Bioinformatics

Jiří Damborský
National Center for Biomolecular Research

jiri@chemi.muni.cz, ph. 41129 377, Kotlarska 2, bld. 7, 2nd floor

Bioinformatics - what is it?

- The term bioinformatics is used to encompass almost all computer applications in biological sciences.
- Information technology applied to the management and analysis of biological data
- originally - analysis of sequence data (80s)
- presently - also analysis of 3D-structures

Bioinformatics - study material

- Introduction to bioinformatics, T.K. Attwood and D.J. Parry-Smith, Longman, Essex, 1999.
- copy of the slides
- http://www.chemi.muni.cz/~jiri
- http://www.bioinf.man.ac.uk/dbbrowser/bioactivity/prefacefrm.html

Bioinformatics - composition

12 lectures per semester

3 hours per week

- 1st and 2nd hour = lectures - theory
- 3rd hour = practical course on computers

Bioinformatics - lectures

- Introduction
- Information networks
- Protein information resources
- Genome information resources
- DNA sequence analysis
- Pairwise sequence alignment
- Multiple sequence alignment
- Secondary database searching
- Analysis packages
- Protein structure modelling

Bioinformatics - practical training

- Biological databases
- Searching and modelling servers
- Building a sequence search protocol
- Case examples
- Protein structure prediction
- Protein modelling

- Follow-up of lectures

Introduction

- history of sequencing
- what is it Bioinformatics?
- sequence to structure deficit
- genome projects
- why is Bioinformatics important?
- patter recognition and prediction
- folding problem
- sequence analysis
- homo/analogy and ortho/paralogy

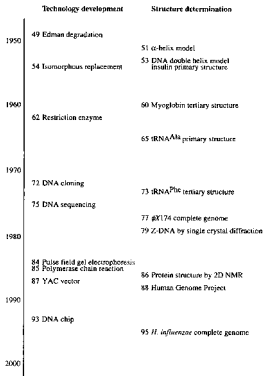
History of sequencing

- Protein sequencing
  - separation of peptides, identification and quantification of amino acids
  - Edman degradation
  - mass-spectrometry - advantage in identification of post-translational modifications
  - 1955 sequencing of peptide insuline
  - 1960 sequencing of enzyme ribonuclease
  - 1980s automated sequencers

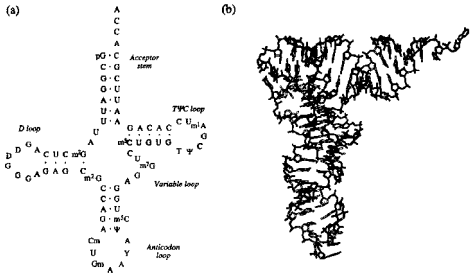
History of sequencing

- Nucleic acid sequencing
  - tRNA - short, could be purified
  - DNA - large (human chromosome 55-250 x 10^6 bp); the longest fragment for sequencing is 500 bp; purification is problematic
  - advent of gene cloning and PCR
  - 1972 DNA cloning
  - 1975 DNA sequencing
  - 1980s and 1990s sequence revolution

The history of technology developments and structure determination



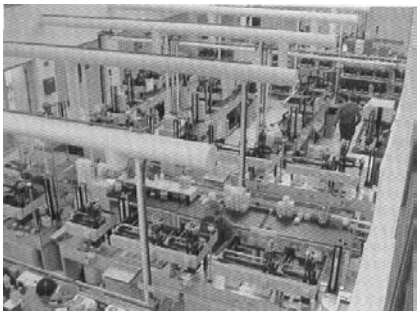
Transfer RNA: (a) primary sequence and (b) tertiary structure



Automatic sequencing machine

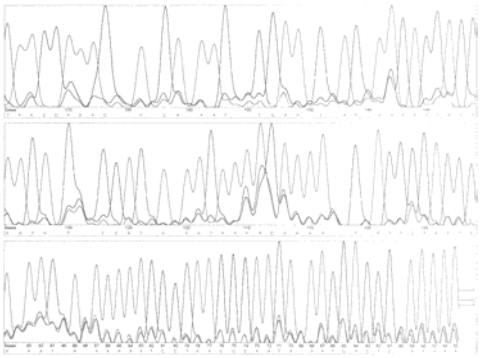
ABI Prism 310, Applied Biosystems

Automated production line in sequencing "factory"



Whitehead Institute, Center for Genome Research, USA

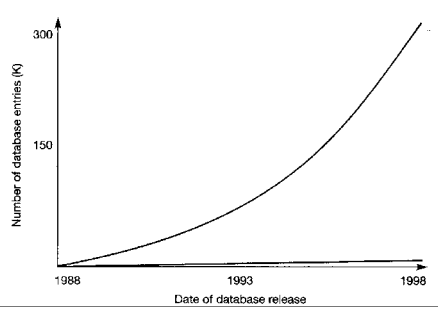
Sequencing chromatogram



What is Bioinformatics?

- improvements in DNA sequencing technologies and computer-based technologies
- originally - analysis of sequence data (1980s)
- presently - also analysis of 3D-structures
- The term bioinformatics is used to encompass almost all computer applications in biological sciences.
- Information technology applied to the management and analysis of biological data.

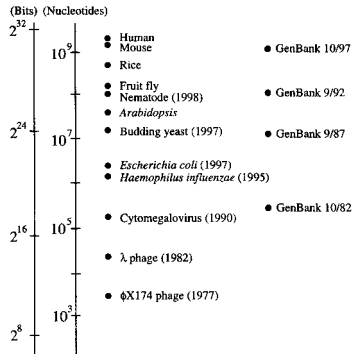
The sequence to structure deficit



Genome projects

- 1977 first complete genome - virus ϕX174, 5000 nucleotides; 11 genes
- 1995 first complete genome of living organism *Haemophilus influenzae*, 1.8 million nucleotides and 1700 genes
- sequencing of model systems: *Escherichia coli, Saccharomyces cerevisiae, Cernorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana, Canis familiaris, Mus Musculus*

The genome size of various species



Comparative genomic analysis of model organisms

	Genome size (Mb)	Gene number	Haploid chromosome number
Bacterium (*Escherichia coli*)	~4	4,403	1
Yeast (*Saccharomyces cerevisiae*)	~12	6,190	16
Worm (*Caenorhabditis elegans*)	97	19,730	6
Fruit Fly (*Drosophila melanogaster*)	120	13,601	4
Mouse (*Mus Musculus*)	3,454	~50,000 (estimated)	20
Human (*Homo sapiens*)	2,910	33,609	23

Human Genome Project

- in mid-1980s initiated Human Genome Project
- estimated 100.000 genes and completion in 2005
- need for automated sequencing and improved computational techniques
- shotgun method

- sequencing of rough draft first
- first draft completed in 2000 by publicly funded the International Consortium Human Genome Project and the company Celera Genomics

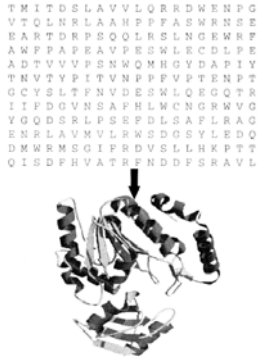
Human Genome Project

- ~33.000 genes
- genes are complex due to alternative splicing
- >1.000.000 proteins (estimated)
- hundreds of genes resulted from horizontal transfer from bacteria (in vertebrate lineage)
- dozen of genes derived from transposable elements (their activity however has declined)
- the mutation rate in male is two-times higher than in female
- >1.400.000 single point polymorphisms (SNPs)

Why is bioinformatics important?

- last 20-30 years - structural biology
- new era - bioinformatics - due to genome projects and sequence/structure deficit
- biological function is not known for about 50% of all genes in every sequenced genome
- role of bioinformatics
  - data management and storage
  - data analysis = conversion of primary sequence to biological knowledge

Pattern recognition *versus* prediction



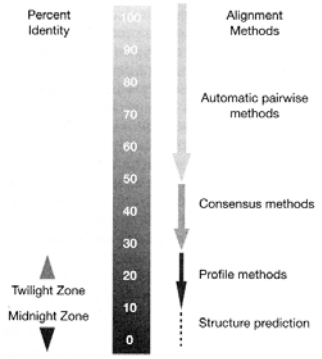
Levels of protein structure

Primary structure:	the linear sequence of amino acids in a protein molecule
Secondary structure:	regions of local regularity within a protein fold (e.g., α-helices, β-turns, β-strands)
Super-secondary structure:	the arrangement of α-helices and/or β-strands into discrete folding units (e.g., β-barrels, βαβ-units, Greek keys, etc.)
Tertiary structure:	the overall fold of a protein sequence, formed by the packing of its secondary and/or super-secondary structure elements
Quaternary structure:	the arrangement of separate protein chains in a protein molecule with more than one subunit
Quinternary structure:	the arrangement of separate molecules, such as in protein–protein or protein–nucleic acid interactions

Homology and analogy

- Sequences are said to be homologous if they are related by divergence from a common ancestor.
- Proteins can share similar folds (e.g., β-barrel) or similar catalytic residues (e.g., serine proteases) without any sequential similarity. Convergence to similar biological solutions from different evolutionary starting points results in analogy.
- Sequence analysis assumes homologous proteins.
- Homology is not a measure of similarity.

