Chapter I Major sequence repositories

- Description of molecular biology databases (USA) /databanks (in UK) :
  - Heterogeneity in their aims, shapes and usage
- Databases architecture :
  - Discrepancy between these databases/databanks and the state of the art in data management
- Databases integration

Why build a database?

- Need to store genome and protein sequences and other related data from biological and computational analysis
- Sequence annotation and comparison allow the identification of family and functional relationships among proteins
- Searching databases is often the first step in the study of a new molecule : needs for
  - a query system (a catalogue, indexed files, SQL)
  - an information system (Entrez)

What is a database?

- A structured collection of data held in computer storage (includes a software to make it accessible in a variety of ways)
  - Structured
  - Searchable (index)
  - Updated periodically
  - Cross-referenced (hyperlinks)
- Database management : the organization and manipulation of data
  - Database management systems (DBMS)
- Text mining is limited by the lack of coherence between databases and semantics problem
  - e.g. gene ontology

Database system

ENA database :

Type of data

- Nucleotides
- EST
- Assembly
- CDS

Dealing with DB

Data storage management

- Flat-file
- **Relational**
- ACEDB
- OPM

Data quality

- Data consistency
- nonredundancy
- updates
- curation
- annotation

In-house

- Secure
- Non shareable

Online

- Shared

- Static
- Dynamic
- Data on analysis tools
- Annotation Reference

Flat-file Sequence formats

- Many molecular Databases consist of text (or GIF) files :
  - Sequences edited with a word processing program
- Database entry are structured explicitly
  - They differ not only in their syntax but also semantically
    - The sequence in a GenBank entry is between ORIGIN and //
    - The sequence in a EMBL entry is between SQ and //

- or implicitly : FASTA format required by analysis tools

```
>sp|P01215|GLHA_HUMAN Glycoprotein hormones
alpha chain OS=Homo sapiens GN=CGA PE=1 SV=1
MDYYRKYAAIFLVTLSVFLHVLHSAPDVQDCPECTLQENPFFSQP
GAPILQCMGCCFSRAYPTPLRSKKTMLVQKNVTSESTCCVAKSYN
RVTVMGGFKVENHTACHCSTCYYHKS
```

Relational Database management systems (DBMS)

- Databases are composed of a set of tables linked by a shared information referred as the key
- Easily searchable

protab1			
Protein-code	Protein-name	Length	Species-origin
P1001	Hemoglobin	145	Bovine
P1002	Hemoglobin	136	Ovine
P1003	Eye Lens Protein	234	Human
.....			

protab2	
Protein-code	Protein-sequence
P1001	MDRTTHGFDLKLLSPRTVNQWLMLALFFGHS...
P1002	MDKTSHGFEIKLLTPKKLQQWLMIAIYFGHT...
P1003	SRTHEEEGKLMQWPPRPLYIALFTEPPYP...
.....	

```
SELECT protein-code, protein-name
FROM protab1
WHERE species-origin='Bovine'
```

Entry : P1001 Hemoglobin

- Sequence retrieval using SQL

protab1			
Protein-code	Protein-name	Length	Species-origin
P1001	Hemoglobin	145	Bovine
P1002	Hemoglobin	136	Ovine
P1003	Eye Lens Protein	234	Human
.....			

protab2	
Protein-code	Protein-sequence
P1001	MDRTTHGFDLKLLSPRTVNQWLMLALFFGHS...
P1002	MDKTSHGFEIKLLTPKKLQQWLMIAIYFGHT...
P1003	SRTHEEEGKLMQWPPRPLYIALFTEPPYP...
.....	

```
SELECT protab1.protein-code, protab1.protein-name,
protab2.protein-sequence

FROM protab1, protab2

WHERE protab1.protein-code = protab2.protein-code

AND protab2.protein-code = 'P1002'
```

P1002 Hemoglobin MDKTSHGFEIKLLTPKK

Arguments against Relational Database management systems for molecular biology DB

- ✓ Database management systems are dispensable ?
  - Data are not subject to modification
  - Cost of porting a flat-file database into a relational database
- ✓ Molecular biology data are often very complex
- ✓ Most molecular biologists are not familiar with database query language such as Structured Query language (SQL)

Conflicts arisen by integrating data from distinct origins:

- Descriptive (model used for items)
- Heterogenity according to the system used for DB management
- Semantics :importance of the vocabulary used

Databases use different terms ⇒ a nomenclature defining the relationships between terms has been created by a consortium of scientists (EBI) = ontology

Specific vocabulary for :

- Biological function (cell/molecular)
- Biological or cell process
- Cell component (localisation, complexes)

Cellular process:

- Cell networks
  - metabolic pathways
  - regulatory networks
  - signal transduction
- Cell division
- Cell structure

Major Molecular biology databanks

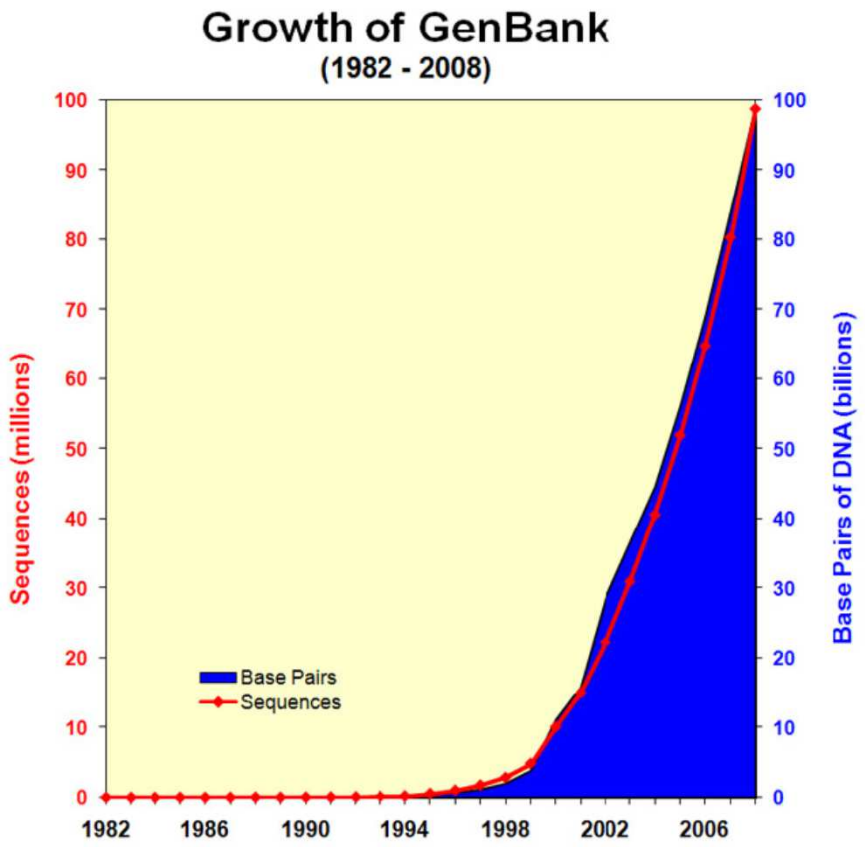
- Primary Sequence Databases
  - Collection and storage of sequences in computer files : ENA (EMBL/GenBank), SwissProt
  - Data integration for reducing redundancy:RefSeq
  - Information retrieval and analysis tools
    - convenient user interface (lack of common scheme)
- Secondary Databases
  - Reorganization of raw data to enable new predictions (domain family): InterPro
  - Species-specific genome databases: SGD, Ensembl
- Web-base interface systems

