL-EBI
Macromolecular Structure Database
contact msd
home > searches > MSDmotif Alpha Test
Search against: 24909 PDBs
this page: about help

Ligand statistics
Atomic bonds statistics
Environment statistics
Pattern/motif statistics

Pick a small motif for search from the list:

Motif	Number	
AlphaBeta Motif	405923	go »
Asx Motif	105024	go »
Asx Turn	28841	go »
Beta Bulge	20652	go »
Beta Bulge Loop	28132	go »
Nest	329519	go »
Schellman loop	78188	go »
ST Motif	70934	go »
ST staple	171582	go »
ST turn	18201	go »

Small motifs were derived by
Prof. James Milner-White
Dr David P. Leader
University of Glasgow

PDB filter:

PDB code: Experiment type: any Release year: from any to any
Authors last names: Keywords: Resolution: any

Reset Search...

Search fields:

Hetero:
Secondary structure pattern (SSP): □ Interacts with a ligand Interacting SS numbers
Sequence or pattern: □ Interacts with a ligand Interacting residue numbers
Phi:
Psi:

Sequence search using:
○ - Blast on PDB sequences
○ - Blast on turns sequences
○ - Pattern match on PDB sequences

Sequence/SSP position:
● - Sequence **inside** SSP
○ - SSP **inside** sequence
○ - Sequence **before** SSP | residues offset from: to:
○ - Sequence **after** SSP
○ - Irrelevant

Hetero examples: 1) NAG|MAN 2) HEM|HEC.FE 3) HEM|HEC.[N|O]

Blast e-value: auto □ Include undefined interaction Phi/Psi deviation: (°): 10 Normalize by -

or Upload PDB file

Browse... Upload and analyse...

The file will be transfered and transformed into the object model which will remain on our server up to the end of your session.
The uploaded data will not be retained or used for any other purpose apart from the query requested.

The output provides info about **ligands interaction**
and **PROSITE patterns matches**
It is available to **form a search criteria** for similar sites/motifs throughout PDB archive

6. MICROARRAY DATABASES

Microarray data analysis

Clustering and class prediction are typical methods currently used in gene expression data analysis .

Expression Profiler developed at the EBI.

Microarray databases

ArrayExpress – microarrays storage from EBI.

ChipDB Gene expression database.

ExpressDB – Relational expression database of RNA in yeasts and E. coli

Gene Expression Atlas – Expression samples in humans and mice.

Gene Expression Omnibus – NCBI database for public acces of genie expression data.

GermOnline – Genes involved in mitosis and meiosis.

Human Gene Expression Index (HuGE Index) - Human gene expression in normal tissues.

List Of Lists Annotated (LOLA) –identification and correlation of genes paired by microarray experiments.

GENIC EXPRESSION DATABASES

Allow comparison among different experiments from different laboratories

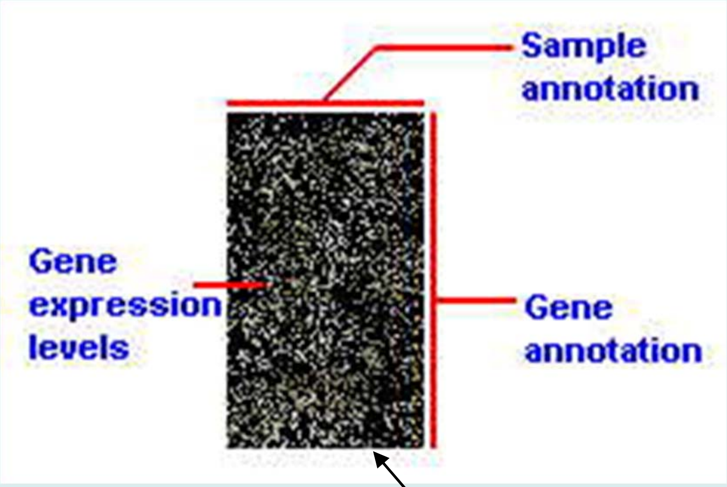
International Standard Efforts:

- Standard definition of minimun data group (MIAME)

- Development of a relational database (ArrayExpress)

- Interchangeable data format (MAGE-ML, formato XML)

- Development of an onthology for microarrays



Genic expression matrix

To be continued in other doctorate course….