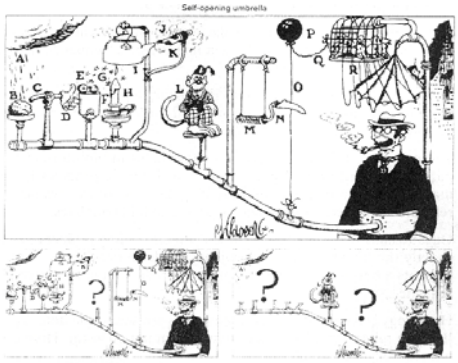
Application areas of different analysis methods



Orthology and paralogy

- Proteins performing the same function in different species - orthologues.
- Proteins performing different, but related functions within same organism - paralogues.
- Sequence comparison of orthologuos proteins - phylogenetic analysis.

Modularity of proteins = difficulties of homology searches



Information networks

- what is the Internet?
- how do computers find each other?
- FTP and Telnet
- what is the Worl Wide Web?
- HTTP, HTML and URL
- EMBnet, EBI, NCBI
- SRS and ENTREZ

What is the Internet?

- Global network of computer networks that link government, academic and business institutions.
- communication by TCP/IP
  (Transmission Control Protocol/Internet Protocol)
- computers - nodes, data - packets
- packets may not be transferred directly from one computer to another

How do computers find each other?

- Each computer is assigned IP address
  147.251.28.2
  machine.site.domain
  bilbo.chemi.muni.cz
- FTP - File Transfer Protocol
- Telnet - remote connection

Example of Internet domains and subdomains

Country-based domains		*Other domains*		*Subdomains*	
Australia	.au	Educational	.edu	Academic	.ac
Denmark	.dk	Commercial	.com	Company	.co
Finland	.fi	Governmental	.gov	Other organisation	.org
France	.fr	Military	.mil	General	.gen
Germany	.de				
Greece	.gr				
Hungary	.hu				
Ireland	.ie				
Israel	.il				
Italy	.it				
Netherlands	.nl				
New Zealand	.nz				
Poland	.pl				
Portugal	.pt				
South Africa	.za				
Spain	.es				
Sweden	.se				
Switzerland	.ch				
United Kingdom	.uk				
USA	.us				

What is the World Wide Web?

- Developed at CERN - the European Laboratory of Particle Physics.
- The purpose was sharing of information.
- Hypermedia based information system.
- The most advanced information system found on the web.
- Very popular - almost synonymous with the Internet.

Web browsers

- Browser is the client communicating with servers using standard protocols.
- Home page is the first point of contact between browser and the server.

Lynx - academic, VT100 terminal
Mosaic - academic, X-windows
Netscape Navigator - commercial
Internet Explorer - commercial

HTTP, HTML and URL

- HTTP - HyperText Transport Protocol
  documents exploited by browsers are written in hypertext and transferred by HTTP
- HTML - HyperText markup Language
  standard language for writing a hypertext
- URL - Uniform Resourse Locator
  unique address for a document
  example: http://www.chemi.muni.cz/~jiri

EMBnet, EBI, NCBI

- 1988 established the network of European biocomputing and bioinformatics laboratories.
- Eliminates the need for multicopies of biology databases and retrieval software.
- Hinxton Hall = Sanger Centre + MRC Human Genome Mapping Project Resource Centre + European Bioinformatics Institute (EBI)
- National Center for Biotechnology Information (NCBI)

SRS, ENTREZ and LinkDB

SRS - The Sequence Retrieval System

- maintained by EBI
- network browser for databases in molecular biology
- allows indexation of flat-file databases
- allows customised search of selected databases
- link databanks: sequence, structure, bibliography, etc.

ENTREZ

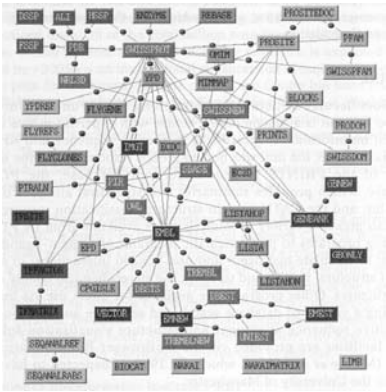
- integrates databases of NCBI
- less flexible then SRS
- valuable concept of neighbouring
- link databanks: DNA and protein sequences, genome data, structural data, PubMed bibliography

SRS, ENTREZ and LinkDB

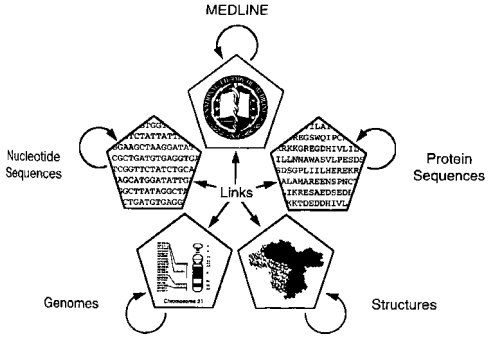
LinkDB

- maintained by Institute for Chemical Reseach, Japan
- network browser for databases in DBGET and KEGG (Kyoto encyclopedia of genes and genomes)
- link databanks: sequence, motifs, structure, amino acid properties, ligands, metabolic pathways

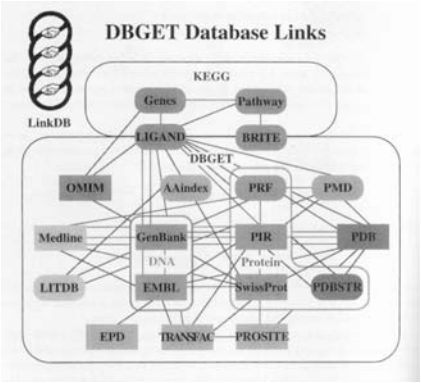
Network of databases linked via SRS



Network of databases linked via ENTREZ



Network of databases linked via LinkDB



Protein information resources

- biological databases - introduction
- primary protein sequence databases
- composite protein sequence databases
- secondary databases
- composite secondary databases
- protein structure databases
- protein structure classification databases

Biological databases - introduction

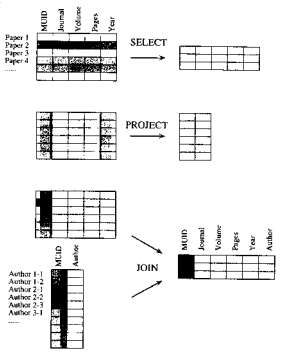
- Vast amounts of data produced - databases must be established for storage of the data.
- Databases must be maintained and disseminated together with the analysis tools.
- Classification of databases
  - flat files
  - relational
  - object-oriented