  Form-based query and browsing interfaces: Entrez

Nucleic Acid Sequence databanks

- ENA (EMBL) European Bioinformatics Institute (EBI), U.K.
- GENBANK : Los Alamos National Laboratory 1979 → National Center for Biotechnology Information (NCBI), U.S.A.
- DDBJ (National Institute of Genetics, Japan) (1984)

} **INSDC**



1982: 680×10^3 base pairs
606 sequences

2000: 10×10^9 base pairs
9×10^6 sequences

2008: 85×10^9 base pairs
83×10^6 sequences

2015: 1.3×10^{12} base pairs
608.5×10^6 sequences

Since 1998 it has been doubling every 22 months

Database entry

- Entries contain :
  - Data:
    - The locus (mnemonic identification code) and accession number (unique identifier for a sequence record)
    - The sequence
  - Additional information referred to as annotation/ref
    - Date of creation and modification (DT)
    - Source organism (OS)
    - Description (DE) : Biological function
    - Literature references (RX)
    - Key cross-reference words (KW)
    - Links to other databases (DR)
    - Features about the sequence (coding sequence, regulatory regions, ...) (FT)

EBI format

- The information is organised into fields (and subfields), each with a keyword (as search indexes)

Listeria ivanovii sod gene

Identifier

Accession number (AC)

Features (FT)

Protein sequence

```
LOCUS       LISOD         756 bp    DNA             BCT       30-JUN-1993
DEFINITION  L.ivanovii sod gene for superoxide dismutase.
ACCESSION   X64011 S78972
VERSION     X64011.1  GI:44010
KEYWORDS    sod gene; superoxide dismutase.
SOURCE      Listeria ivanovii.
ORGANISM    Listeria ivanovii
            Bacteria; Firmicutes; Bacillus/Clostridium group;
            Bacillus/Staphylococcus group; Listeria.
REFERENCE   1  (bases 1 to 756)
  AUTHORS   Haas,A. and Goebel,W.
  TITLE     Cloning of a superoxide dismutase gene from Listeria ivanovii by
            functional complementation in Escherichia coli and characterization
            of the gene product
  JOURNAL   Mol. Gen. Genet. 231 (2), 313-322 (1992)
  MEDLINE   92149371
REFERENCE   2  (bases 1 to 756)
  AUTHORS   Kreft,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-APR-1992) J. Kreft, Institut f. Mikrobiologie,
            Universitaet Wuerzburg, Biozentrum Am Hubland, 8700 Wuerzburg, FRG
FEATURES             Location/Qualifiers
     source          1..756
                     /organism="Listeria ivanovii"
                     /strain="ATCC 19119"
                     /db_xref="taxon:1638"
     RBS             95..100
                     /gene="sod"
     gene            95..746
                     /gene="sod"
     CDS             109..717
                     /gene="sod"
                     /EC_number="1.15.1.1"
                     /codon_start=1
                     /transl_table=11
                     /product="superoxide dismutase"
                     /protein_id="CAA45406.1"
                     /db_xref="GI:44011"
                     /db_xref="SWISS-PROT:P28763"
                     /translation="MTYELPKLPYTYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVS
                     GHAELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGGAPTGNLK
                     AAIESEFGTFDEFKEKFNAAAAARFGSGWAWLVVNNGKLEIVSTANQDSPLSEGKTPV
                     LGLDVWEHAYYLKFQNRRPEYIDTFWNVINWDERNKRFDAAK"
     terminator      723..746
                     /gene="sod"
BASE COUNT      247 a    136 c    151 g    222 t
ORIGIN
        1 cgttatttaa ggtgttacat agttctatgg aaatagggtc tataccttc gccttacaat
       61 gtaatttctt ttcacataaa taataaacaa tccgaggagg aatttttaat gacttacgaa
      121 ttaccaaaat taccttatac ttatgatgct ttgaagccga attttgataa agaaacaatg
```

- Some definitions :
  - Locus: Unique string of 10 letters and numbers. Not maintained amongst databases
  - Accession: A unique identifier to that record, citable entity; does not change when record is updated)
  - Version: new system where the accession and version play the same function as the accession and gi number
  - Nucleotide or protein gi : GenInfo identifier (gi); a unique integer which will change every time the sequence changes
  - PID : protein identifier: g, e or d prefix to gi number. Can have one or two on one CDS

Type of files

- From one-gene investigators
  - A well annotated genomic DNA segment or cDNA
  - A mitochondrial DNA or virus
- From population/phylogenetic analysis
  - rRNA amplicon from environmental sampling
- From genome Centers
  - Gene expression (Expressed Sequence tags or ESTs)
- Genome sequencing projects
  - WGS eg., AAAA01072744 (project version contig number)

- Expressed Sequence tags : Short (300-500 bp) single reads from mRNA (cDNA) ; they represent a snapshot of what is expressed in a given tissue and developmental stage
- Sequenced tagged Site (STS) : 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 rather than mRNA
- HTG : High Throughput Genome Sequences are unfinished genome sequencing effort records. (see Genbank HTC division for unfinished cDNA sequencing
- WGS : contigs from ongoing Whole Shotgun sequencing projects