  - primary
  - secondary
  - composite

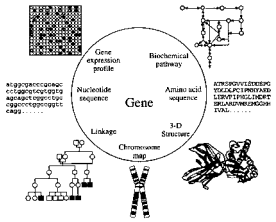
Entry from a flat file database

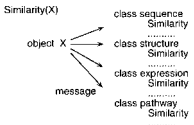
```
LOCUS       DRODPPC      4001 bp    mRNA            INV       15-MAR-1990
DEFINITION  D.melanogaster decapentaplegic gene complex (DPP-C), complete cds.
ACCESSION   M30116
NID         g157291
KEYWORDS    .
SOURCE      D.melanogaster, cDNA to mRNA.
  ORGANISM  Drosophila melanogaster
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Arthropoda;
            Tracheata; Insecta; Pterygota; Diptera; Brachycera; Muscomorpha;
            Ephydroidea; Drosophilidae; Drosophila.
REFERENCE   1  (bases 1 to 4001)
  AUTHORS   Padgett,R.W., St Johnston,R.D. and Gelbart,W.M.
  TITLE     A transcript from a Drosophila pattern gene predicts a protein
            homologous to the transforming growth factor-beta family
  JOURNAL   Nature 325, 81-84 (1987)
  MEDLINE   87090408
COMMENT     The initiation codon could be at either 1188-1190 or 1587-1589.
FEATURES             Location/Qualifiers
     source          1..4001
                     /organism="Drosophila melanogaster"
                     /db_xref="taxon:7227"
     mRNA            <1..3918
                     /gene="dpp"
                     /note="decapentaplegic protein mRNA"
                     /db_xref="FlyBase:FBgn0000490"
     gene            1..4001
                     /note="decapentaplegic"
                     /gene="dpp"
                     /allele=""
                     /db_xref="FlyBase:FBgn0000490"
     CDS             1188..2954
                     /gene="dpp"
                     /note="decapentaplegic protein (1188 could be 1587)"
                     /codon_start=1
                     /db_xref="FlyBase:FBgn0000490"
                     /db_xref="PID:g157292"
                     /translation="MRAWLLLLAVLATFQTIVRVASTEDISQRFIAAIAPVAAHIPLA
                     SASGSGSGRSGSRSVGASTSTALAKAFNPFSEPASFSDSDKSHRSKTNKKPSKSDANR
                     ..........
                     LGYDAYYCHGKCPFPLADHFNSTNHAVVQTLVNNMNPGKVPKACCVPTQLDSVAMLYL
                     NDQSTVVLKNYQEMTVVGCGCR"
BASE COUNT     1170 a   1078 c    956 g    797 t
ORIGIN
        1 gtcgttcaac agcgctgatc gagtttaaat ctataccgaa atgagcggcg gaaagtgagc
       61 cacttggcgt gaacccaaag ctttcgaggc aaattctcgg acccccatat acaaatatcg
      121 gaagaagtac cgaacagttt cgcgacgcga agcgttaaga tcgtcaaaag atctccgtgc
      181 ggaaacaaag aaattgaggc actattaaga gattgctgct gtgcgcgagt gtgtgtcttc
      241 agctgggtgt gtggaatgtc aactgacggg ttgtaaaggg aaaccctgaa atccgaacgg
      301 ccagccaaag caaataaagc tgtgaatacg aattaagtac aacaaacagt tactgaaaca
      361 gatacagctt cggattcgaa tagagaaaca gatactggag atgcccccag aaacaattca
      421 attgcaaata tagtgcgttg cgcgagtgcc agtggaaaaa tatgtggatt acctgcgaac
      481 cgtccgccca aggagccgcc gggtgacagg tgtacccccc aggataccaa cccgagccca
      541 gaccgagatc cacatccaga tcccgaccgc agggtgctag tgtgtcatgt gccgcggcat
      601 accgaccgca gccacatcca ccgaccaggt gcgcctcgaa tgcggcaaca caatttcaaa
          ..........
     3841 aactgtataa acaaaacgta tgccctataa atatatgaat aactatctac atcgttatgc
     3901 gttctaagct aagctcgaat aaatccgtac acgttaatta atctagaatc gtaagaccta
     3961 acgcgtaagc tcagcatgtt ggataaatta atagaaacga g
//
```

Relational database

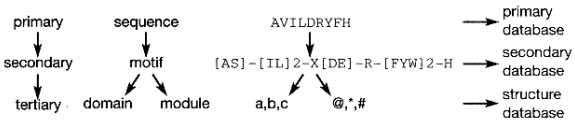


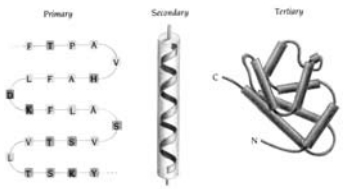
Object-oriented database





Levels of protein structure and corresponding databases





Primary protein sequence databases

- PIR
- MIPS
- SWISS-PROT
- TrEMBL
- NRL-3D

Store biomolecular sequences and annotations.

Primary protein sequence databases

- PIR - Protein Sequence Database
  - 1960s by Margaret Dayhoff
  - maintained by international consortium
  - four sections PIR1-PIR4
    PIR1 - fully classified and annotated entries
    PIR2 - preliminary entries
    PIR3 - unverified entries
    PIR4 - conceptual translations of artefactual sequences, non-transcribed, non-translated
- MIPS - Martinsried Institute for Protein Sequences
  - collects and processes sequence data for PIR

Primary protein sequence databases

- SWISS-PROT
  - University Geneva → EBI → Swiss Inst. of Bioinformatics
  - high-level annotations including description of the function, structure and domains, post-translational modifications, variants, etc.
  - annotated manually (high quality)
  - automatically annotated = TrEMBL
  - minimally redundant
  - interlinked with many other sources
  - efficient searching of selected fields only
  - most widely used protein sequences database

Primary protein sequence databases

- TrEMBL - Translated EMBL
  - computer-annotated supplement of SWISS-PROT
  - contains translations of all coding sequences in EMBL
  - SP-TrEMBL (SWISS-PROT TrEMBL), REM-TrEMBL
- NRL-3D
  - produced by PIR from sequences extracted from Brookhaven Protein Databank (PDB)
  - annotations in PIR format including structural information extracted from PDB: secondary elements, active site AAs, experimental method, resolution
  - makes sequence information in PDB searchable by keywords and similarity

Entry from the SWISS-PROT database

```
ID   DECA_DROME     STANDARD;      PRT;   588 AA.
AC   P07713;
DT   01-APR-1988 (REL. 07, CREATED)
DT   01-APR-1988 (REL. 07, LAST SEQUENCE UPDATE)
DT   01-FEB-1995 (REL. 31, LAST ANNOTATION UPDATE)
DE   DECAPENTAPLEGIC PROTEIN PRECURSOR (DPP-C PROTEIN).
GN   DPP.
OS   DROSOPHILA MELANOGASTER (FRUIT FLY).
OC   EUKARYOTA; METAZOA; ARTHROPODA; INSECTA; DIPTERA.
RN   [1]
RP   SEQUENCE FROM N.A.
RM   87090408
RA   PADGETT R.W., ST JOHNSTON R.D., GELBART W.M.;
RL   NATURE 325:81-84(1987).
RN   [2]
RP   CHARACTERIZATION, AND SEQUENCE OF 457-476.
RM   90258853
RA   PANGANIBAN G.E.F., RASHKA K.E., NEITZEL M.D., HOFFMANN F.M.;
RL   MOL. CELL. BIOL. 10:2669-2677(1990).
CC   -!- FUNCTION: DPP IS REQUIRED FOR THE PROPER DEVELOPMENT OF THE
CC       EMBRYONIC DORSAL HYPODERM, FOR VIABILITY OF LARVAE AND FOR CELL
CC       VIABILITY OF THE EPITHELIAL CELLS IN THE IMAGINAL DISKS.
CC   -!- SUBUNIT: HOMODIMER, DISULFIDE-LINKED.
CC   -!- SIMILARITY: TO OTHER GROWTH FACTORS OF THE TGF-BETA FAMILY.
DR   EMBL; M30116; DMDPPC.
DR   PIR; A26158; A26158.
DR   HSSP; P08112; 1TFG.
DR   FLYBASE; FBGN0000490; DPP.
DR   PROSITE; PS00250; TGF_BETA.
KW   GROWTH FACTOR; DIFFERENTIATION; SIGNAL.
FT   SIGNAL        1      ?        POTENTIAL.
FT   PROPEP        ?    456
FT   CHAIN       457    588        DECAPENTAPLEGIC PROTEIN.
FT   DISULFID    487    553        BY SIMILARITY.
FT   DISULFID    516    585        BY SIMILARITY.
FT   DISULFID    520    587        BY SIMILARITY.
FT   DISULFID    552    552        INTERCHAIN (BY SIMILARITY).
FT   CARBOHYD    120    120        POTENTIAL.
FT   CARBOHYD    342    342        POTENTIAL.
FT   CARBOHYD    377    377        POTENTIAL.
FT   CARBOHYD    529    529        POTENTIAL.
SQ   SEQUENCE   588 AA;  65850 MW;  1768420 CN;
     MRAWLLLLAV LATFQTIVRV ASTEDISQRF IAAIAPVAAH IPLASASGSG SGRSGSRSVG
     ASTSTALAKA FNPFSEPASF SDSDKSHRSK TNKKPSKSDA NRQFNEVHKP RTDQLENSKN
     KSKQLVNKPN HNKMAVKEQR SHKKSHHHR SHQPKQASAS TESHQSSSIE SIFVEEPTLV
     LDREVASINV PANAKAIIAE QGPSTYSKEA LIKDKLKPDP STLVEIEKSL LSLFNMKRPP
     KIDRSKIIIP EPMKKLYAEI MGHELDSVNI PKPGLLTKSA NTVRSFTHKD SKIDDRFPHH
     HRFRLHFDVK SIPADEKLKA AELQLTRDAL SQQVVASRSS ANRTRYQVLV YDITRVGVRG
     QREPSYLLLD TKTVRLNSTD TVSLDVQPAV DRWLASPQRN YGLLVEVRTV RSLKPAPHHH
     VRLRRSADEA HERWQHKQPL LFTYTDDGRH KARSIRDVSG GEGGGKGGRN KRHARRPTRR
     KNHDDTCRRH SLYVDFSDVG WDDWIVAPLG YDAYYCHGKC PFPLADHFNS TNHAVVQTLV
     NNMNPGKVPK ACCVPTQLDS VAMLYLNDQS TVVLKNYQEM TVVGCGCR
//
```

Composite protein sequence databases

- NRDB
- OWL
- MIPSX
- SWISS-PROT+TrEMBL

Amalgamates a number of primary sources, using a set of clearly defined criteria.

Composite protein sequence databases

- NRDB - Non-Redundant DataBase
  - developed and maintained by NCBI
  - composite: GenPept (CDS translations of GenBank), GenPeptupdate, PDB sequences, SWISS-PROT, SWISS-PROTupdate, RIR
  - advantages: comprehesive and up-to date
  - disadvantages: not fully redundant (only identical copies removed), occurence of multiple entries due to polymorphism, incorrect sequences amended in SWISS-PROT re-introduced by translation of GenBank
  - default database of the NCBI BLAST (ENTREZ/NCBI)

Composite protein sequence databases

- OWL
  - developed and maintained by University of Leads
  - composite: SWISS-PROT, PIR1-4, GenBank, NRL-3D
  - SWISS-PROT the highest priority for annotation
  - advantages: less redundant, fully indexed (fast)
  - disadvantages: not up-to-date (released every 6-8 weeks), incorrect sequences
  - available from SEQNET of UK EMBnet

Composite protein sequence databases

- MIPSX
  - developed by Max-Planck Institute in Martinsried
  - composite: PIR1-4, MIPS, NRL-3D, SWISS-PROT, TrEMBL, GenPept, Kabat, PSeqIP
  - identical entries and subsequences removed
- SWISS-PROT+TrEMBL
  - developed and maintained by EBI
  - composite: SWISS-PROT, TrEMBL
  - advantages: comprehensive, minimally redundant, fewer errors
  - disadvantages: not as up-to-date as NRDB
  - available from SRS of EBI

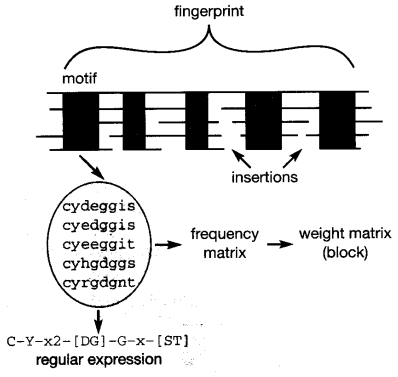
Overview of primary sources of composite databases

NRDB	*OWL*	*MIPSX*	*SP+TrEMBL*
PDB	SWISS-PROT	PIR1–4	SWISS-PROT
SWISS-PROT	PIR	MIPSOwn	TrEMBL
PIR	GenBank	MIPSTrn	
GenPept	NRL-3D	MIPSH	
SWISS-PROTupdate		PIRMOD	
GenPeptupdate		NRL-3D	
		SWISS-PROT	
		EMTrans	
		GBTrans	
		Kabat	
		PseqIP	

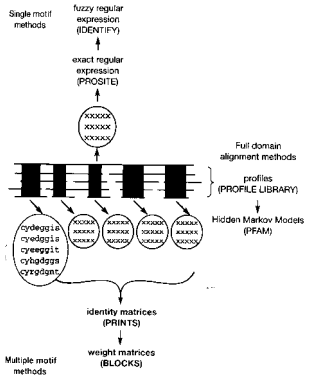
Secondary databases

- Contains information derived from primary sequence data, typically in the form of abstractions: regular expressions, fingerprints, blocks, profiles or Hidden Markov Models.
- These abstractions represent distillations of the most conserved features of multiple alignments.
- The abstractions are useful for discrimination of family membership for newly determined sequences.

Terms used in sequence analysis methods

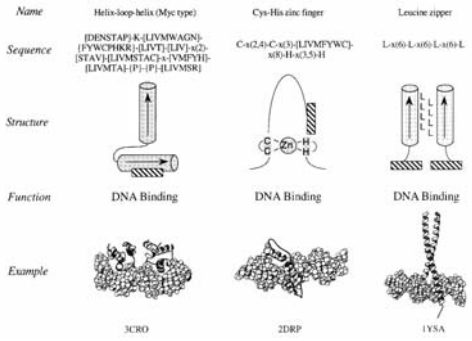


Three principal methods for building secondary databases



Implication of function from a sequence

DNA-binding proteins



Secondary databases

- PROSITE
- PRINTS
- BLOCKS
- Profiles
- Pfam
- IDENTIFY

Secondary databases

- PROSITE
  - historically the first secondary database
  - maintained by Swiss Institute of Bioinformatics
  - motivation: identification of protein families
  - abstraction: regular expressions (patterns)
  - construction: automatic multiple alignment and manual extraction of conserved regions
  - ideally patterns should identify only true-positives (not false-positives)
  - entries deposited as two distinct files: pattern file and documentation files
  - primary source: SWISS-PROT

Pattern file of a entry from the PROSITE database

```
ID  OPSIN; PATTERN.
AC  PS00238;
DT  APR-1990 (CREATED); NOV-1997 (DATA UPDATE); NOV-1997 (INFO UPDATE).
DE  Visual pigments (opsins) retinal binding site.
PA  [LIVMW]-[PGC]-x(3)-[SAC]-K-[STALIM]-[GSACNV]-[STACP]-x(2)-[DENQ]-[AP]-
PA  x(2)-[IY].
NR  /RELEASE=32,49340;
NR  /TOTAL=53(53); /POSITIVE=53(53); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR  /FALSE_NEG=0; /PARTIAL=1;
CC  /TAXO-RANGE=??E??; /MAX-REPEAT=1;
CC  /SITE=5,retinal;
DR  P06002, OPS1_DROME, T; P28678, OPS1_DROPS, T; P22269, OPS1_CALVI, T;
DR  P08699, OPS2_DROME, T; P28679, OPS2_DROPS, T; P04950, OPS3_DROME, T;
DR  P28680, OPS3_DROPS, T; P08255, OPS4_DROME, T; P29404, OPS4_DROPS, T;
DR  P17646, OPS4_DROVI, T; P35362, OPSD_EPHSP, T; P41591, OPSD_ANOCA, T;
DR  P41590, OPSD_ASTFA, T; P02699, OPSD_BOVIN, T; P32308, OPSD_CANFA, T;
DR  P32309, OPSD_CARAU, T; P22328, OPSD_CHICK, T; P28681, OPSD_CRIGR, T;
DR  P08100, OPSD_HUMAN, T; P15409, OPSD_MOUSE, T; P35403, OPSD_PONPI, T;
DR  P02700, OPSD_SHEEP, T; P29403, OPSD_XENLA, T; P22671, OPSD_LAMJA, T;
DR  P31355, OPSD_RANPI, T; P24603, OPSD_LOLPO, T; P09241, OPSD_OCTDO, T;
DR  P35356, OPSD_PROCL, T; P31356, OPSD_TODPA, T; P35360, OPSI_LIMPO, T;
DR  P35361, OPSJ_LIMPO, T; P32310, OPSB_CARAU, T; P28682, OPSB_CHICK, T;
DR  P35357, OPSB_GEKGE, T; P03999, OPSB_HUMAN, T; P28684, OPSV_CHICK, T;
DR  P22330, OPSC_ASTFA, T; P22331, OPSH_ASTFA, T; P32311, OPSG_CARAU, T;
DR  P32312, OPSR_CARAU, T; P28683, OPSG_CHICK, T; P35358, OPSG_GECGE, T;
DR  P04001, OPSG_HUMAN, T; P41592, OPSR_ANOCA, T; P22332, OPSR_ASTFA, T;
DR  P32313, OPSR_CARAU, T; P22329, OPSR_CHICK, T; P04000, OPSR_HUMAN, T;
DR  P34989, OPSL_CALJA, T; P35359, OPSU_BRARE, T; P23820, REIS_TODPA, T;
DR  P47803, RGR_BOVIN , T; P47804, RGR_HUMAN , T;
DR  P17645, OPS3_DROVI, P;
DO  PDOC00211;
```