Protein Sequence databanks (PSD)

Historical view

- 1965: Atlas of protein sequences, Margaret Dayhoff (National Biomedical Research Foundation, Washington DC)
  - 50 entries
  - paper copy until 1987, then digital version
- 1984 : creation of PIR-NBRF
  - Collaboration with MIPS and JIPID (PIR-IPSD)
- 1986: creation of SwissProt
  - Collaboration with SIB and EBI
- 2003 : Creation of UniProt 5.4×10^5 entries in 2014

- Experimentally determined (sequence and 3D-structure) and translated Amino acid sequences

- **SwissProt** (Swiss Institute of Bioinformatics and EBI)
  : minimal redundancy, well-annotated and cross-referencing
- **PIR (Protein Identification Resource)-International PSD**
  - National Biomedical Research Foundation, (NBRF) U.S.A.
  - Martinsried Institute for protein Sequences (MIPS), Munich
  - Japan International Protein Information Database (JIPID), Tsukuba
- **GENPEPT** (NCBI)
- **TREMBL** (EBI/SIB)

} *Automated translation of CDS in Genbank and EMBL databases*

Typical SwissProt entry (α subunit of human chorionic gonadotropin

```
ID   GLHA_HUMAN     STANDARD;      PRT;   116 AA.
AC   P01215;
DT   21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT   21-JUL-1986, sequence version 1.
DT   05-SEP-2006, entry version 74.
DE   Glycoprotein hormones alpha chain precursor
(Anterior pituitary
DE   glycoprotein hormones common subunit alpha)
(Follitropin alpha chain)
DE   (Follicle-stimulating hormone alpha chain) (FSH-
alpha) (Lutropin alpha
DE   chain) (Luteinizing hormone alpha chain) (LSH-
alpha) (Thyrotropin
DE   alpha chain) (Thyroid-stimulating hormone alpha
chain) (TSH-alpha)
DE   (Choriogonadotropin alpha chain) (Chorionic
gonadotrophin alpha
DE   subunit) (CG-alpha).
GN   Name=CGA;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates;
Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE.
RX   MEDLINE=80011660; PubMed=481597;
RA   Fiddes J.C., Goodman H.M.;
RT   "Isolation, cloning and sequence analysis of the
cDNA for the alpha-
RT   subunit of human chorionic gonadotropin.";
RL   Nature 281:351-356(1979).
RN   [2]
RP   NUCLEOTIDE SEQUENCE.
RX   MEDLINE=94254248; PubMed=8196184;
RA   Miyoshi I., Kasai N., Hayashizaki Y.;
RT   "Structure and regulation of human thyroid-
stimulating hormone (TSH)
```

```
DR   RZPD-ProtExp; RZPDo834A0815; -.
DR   GO; GO:0005625; C:soluble fraction; TAS.
DR   GO; GO:0005179; F:hormone activity; TAS.
DR   GO; GO:0007267; P:cell-cell signaling; TAS.
DR   GO; GO:0007165; P:signal transduction; TAS.
DR   InterPro; IPR002400; GF_cysknot.
DR   InterPro; IPR000476; Glyco_hormone.
DR   PANTHER; PTHR11509; Glyco_hormone; 1.
DR   Pfam; PF00236; Hormone_6; 1.
DR   PRINTS; PR00438; GFCYSKNOT.
DR   PRINTS; PR00274; GLYCOHORMONE.
DR   ProDom; PD002047; Glyco_hormone; 1.
DR   SMART; SM00067; GHA; 1.
DR   PROSITE; PS00779; GLYCO_HORMONE_ALPHA_1; 1.
KW   3D-structure; Direct protein sequencing; Glycoprotein;
Hormone;
KW   Signal.
```

Feature's table

```
FT   SIGNAL        1     24
FT   CHAIN        25    116       Glycoprotein hormones alpha
chain.
FT                                 /FTId=PRO_0000011640.
FT   CARBOHYD     76     76       N-linked (GlcNAc...).
FT                                 /FTId=CAR_000036.
FT   CARBOHYD    102    102       N-linked (GlcNAc...).
FT                                 /FTId=CAR_000037.
FT   DISULFID     31     55
FT   DISULFID     34     84
FT   DISULFID     52    106
FT   DISULFID     56    108
FT   DISULFID     83    111
FT   CONFLICT     29     29       Q -> E (in Ref. 9).
FT   CONFLICT    108    109       CS -> SC (in Ref. 6 and 7).
FT   STRAND       30     30
SQ   SEQUENCE   116 AA;  13075 MW;  F0623CD8CC90CFCD CRC64;
     MDYYRKYAAI FLVTLSVFLH VLHSAPDVQD CPECTLQENP FFSQPGAPIL
QCMGCCFSRA
     YPTPLRSKKT MLVQKNVTSE STCCVAKSYN RVTVMGGFKV ENHTACHCST
CYYHKS
//
```

- **Composite sequence databases**
  - **UniProt** (Universal Protein resource)
    - # *UniProt Knowledgebase* (UniProtKB): Swiss-Prot +PIR +TrEMBL
    - # *UniMES*: metagenomic and environmental sequences
    - # *UniProt Non-redundant Reference* (UniRef): closely related sequences are combined into a single record (100, 90, 50%)
    - # *Uniprot Archive* (UniParc): comprehensive and non redundant repository (UPI: same protein from UniProt, RefSeq, PDB, …)

  - **RefSeq** (NCBI): provides separate and linked records for the genomic DNA, mRNA, and the protein arising from the transcript:NM_000646.1

Protein structure databases

- The Protein Data Bank (PDB)
  - Research Collaboratory for Structural Bioinformatics (RCSB)
  - 46151 entries (82% from X-ray crystallography, 7% from NMR)
  - Entry ID : 1EUV (annotation, coordinates and connectivities)

RCSB PDB PROTEIN DATA BANK

A MEMBER OF THE PDB
An Information Portal to Biological Macromolecular Structures
As of Tuesday Sep 02, 2008 there are 52821 Structures | PDB Statistics

CONTACT US | HELP | PRINT PAGE PDB ID or keyword Author Site Search | Advanced Search

Home | Search | Structure | Queries

- 1EUV
- Download Files
- FASTA Sequence
- Download Original Files
- Display Files
- Display Molecule
- Structural Reports
- External Links
- Structure Analysis
- Help

Quick Tips:
To view **sequence details** of this structure click on the *Sequence Details* tab above the summary page.