Documentation file of a entry from the PROSITE database

```
{PDOC00211}
{PS00238; OPSIN}
{BEGIN}
********************************************
* Visual pigments (opsins) retinal binding site *
********************************************

Visual pigments [1,2] are the light-absorbing molecules that mediate vision. They
consist of an apoprotein, opsin, covalently linked to the chromophore cis-retinal.
Vision is effected through the absorption of a photon by cis-retinal which is
isomerized to trans-retinol. This isomerization leads to a change of conformation of
the protein. Opsins are integral membrane proteins with seven transmembrane
regions that belong to family 1 of G-protein coupled receptors (see <PDOC00210>)

In vertebrates four different pigments are generally found. Rod cells, which mediate
vision in dim light, contain the pigment rhodopsin. Cone cells, which function in
bright light, are responsible for color vision and contain three or more color pig-
ments (for example, in mammals: red, blue and green).

In Drosophila, the eye is composed of 800 facets or ommatidia. Each ommatidium
contains eight photoreceptor cells (R1-R8): the R1 to R6 cells are outer cells, R7
and R8 inner cells. Each of the three types of cells (R)-R6, R7 and R8) expresses a
specific opsin.

Proteins evolutionary related to opsins include squid retinochrome, also known as
retinal photoisomerase, which converts various isomers of retinal into 11-cis retinal
and mammalian retinal pigment epithelium (RPE) RGR [3], a protein that may also
act in retinal isomerization.

The attachment site for retinal in the above proteins is a conserved lysine residue in
the middle of the seventh transmembrane helix. The pattern we developed includes
this residue.

-Consensus pattern:  [LIVMW]-[PGC]-x(3)-[SAC]-K-[STALIM]-[GSACNV]-
                     [STACP]-x(2)-[DENQ*]-[AP]-x(2)-[IY]
                     [K is the retinal binding site]
-Sequences known to belong to this class detected by the pattern: ALL.
-Other sequence(s) detected in SWISS-PROT: NONE.
-Last update: November 1997 / Pattern and text revised.

[ 1] Applebury M.L., Hargrave P.A.
     Vision Res. 26:1881-1895(1986).
[ 2] Feyereil K.J., Meyerowitz E.M.
     J. Mol. Evol. 35:567-578(1992).
[ 3] Shen D., Jang M., Hao W., Tao L., Salazar M., Fong H.K.W.
     Biochemistry 33:13147-13125(1994).
{END}
```

Secondary databases

- PRINTS
  - developed at University College London
  - motivation: identification of protein families by more than one pattern
  - abstraction: fingerprints (aligned motifs)
    fingerprints store original sequence information
  - construction: sequence information in a seed motifs are augmented through iterative database scanning
  - construction of fingerprints done manually
  - primary source (original): OWL
  - primary source (new): SWISS-PROT and SP-TrEMBL

Pattern file of a entry from the PRINTS database (I)

```
(a)
OPSIN              OPSIN SIGNATURE
 Type of fingerprint: COMPOUND with 3 elements
Links:
   PRINTS; PR00237 GPCRRHODOPSN; PR00247 GPCRCAMP; PR00248 GPCRMGR
   PRINTS; PR00249 GPCRSECRETIN; PR00250 GPCRSTE2; PR00251 BACTRLOPSIN
   PROSITE; PS00238 OPSIN; PS00237 G_PROTEIN_RECEPTOR
   BLOCKS; BL00238
   SBASE; OPSD_HUMAN
   GCRDB; GCR_0085
   Creation date 20-DEC-1993; UPDATE 2-JUL-1996

   1. APPLEBURY, M.L. and HARGRAVE, P.A.
   Molecular biology of the visual pigments.
   VISION RES. 26 (12) 1881-1895 (1986).

(b)
 SUMMARY INFORMATION
  73 codes involving 3 elements
   1 codes involving 2 elements

 COMPOSITE FINGERPRINT INDEX
   3|  73   73   73
   2|   0    1    1
  --+----------------
    |   1    2    3
```

Pattern file of a entry from the PRINTS database (II)

```
(c)
 INITIAL MOTIF SETS
 OPSIN1 Length of motif = 13 Motif number = 1
 Opsin motif I - 1
                          PCODE        ST    INT
 YVTVQHKKLRTPL            OPSD_BOVIN   60    60
 YVTVQHKKLRTPL            OPSD_HUMAN   60    60
 YVTVQHKKLRTPL            OPSD_SHEEP   60    60
 AATMKFKKLRHPL            OPSG_HUMAN   76    76
 AATMKFKKLRHPL            OPSR_HUMAN   76    76
 YIPATTKSLRTPA            OPS1_DROME   73    73
 VATLKYKKLKQPL            OPSB_HUMAN   57    57
 YIPDGTKSLRTPA            OPS2_DROME   80    80
 WVFSAAKSLRTPS            OPS3_DROME   81    81
 WIFSTSKSLRTPS            OPS4_DROME   77    77
 YLFSKTKSLQTPA            OPSD_OCTDO   58    58
 YLFTKTKSLQTPA            OPSD_LOLFO   57    57

 OPSIN2 Length of motif = 13 Motif number = 2
 Opsin motif II - 1
                          PCODE        ST     INT
 GWSRYIPEGMQCS            OPSD_BOVIN   174    101
 GWSRYIPEGLQCS            OPSD_HUMAN   174    101
 GWSRYIPQGMQCS            OPSD_SHEEP   174    101
 GWSRYWPHGLKTS            OPSG_HUMAN   190    101
 GWSRYWPHGLKTS            OPSR_HUMAN   190    101
 GWSRYVPEGNLTS            OPS1_DROME   187    101
 GWSRFIPEGLQCS            OPSB_HUMAN   171    101
 GWSAYVPEGNLTA            OPS2_DROME   194    101
 TWGRFVPEGYLTS            OPS3_DROME   194    100
 FWDRFVPEGYLTS            OPS4_DROME   190    100
 NWGAYVPEGILTS            OPSD_OCTDO   174    103
 GWGAYTLEGVLCN            OPSD_LOLFO   173    103
```

Secondary databases

- BLOCKS (abstraction: blocks)
- Profiles (abstraction: profiles)
- Pfam (abstraction: Hidden Markov Models)
- IDENTIFY
  - developed at Stanford University
  - abstraction: motifs encoded by fuzzy approach (alternative residues are tolerated in motifs)
  - construction: automatically derived using the program eMOTIF
  - primary sources: PRINTS and BLOCKS

Properties of amino acids used in eMOTIF

Residue property	*Residue groups*
Small	Ala, Gly
Small hydroxyl	Ser, Thr
Basic	Lys, Arg
Aromatic	Phe, Tyr, Trp
Basic	His, Lys, Arg
Small hydrophobic	Val, Leu, Ile
Medium hydrophobic	Val, Leu, Ile, Met
Acidic/amide	Asp, Glu, Asn, Gln
Small/polar	Ala, Gly, Ser, Thr, Pro

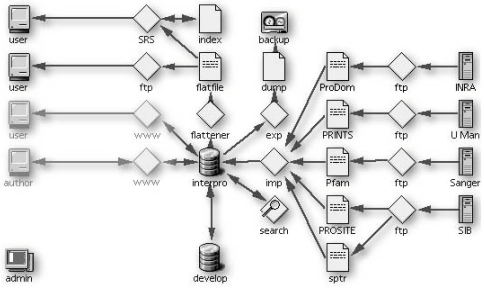
Overview of primary sources and stored information in secondary databases

Secondary database	*Primary source*	*Stored information*
PROSITE	SWISS-PROT	Regular expressions (patterns)
Profiles	SWISS-PROT	Weighted matrices (profiles)
PRINTS	OWL*	Aligned motifs (fingerprints)
Pfam	SWISS-PROT	Hidden Markov Models (HMMs)
BLOCKS	PROSITE/PRINTS	Aligned motifs (blocks)
IDENTIFY	BLOCKS/PRINTS	Fuzzy regular expressions (patterns)

Composite secondary databases

- INTERPRO - Integrated resource of Protein Families, Domains and Sites
  - developed by EBI, SIB, University of Manchester, Sanger Centre, GENE-IT, CNRS/INRA, LION Bioscience AG and University of Bergen (European Research Project)
  - provides an integrated view of the commonly used secondary databases: PROSITE, PRINTS, SMART, Pfam and ProDom
  - accessible by ftp, www and via member databases

InterPro dataflow scheme



Protein structure databases

- PDB
- PDBsum

Protein structure classification databases

- SCOP
- CATCH

Genome information resources

- primary DNA sequence databases
- specialised DNA sequence databases

Primary DNA sequence databases

- EMBL
- DDBJ
- GenBank
- dbEST
- GSDB

Store DNA sequences and annotations.

Primary DNA sequence databases

- EMBL - European Molecular Biology Laboratory
  - European Bioinformatics Institute (EBI)
  - collaboration with DDBJ and GenBank - exchange of new entries on daily basis
  - source of sequences: direct author submissions, genome projects, scientific literature, patents
  - rate of growth is exponential with doubling time ~9-12 months
  - most entries from model organisms
  - retrieval through SRS

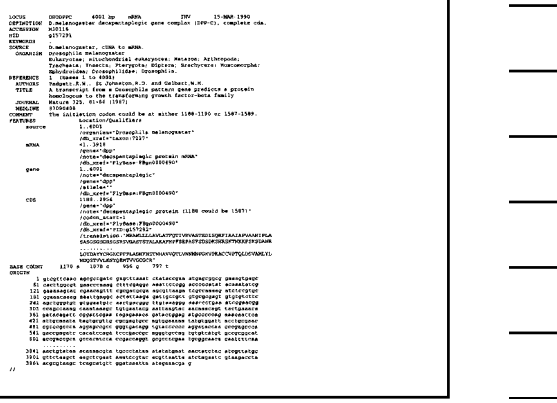
Primary DNA sequence databases

- DDBJ - DNA Data Bank of Japan
  - National Institute of Genetics
  - collaboration with EMBL and GenBank
  - retrieval through DBGet
- GenBank
  - National Center for Biotechnology Information (NCBI)
  - collaboration with DDBJ and EMBL
  - data split into 17 divisions
  - retrieval through Entrez

Codes for 17 divisions of GenBank

Division	*Sequence subset*
PRI	Primate
ROD	Rodent
MAM	Other mammalian
VRT	Other vertebrate
INV	Invertebrate
PLN	Plant, fungal, algal
BCT	Bacterial
RNA	Structural RNA
VRL	Viral
PHG	Bacteriophage
SYN	Synthetic
UNA	Unannotated
EST	EST (Expressed Sequence Tags)
PAT	Patent
STS	STS (Sequence Tagged Sites)
GSS	GSS (Genome Survey Sequences)
HTG	HTG (High Throughput Genomic Sequences)

Entry from the GenBank database



Primary DNA sequence databases

- dbEST
  - National Center for Biotechnology Information (NCBI)
  - maintains only Expressed Sequence Tag (EST) data
- GSDB - Genome Sequence DataBase
  - National Center for Genome Resourses
  - complete collections of DNA sequence for genome-sequencing laboratories
  - on-line submission of large-scale data
  - quality checks
  - format consistent with GenBank + GSDBID

Specialised DNA sequence databases

- SGD
- UniGene
- TDB
- ACeDB

Store species-specific and technique-specific DNA sequences.

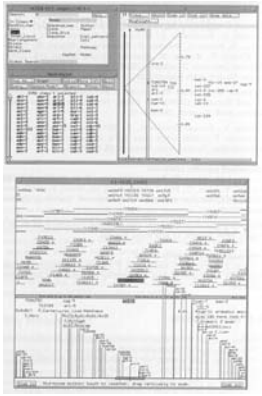
Specialised DNA sequence databases

- SGD - *Saccharomyces* Genome Database
  - molecular biology and genetics of *S. cerevisiae*
  - complete genome, genes, proteins, phenotypes
  - first eukaryotic genome sequenced (1998)
  - sequence analysis, register of genes, 3D structural data, primer sequences for cloning
- UniGene
  - collection of genes encoding proteins (transcript map)
  - non-redundant; derived from GenBank
  - data organised in clusters (1 cluster = 1 unique gene)
  - gene-mapping projects and gene expression analysis

Specialised DNA sequence databases

- TDB - TIGR Database
  - suite of databases: DNA and protein sequences, gene expression, protein families, taxonomic data
  - links: TIGR microbial genome sequencing projects, parasite databases, gene index projects, *A. thaliana* database, human genomic dataset
- ACeDB - A *Cernorhabditis elegans* DataBase
  - *C. elegans* genome project
  - restriction maps, gene structural information, cosmid maps, sequence data, bibliographic information
  - software to organise data ACEDB: CGI script and perl

ACEDB software for organisation of genomic data



DNA sequence analysis

- why to analyse DNA?
- gene structure
- gene sequence analysis
- expression profile, cDNA, EST
- EST sequences analysis

Why to analyse DNA?

- The most sensitive comparisons between sequences are on protein level because of redundancy of the genetic code.
- The loss of degeneracy is accompanied by a loss of information directly linked to the evolution - proteins are only functional abstractions of genetic events at DNA level.
- Silent mutations, important for phylogenetic analysis, can not be detected at protein level.
- Exon/intron analysis, open reading frame [ORF] analysis can not be performed at protein level.

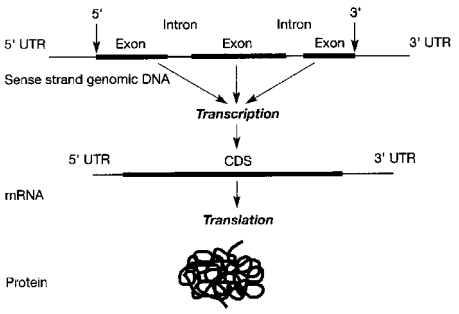
The genetic code

	T		C		A		G		
T	TTT	Phe	TCT	Ser	TAT	Try	TGT	Cys	T
	TTC		TCC		TAC		TGC		C
	TTA	Leu	TCA		TAA	**Stop**	TGA	**Stop**	A
	TTG		TCG		TAG		TGG	Trp	G
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	Gln	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	Lys	AGA	Arg	A
	ATG	**Met**	ACG		AAG		AGG		G
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	Glu	GGA		A
	GTG		GCG		GAG		GGG		G

Gene structure

- Eukaryotic genes are more complex then prokaryotic due to presence of introns.
- DNA databases typically contain genomic data: untranslated sequences, introns+exons, mRNA, cDNA.
- Gene products (proteins) can be of different length, because not all exons can be present in final mRNA.
- The proteins of different length originating from single sequence are called splice variants.

Central dogma of molecular biology



Gene structure

- Untranslated regions (UTRs)
  - portions of the sequence flanking the coding sequence (CDS) not translated into protein
  - UTRs (especially 3' end) is highly gene/species specific
- Exons
  - protein-coding DNA sequences of a gene
- Introns
  - DNA sequences interrupting protein-coding DNA sequence of a gene
  - transcribed into RNA but are edited out during post-transcriptional modifications

Gene sequence analysis

- Conceptual translation - theoretical translation of the DNA sequence to the protein sequence using DNA code without biochemical support.
- Six-frame translation results in six potential protein sequences (ORF analysis).
- ORF analysis
  - codon for methionine - initial codon in the CDS
  - sufficient CDS lenght - long CDS are rare
  - pattern of codon usage - species specific
  - bias towards G/C in the third base of a codon - species specific

Expression profile, cDNA, EST

- Hierarchy of genomic information
  - human genome consists of ~3 billion bp
  - ~3% of the DNA is coding sequence →mRNA→ protein
  - rest of the genome needed for compact structure of chromosomes, replication, control of transcription, etc.
  - 1. chromosomal genome (genome) - genetic information common to every cell in the organism
  - 2. expressed genome (transcriptome) - part of genome expressed in a cell at specific stage in its development
  - 3. proteome - protein molecules that interact to give the cell its individual character

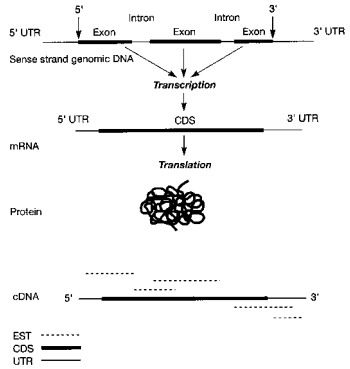
Expression profile, cDNA, EST

- Expression profile
  - characteristic range of genes expressed at particular stage of development and functioning
  - goal of genome projects is to sequence entire (chromosomal) genome
  - having complete sequences and knowing what they mean - two distinct stages of understanding genome
  - alternative approach is analysis of parts of genome expressed in a cell at specific stage in its development
  - comparison of expression profiles: identification of abnormal expressions, expression levels
  - interesting for industry - gene discovery, drug design