Are you missing data updates? The PDB archive has moved to ftp://ftp.wwpdb.org.
For more information click here.

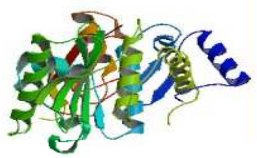
Help | Structure Summary | Biology & Chemistry | Materials & Methods | Sequence Details | Geometry

1euv DOI 10.2210/pdb1euv/pdb

Red - Derived Information

Title	X-RAY STRUCTURE OF THE C-TERMINAL ULP1 PROTEASE DOMAIN IN COMPLEX WITH SMT3, THE YEAST ORTHOLOG OF SUMO.
Authors	Mossessova, E., Lima, C.D.
Primary Citation	Mossessova, E., Lima, C.D. (2000) Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast. *Mol.Cell* **5:** *865-876* [Abstract]
History	Deposition 2000-04-17 Release 2000-06-07
Experimental Method	Type X-RAY DIFFRACTION Data N/A
Parameters	Resolution[Å] 1.60; R-Value 0.193 (obs.); R-Free 0.251; Space Group $P\,2_1\,2_1\,2$
Unit Cell	Length [Å] a 125.77 b 53.17 c 54.26; Angles [°] alpha 90.00 beta 90.00 gamma 90.00
Molecular Description Asymmetric Unit	Polymer: 1 Molecule: ULP1 PROTEASE Fragment: C-TERMINAL PROTEASE DOMAIN Chains: A Polymer: 2 Molecule: UBITQUTIN-LIKE PROTEIN SMT3 Fragment: SMT3 RESIDUES 13-98 Chains: B
Classification	Hydrolase
Source	Polymer: 1 Scientific Name: Saccharomyces cerevisiae Common Name: Yeast Expression system: Escherichia coli Polymer:

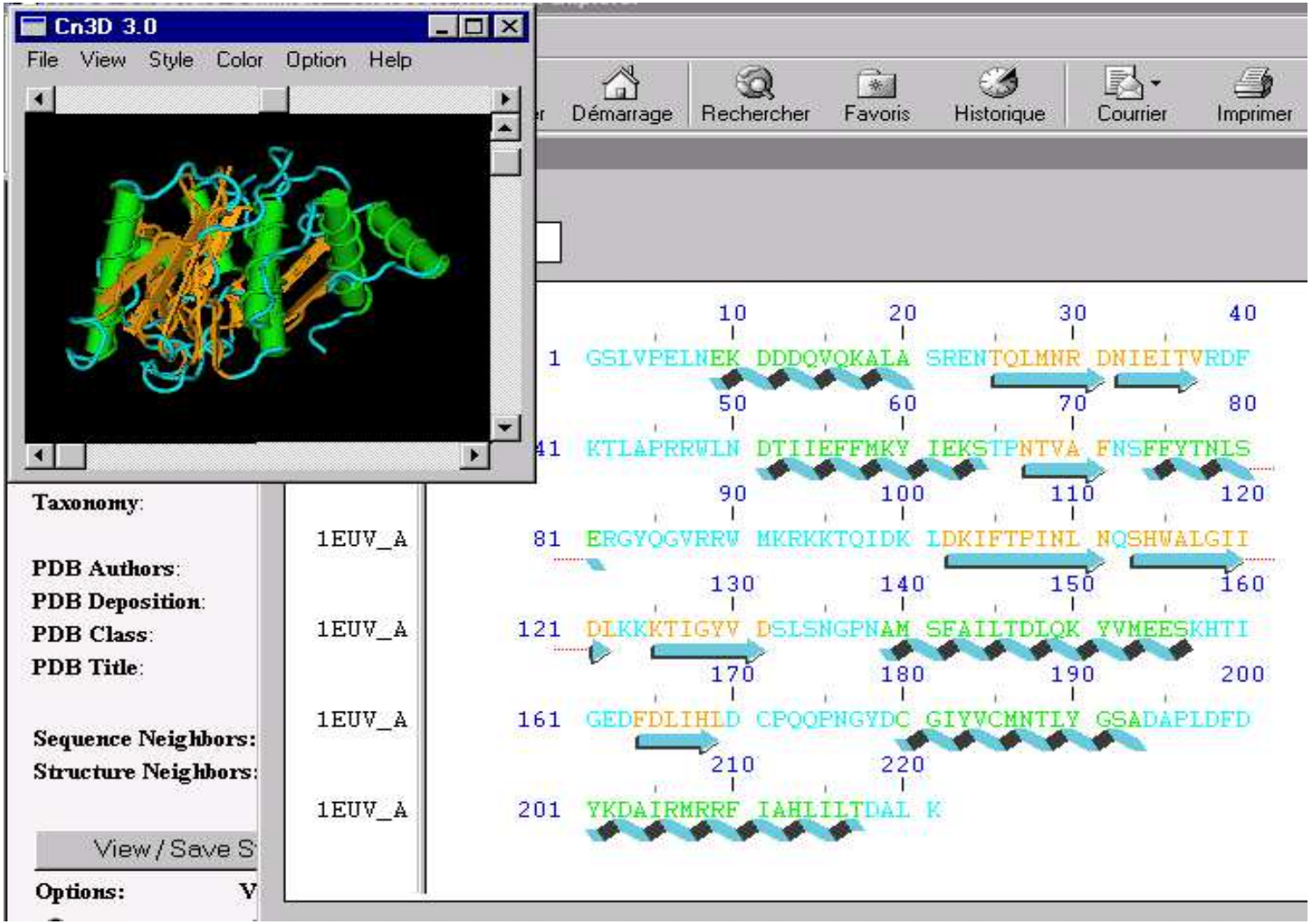
Images and Visualization
<< Biological Molecule >>



Display Options
KiNG
Jmol
WebMol
MBT SimpleViewer*
MBT Protein Workshop
QuickPDB
All Images
* Capable of displaying biological molecules.

- Molecular Modelling Database (MMDB) and PDBSum

Different view of the PDB data with additional features



MMDB File from NCBI *Entrez*

- **The Structural Classification of proteins (SCOP)**
  - Proteins of known structure are classified according to their evolutionary and structural relationships
  - Hierarchical classification into families, superfamilies, folds and classes

- **The Class, Architecture, Topology, Homology database (CATH)**

Secondary databases

Collection of conserved amino-acid patterns which are derived from primary sequence databases by different methods

- Motif/pattern

Short sequence (up to 10 residues) that is often found within the active site of proteins that have a similar biochemical activity.

D-K-T-G-T-[LIVM]-[TI] phosphorylation site of P-type ATPase

Note: ≠ from structural motif : bZIP, HLH

- Domain /profile

Combination of several secondary structural elements (50-100 residues) or conserved aligned residues

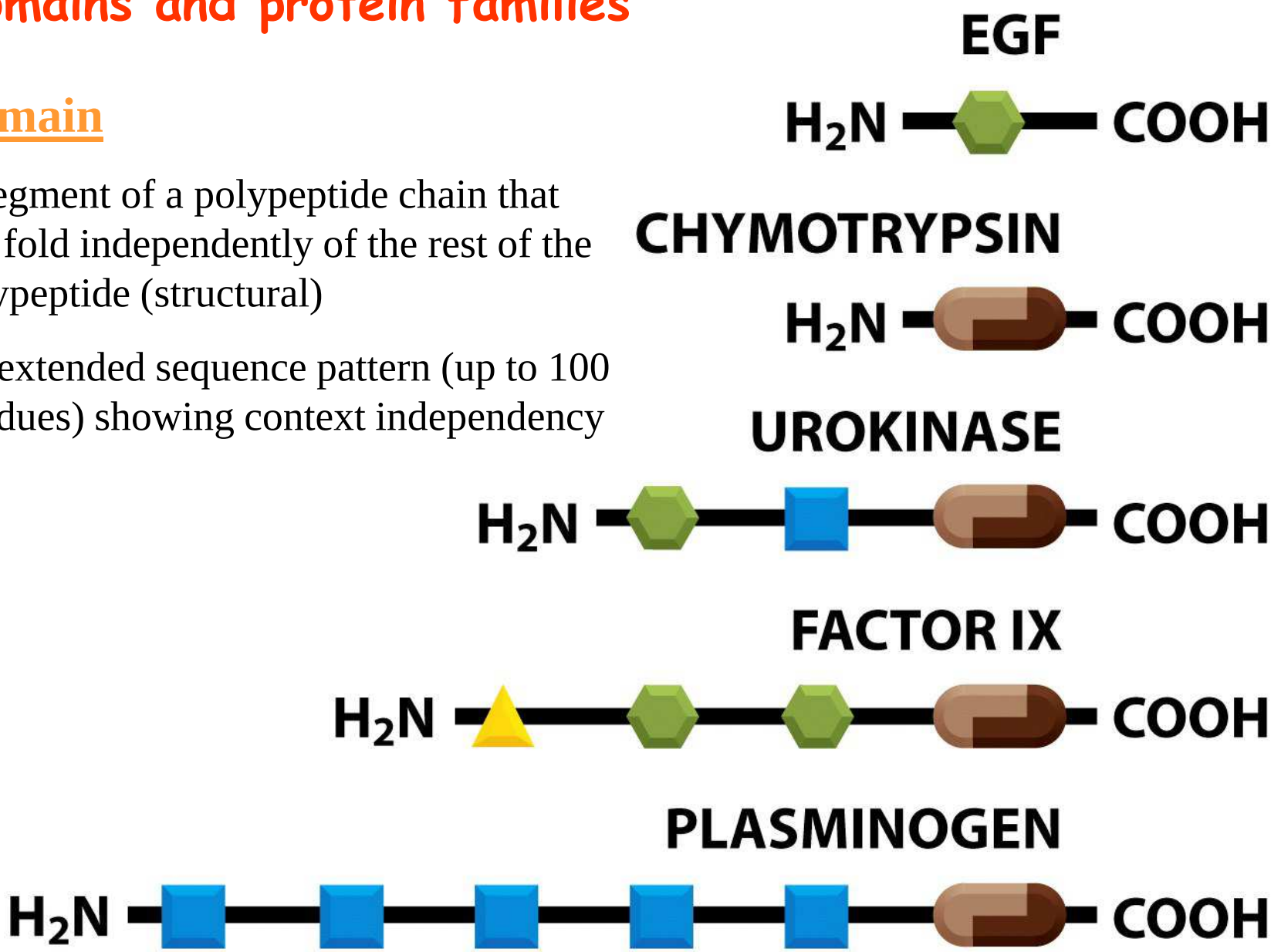
- Family/HMM

Domains and protein families

Domain

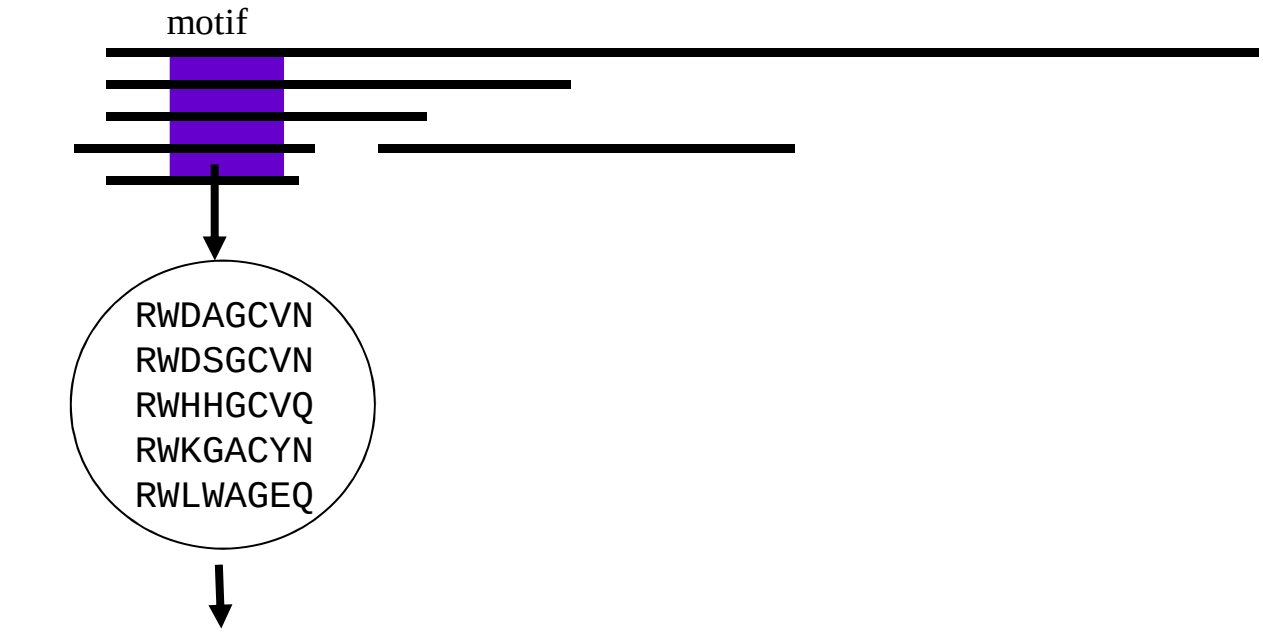
A segment of a polypeptide chain that can fold independently of the rest of the polypeptide (structural)