Expression profile, cDNA, EST

- Complementary DNA (cDNA)
  - DNA that is synthesised from a messenger RNA template using the enzyme reverse transcriptase
  - cDNA captures expression profile
  - preparation: cultivation/isolation of cells, mRNA extraction, reverse transcription of mRNA to cDNA, transformation of cDNA into library, sequencing of randomly chosen clones (100.000 out of 2 mil.)
  - ideally 100.000 sequences 200-400 bp length - expressed sequence tags (ESTs)
  - in reality many failures, number of sequences lower
  - number of clones constructed and sequenced must be large enough to represent expression profile

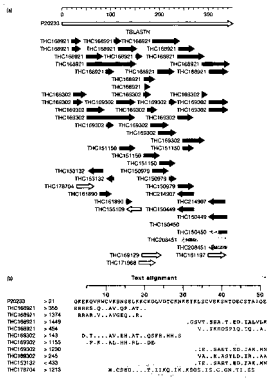
Origin of complementary DNA and expression sequence tags



Expression profile, cDNA, EST

- Libraries of ESTs
  - Merck/IMAGE - 300 000 ESTs from a variety of normalised libraries - higher chance to capture different genes; expression levels not known; sequences deposited to dbEST
  - Incyte - quantitative information on expression levels - standardised libraries; expression profiles in healthy and diseased tissues; sequences form the commercial database LifeSeq
  - TIGR - TIGR Human Gene Index - integrates results from human gene projects [dbEST+GenBank] - purpose is to identify all possible human genes by sequence assembly - creates Tentative Human Consensus (THC) sequences and contigs

Search result from TIGR Human Genom Index database



EST sequences analysis

EST production is highly automated (fluorescent laser systems and computer analysis of chromatograms) influencing the quality of sequences. Specific character of ESTs must be respected during their analysis:

- EST alphabet
- Insertions, deletions, frameshifts
- Splice variants in EST
- Non-coding regions

EST sequences analysis

- EST alphabet
  - automated computer analysis of chromatograms
  - program is sometimes unable to decide base for particular position and inserts ambiguous base N
  - should be <5% of total length
- Insertions, deletions and frameshifts
  - automated base-calling software assumes regular intervals among peaks - not always the case
  - phantom INDELs (insertions and deletions)
  - identification of INDELs by sequence comparisons

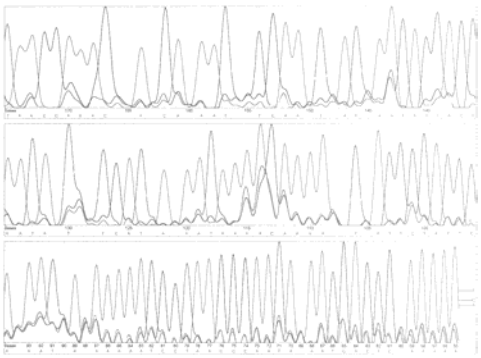
List of base-ambiguity symbols defined by IUB-IUPAC

IUB symbol	*Represented bases*
A	A
C	C
G	G
T/U	T
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
V	A or C or G
H	A or C or T
D	A or G or T
B	C or G or T
X/N	G or A or T or C

EST sequences analysis

- Splice variants
  - splice variants are represented by deletions arising from non-inclusion of exons
  - in EST maybe missing bases due to sequencing errors
  - partially good match = splice form or sequence error?
- Non-coding regions
  - question: does this EST represent a new gene?
  - search of DNA database for similar non-coding regions
  - no hit found = the EST represents a new gene (CDS) or the EST represents non-coding sequence not present in the database

Sequencing chromatogram



EST sequences analysis

Three categories of EST analysis tools:

- Sequence similarity search tools
- Sequence assembly tools
- Sequence clustering tools

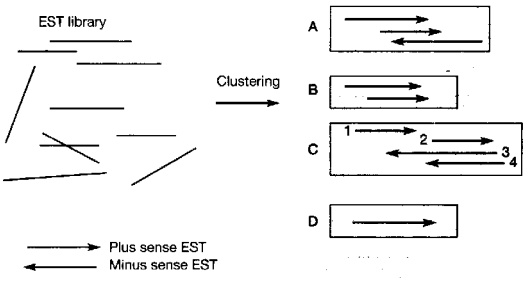
EST sequences analysis

- Sequence similarity search tools
  - current database search programs are designed to cope with EST: TBLASTN (translate DNA databases), BLASTX (translate input sequence), TBLASTX (translate both)
- Sequence assembly tools
  - search of the databases reveals several ESTs matching the query sequence
  - alignment of hits and construction of consensus
  - search with consensus, aligment, ….
  - iterative sequence alignment = sequence assembly

EST sequences analysis

- Sequence clustering tools
  - clustering of EST sequences reduces redundancy and saves the search time
  - enables estimation of genes in the EST database
  - approach 1: clustering based on sequences from comprehensive DNA database
  - approach 2: clustering of all ESTs, construction of consensus sequences representing each cluster, DNA database search using consensus sequences only
  - result = ESTs that do not match any of the database sequences

Clustering of EST library



Pairwise sequence alignment

- database searching
- alphabets and complexity
- algorithms and programs
- sequences and sub-sequences
- identity and similarity
- dotplot
- local and global similarity
- pairwise database searching

Database searching

- Database search can take a form of text queries or sequence similarity searches.
- Text queries are problematic due to missing annotations in many sequences.
- query sequence = probe
  searched sequence = subject
- The purpose of searches is to identify evolutionary relationships (homology) from sequence similarity. Important for search of analogous family members in different species.

Alphabets and complexity

- A sequence consists of letters from an alphabet.
- The complexity of the alphabet is defined by the number of letters it contains:
  - DNA = 4
  - EST = 5
  - proteins = 20
- Special letters can be used for ambiguous bases (N) or residues (X). Sequence searching programs must be able to deal with them.

Algorithms and programs

- Algorithm is a set of steps that define a certain computational process.
- Program is a the implementation of the algorithm.
- Same algorithm may be implemented in many programs.

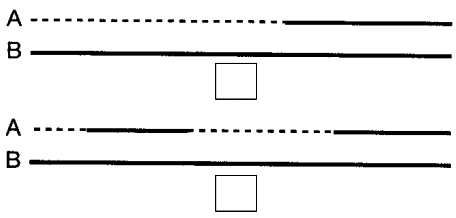
Sequences and sub-sequences

- Alignment of two short sequences:

```
Unaligned                              score = 6
Sequence 1 (query)         AGGVLIIQVG
                           ||||||
Sequence 2 (subject)       AGGVLIQVG

Aligned                                score = 9
Sequence 1 (query)         AGGVLIIQVG
                           |||||| |||
Sequence 2 (subject)       AGGVLI-QVG
```

- Score increases by the insertion of a gap. The gap increases the number of aligned identical residues.

Alignment of a sub-sequence with full sequence



Identity and similarity

- Introduction of gaps solely to maximise identities is not biologically meaningful.
- Scoring penalties are introduced to minimise opening and extension of gaps.
- Unitary matrix (counting identities) is replaced by similarity matrix (counting similarities) = high-scoring matches are replaced by biologically meaningful low-scoring matches.
- Diagnostic power of similarity matrices is higher.

Unitary scoring matrices: (a) DNA and (b) protein

(a)

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

(b)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	B	Z	X
C	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Identity and similarity

- Dayhoff Mutation Data Matrix
  - score is based on the concept of Point Accepted Mutation (PAM)
  - evolutionary distance 1 PAM = probability of a residue mutating during a distance in which 1 point mutation is accepted per 100 residues
  - 250 PAM matrix - similarity score equivalent to 20% matches remaining between two sequences = suitable for identification of similarities in twilight zone
  - limitation: derived from alignment of sequences >85% identical

Mutation Data Matrix for 250 PAMs

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	−2	1	3																	
P	−3	1	0	6																
A	−2	1	1	1	2															
G	−3	1	0	−1	1	5														
N	−4	1	0	−1	0	0	2													
D	−5	0	0	−1	0	1	2	4												
E	−5	0	0	−1	0	0	1	3	4											
Q	−5	−1	−1	0	0	−1	1	2	2	4										
H	−3	−1	−1	0	−1	−2	2	1	1	3	6									
R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6								
K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5							
M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6						
I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5					
L	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6				
V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4			
F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9		
Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	−2	7	10	
W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	−3	−4	−5	−2	−6	0	0	17

Identity and similarity

- BLOSUM matrices
  - BLOcks SUbstitution Matrix
  - derived from blocks of aligned sequences in BLOCKS database - represents distant relationships implicitly
  - bias from identical sequences is removed by clustering
  - BLOSUM62 = matrix derived from sequences clustered at 62% or greater identity