An extended sequence pattern (up to 100 residues) showing context independency



PROSITE

- A database of protein active sites
- Motives are expressed as regular expressions or profiles
- A great tool for predicting the existence of active sites in an unknown protein



R-W-x(2)-[AG]-C-x-[NQ]

regular expression (PROSITE)

Example of a PROSITE entry

Documentation entry

accession

>PDOC00139 PS00154 **ATPASE_E1_E2** E1-E2 ATPases phosphorylation site [pattern].

```
378 - 384  DKTGTLT
```

Documentation

E1-E2 ATPases (also known as P-type) are cation transport ATPases which form an aspartyl phosphate intermediate in the course of ATP hydrolysis. ATPases which belong to this family are listed below [1,2,3].

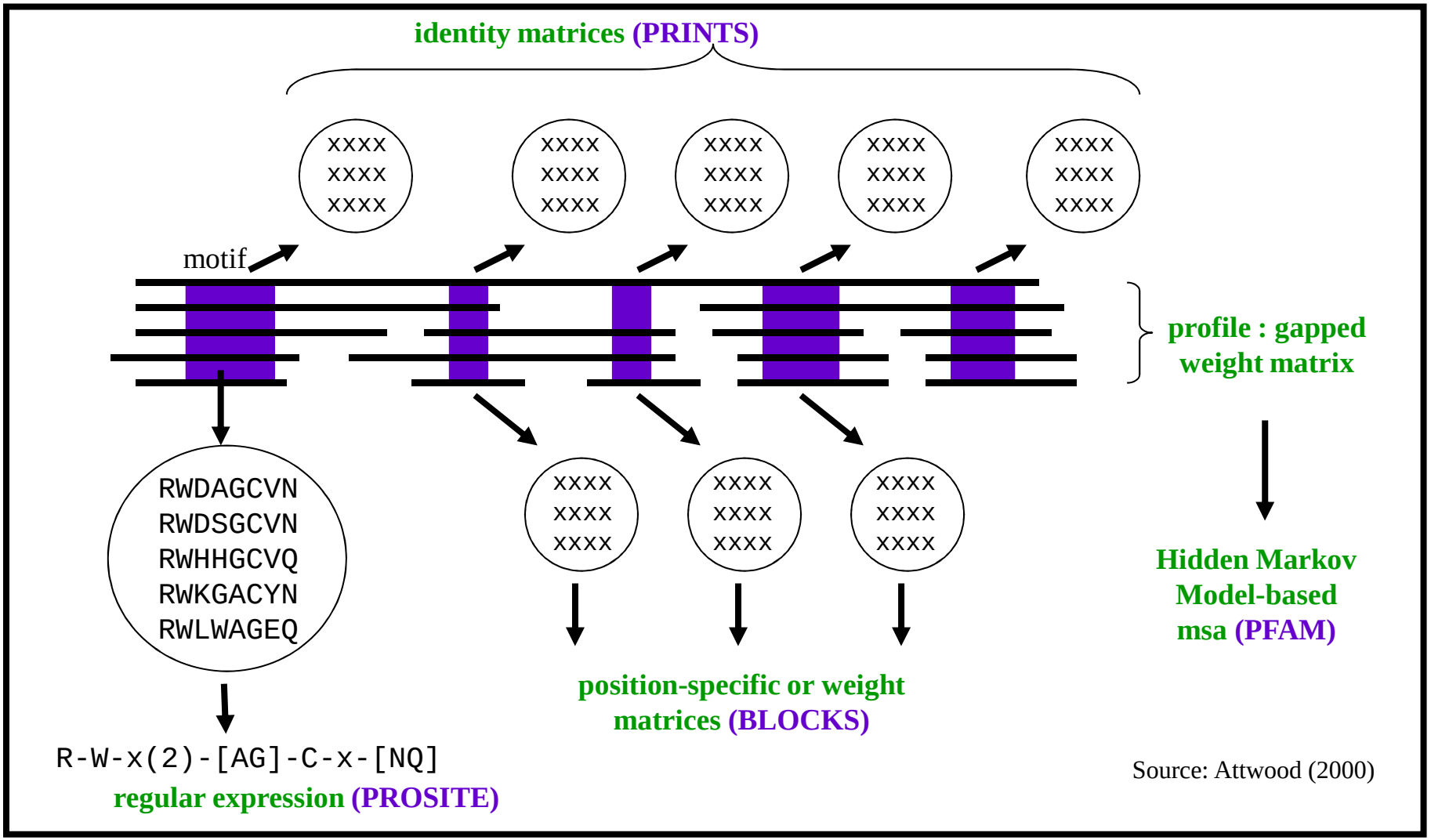
- Fungal and plant plasma membrane (H+) ATPases [reviewed in 4].
- Vertebrate (Na+, K+) ATPases (sodium pump) [reviewed in 5,6].
- Gastric (K+, H+) ATPases (proton pump).
- Calcium (Ca++) ATPases (calcium pump) from the sarcoplasmic reticulum (SR), the endoplasmic reticulum (ER) and the plasma membrane.

Description of pattern(s) and/or profile(s)

Consensus pattern	D-K-T-G-T-[LIVM]-[TI] [D is phosphorylated]
Sequences known to belong to this class detected by the pattern	ALL.

PRINTS and BLOCKS

Several local alignments(fingerprints) expressed as frequency (PRINTS) or weight (BLOCKS) scoring matrices



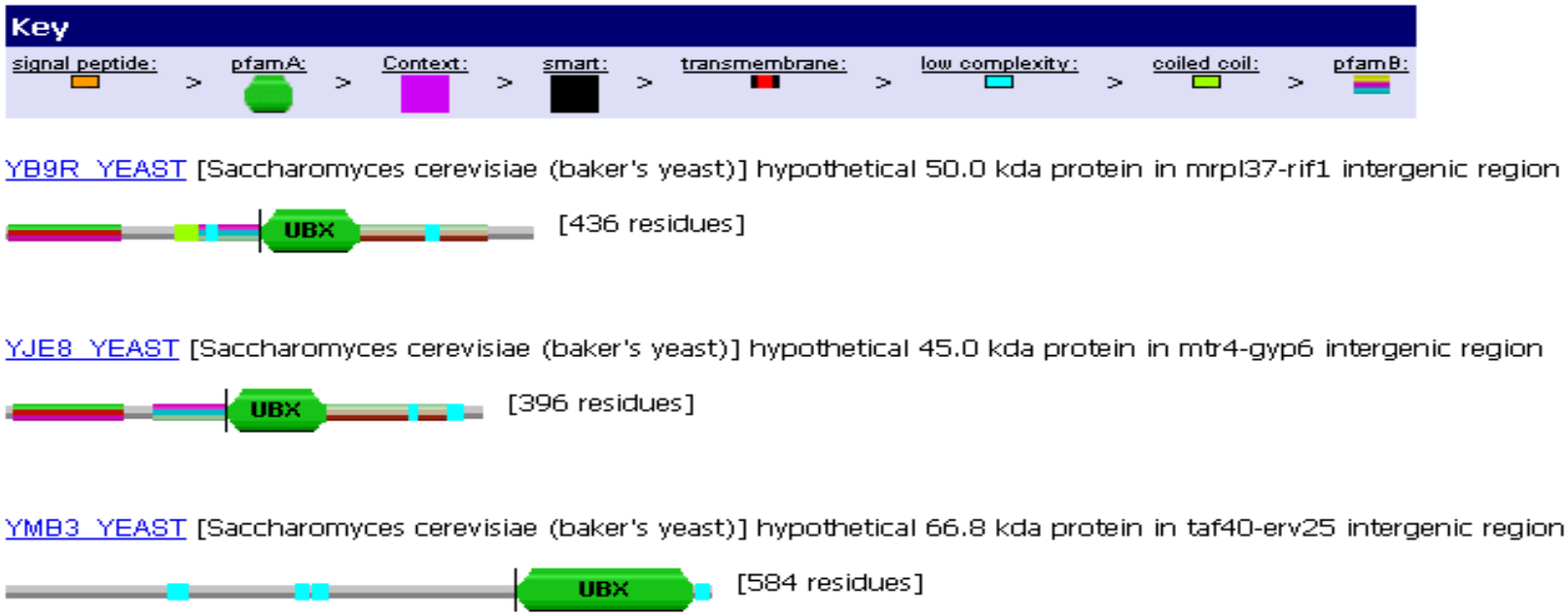
Source: Attwood (2000)

PFAM (Protein Family) (Sanger Institute)

- A database of protein families defined as domains (contiguous segments of protein sequences)
- 12273 multiple alignments including gaps (2011)
- Based on Hidden Markov Models (HMMs)

Pfam entry for the UBX domain (Sanger)

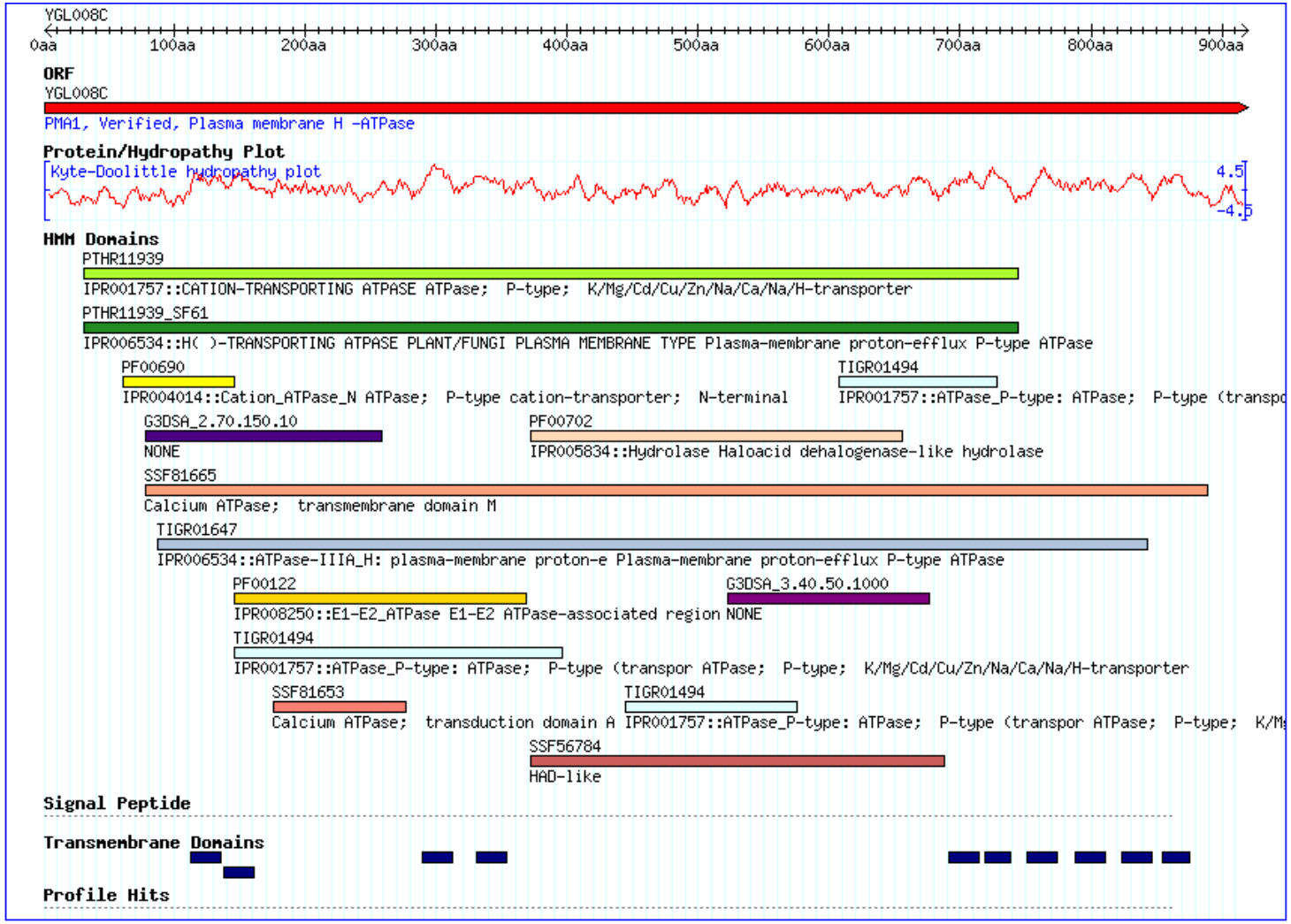
A) Schematic representation



B) Multiple sequence alignment of UBX containing proteins

```
O77047/367-444      ADAVGAIAVVFKLPSGTRLERRFN.QTDSVLDVYHYLFCHPDSPDE.....................
Q06682/414-494      NKPGITTRIQIRTGDGSRLVRRFNALEDTVRTIYEVIKTEMDGFADSR...................
O14048/348-426      SPGPNVTRIQIRMPNGARFIRRFS.LTDPVSKVYAYVKGVAEGADKQP...................
O35987/290-369      NEAEPTTNIQIRLADGGRLVQKFN.HSHRISDIRLFIVDARPAMAATS...................
O81456/229-308      DETVPTTSIQLRLADGTRLVAKFN.HHHTVNDIRGFIDSSRPGASLN....................
O23394/342-423      DPAAPTTSIQLRLADGTRLVSRFN.NHHTVRDVRGFIDASRPGGSKE....................
SHP1_YEAST/343-422  EPKQGDTSIQIRYANGKREVLHCN.STDTVKFLYEHVTSNANTDPSRN...................
O23283/641-723      CDRSVVCSLCVRFPDGRRKQRKFL.KSEPIQLLWSFCYSHIDESEKKA...................
O23283/298-380      CDRSVVCSICVRFPDGRRKQRKFL.KSEPIQLLWSFCYSHMEESEKKE....................
O74498/323-425      SEDE.PARLSIRFPDGSRAVRRFK.KDDTVESVYNYVDYMLFEKEEPEEFGRATSSS..........
Q12229/355-452      SSDKDASKVAIRLENGQRLVRKFD.ASLPTEEIYAFVELQLHDMLNSENDTLPVYQPANY.......
YMB3_YEAST/426-570  ETTGKQATLQFRTSSGKRFVKKFP.SMTTLYQIYQSIGCHIYLAVYSSDPAEWSNALQDKIRQLSAD
YB9R_YEAST/211-292  KLHSSKCVLQIRMTDGKTLKHEFN.SSETLNDVRKWVDVNRTDGDCP.....................
YJE8_YEAST/185-266  FLAQNYCTLQLKLPNGYTISNTFP.PQTKLHKVRMWLDYNCYDDGTP.....................
Q9ZW74/311-392      SKKASDVHLNIRLPDGSSLQEKFS.VTSILRMVKDYVNSNQTIGLGA....................
Y33K_HUMAN/209-294  KREYDQCRIQVRLPDGTSLTQTFR.AREQLAAVRLYVELHRGEELGGGQDP.................
YOJ8_CAEEL/277-358  AVPSDRCRLQVRLPDGSTFVEEFP.SNDVLNSLVEIIRQKPSIAGTT.....................
Q92575/335-416      RERSTVARIQFRLPDGSSFTNQFP.SDAPLEEARQFAAQTVGNTYGN.....................
O82483/384-465      EKGPDVTQVLVRFPNGERKGRMFK.SETKIQTLYDYVDSLGLLDTE.....................
```

- Interpro (Integrative Protein) database
  - Composite database composed of Pfam + Prints + ProDom + Smart + PROSITE (relational DBMS)
  - Version 33 contains 21749 entries and 14633 families



Genome Resources

- The Genomics, proteomics and Bioinformatics Knowledge Base : http://www.123genomics.com
- The Ensembl project: www.ensembl.org/index.html
- SGD : *Saccharomyces* genome database Stanford (USA)
  - genetics and chromosomal maps
  - phenotypes
  - micro-array transcription analysis, two-hybrid protein-protein interaction
  - evolutionary relationships with related yeast species
- FlyBase: a database for Drosophila genetics and molecular biology
- The Plant Genome Information Resource (PlantGDB)
- Online Mendelian Inheritance in Man (OMIM)

  Johns Hopkins University, School of Medicine , Baltimore, USA

Ensembl databases

Genome assembly

Gene annotation

Comparative genomics

Regulation

Variation

Data retrieval systems

Aims: Find new and relevant information

Constraints:

- Not too much, but everything on the topic
- Easy to use
- Fast response
- Linked to analysis tools

- Databanks' own browser
  - SIB provides EXPASY (Expert Protein Analysis System) for UNIPROT
- Entrez
  - Integration system provided by the NCBI (http://www.ncbi.nlm.nih.gov)
- SRS (Sequence Retrieval System)
  - Developped within EMBnet (EBI) , service stopped in 2013

Entrez

- Owner interface
- Boolean operators (AND, OR, NOT)
- Field names within brackets homo sapiens [organism]
- History
- Limits
- Pre-computed similarity searches (neighbors) are available for most database records producing a list of related sequences, structure neighbors, as well as related articles.

The Entrez database and analysis tools

Medical Subject Headings : comprehensive controlled vocabulary for the purpose of indexing journal articles and books in the life sciences