Identity and similarity

- Statistical measures of alignment significance
  - performing sequence alignment computationally = creating match according to mathematical model
  - adjustable parameters: gap penalties, impact of sequence length, effect of alphabet complexity
  - level of confidence to constructed alignment is quantified by statistical parameters:

    probability (p) - probability that the constructed alignment arose by chance [should approach 0]

    expected frequency (E) - number of hits one can expect to see by chance [should be <0.001]

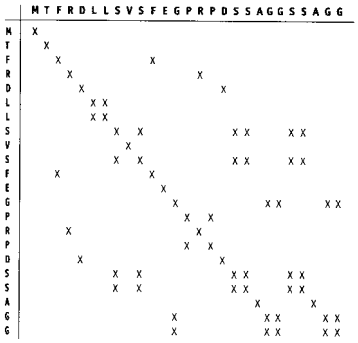
Example hit list from a database search

Sequences producing significant alignments:	Score (bits)	E Value
sp|P51698|LINB_PSEPA (LINB)1,3,4,6-tetrachloro-1,4-cyclohexadie...	616	e-176
sp|Q50642|YP79_MYCTU (RV2579..)Hypothetical 33.7 kDa protein Rv...	450	e-126
sp|P27652|LUCI_RENRE Renilla-luciferin 2-monooxygenase (EC 1.13...	218	2e-56
sp|Q50600|YJ33_MYCTU (RV1833C..)Hypothetical 32.2 kDa protein R...	102	8e-22
sp|Q50670|YM96_MYCTU (RV2296..)Putative haloalkane dehalogenase...	93	7e-19
sp|P22643|HALO_XANAU (DHLA)Haloalkane dehalogenase (EC 3.8.1.5)...	87	5e-17
sp|P34913|HYES_HUMAN (EPHX2)Soluble epoxide hydrolase (SEH) (EC...	49	2e-05
sp|O07214|YR15_MYCTU (RV2715..)Hypothetical 36.9 kDa protein Rv...	47	5e-05
sp|Q50599|YI34_MYCTU (RV1834..)Hypothetical 31.7 kDa protein Rv...	45	2e-04
sp|O31158|PRXC_PSEFL (CPO..)Non-heme chloroperoxidase (EC 1.11....	45	2e-04
sp|P22862|ESTE_PSEFL Arylesterase (EC 3.1.1.2) (Aryl-ester hydr...	44	6e-04
sp|P23106|XYLF_PSEPU (XYLF)2-hydroxymuconic semialdehyde hydrol...	40	0.008
sp|P29715|BPA2_STRAU (BPOA2)Non-haem bromoperoxidase BPO-A2 (EC...	39	0.011
sp|P49323|PRXC_STRLI (CPO..)Non-heme chloroperoxidase (EC 1.11....	37	0.054
sp|P54549|YQJL_BACSU (YQJL)Hypothetical 28.2 kDa protein in GLN...	36	0.093
sp|P48972|MYBB_MOUSE (MYBL2..)Myb-related protein B (B-Myb).[Mu...	36	0.12
sp|Q55921|PRXC_SYNY3 (SLR0314)Putative non-heme chloroperoxidas...	36	0.16
sp|Q9JZR6|PIP_NEIMB (PIP..)Proline iminopeptidase (EC 3.4.11.5)...	36	0.16
sp|O13912|YDW6_SCHPO (SPAC23C11.06C)Hypothetical 60.1 kDa prote...	34	0.47
sp|Q59695|ACOC_PSEPU (ACOC)Dihydrolipoamide acetyltransferase c...	34	0.62
sp|P46544|PIP_LACDE (PEPIP)Proline iminopeptidase (EC 3.4.11.5)...	33	1.1
sp|P46542|PIP_LACDL (PIP..)Proline iminopeptidase (EC 3.4.11.5)...	32	1.4
sp|P10244|MYBB_HUMAN (MYBL2..)Myb-related protein B (B-Myb).[Ho...	30	9.2
sp|Q15811|ITSN_HUMAN (ITSN..)Intersectin (SH3 domain-containing...	30	9.2

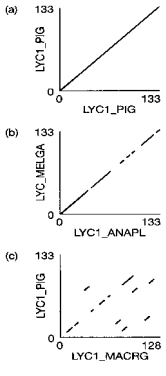
Dotplot

- The most basic visual method for comparison of two sequences.
- Separates noise (random dots) from the signal (adjacent dots).
- Identical sequences are represented by single central diagonal line, similar sequences by a broken diagonal and dissimilar sequences by random dots.
- Advanced dotplots utilise similarity matrices for calculation of cell scores.

Construction of the dotplot matrix



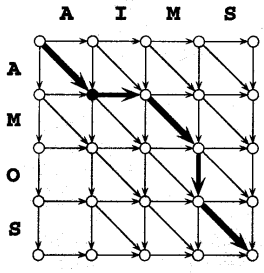
Dotplot of (a) identical, (b) similar and (c) related sequences



Local and global similarity

- Alignments are mathematical models whose behaviour can be modified through the use of adjustable parameters. The models constructed by dynamic programming algorithms - finding solution of a problem by solving smaller, but similar sub-problems.
- Global alignment - considers similarity across the entire sequence.
- Local alignment - considers similarity in parts of sequences only.

Path matrix with optimal path by dynamic programming



Alignment **AIM-S**
A-MOS

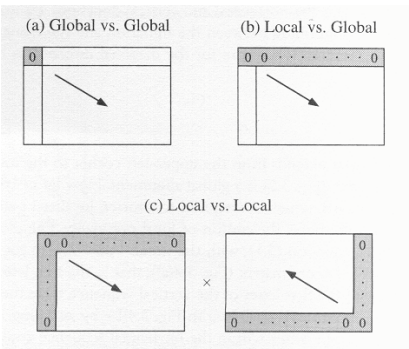
Local and global similarity

- Global alignment
  - Needleman and Wunsch algorithm
  - suitable for sequences similar across most of their length (usually closely related)
  - 1. construction of 2D similarity matrix ("dotplot")
  - 2. successive summation of the cells in the matrix starting from N-terminal end → progressing through the sequence
  - 3. construction of maximum-match path through the entire sequence

Local and global similarity

- Local alignment
  - Smith-Waterman algorithm
  - suitable for distantly related sequences displaying local regions of similarity (functionally-relevant or structurally-relevant)
  - each point of the matrix defines the end point of a potential alignment = edge cells of the matrix are initialised to 0
  - possibility for ending the alignment are calculated for every cell
  - algorithm is much faster compared to global similarity algorithms

Concepts of global and local optimality in the pairwise sequence alignment



Pairwise database searching

- Extension of the pairwise sequence alignments.
- Large database searches can not be performed using the original Needleman and Wunsch or Smith-Waterman algorithms due to time limitations.
- Very fast local-similarity search methods employing heuristics = FastA and BLAST. These methods concentrates on finding short identical matches.

Pairwise database searching

- FastA
  - algorithm by Lipman and Pearson (1985)
  - identifies short words (k-tuples) common to both sequences
  - k-tuples for proteins: 1-2 residues
  - k-tuples for DNA: up to 6 bases
  - k-tuples lying close to each other on the same diagonal joined by heuristics → gapped alignments computed by dynamic programming

Output from FastA search

```
FASTA searches a protein or DNA sequence data bank
 version 3.3t09 May 18, 2001
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

8:1-: 296 aa
 EMBOSS_001
 vs  SWISS-PROT Protein Sequence Database library
searching /ebi/services/idata/fastadb/swissprot library

37135523 residues in 101247 sequences
 statistics extrapolated from 60000 to 101082 sequences
  Expectation_n fit: rho(ln(x))= 5.8158+/-0.000184; mu= 4.0375+/- 0.010
 mean_var=74.4386+/-14.720, 0's: 132 Z-trim: 20  B-trim: 0 in 0/65
 Lambda= 0.1487

FASTA (3.39 May 2001) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 36, opt: 24, gap-pen: -12/ -2, width:  16
 Scan time:  1.930
The best scores are:                                      opt bits E(101082)
SW:LINB_PSEPA P51698 1,3,4,6-TETRACHLORO-1,4-CYCL  ( 296) 2041  447 2.4e-125
SW:YF79_MYCTU Q50642 HYPOTHETICAL 33.7 KDA PROTEI  ( 300) 1494  330 5.1e-90
SW:LUCI_RENRE P27652 RENILLA-LUCIFERIN 2-MONOOXYG  ( 311)  744  169 1.4e-41
SW:DMPD_PSESP P19076 2-HYDROXYMUCONIC SEMIALDEHYD  ( 283)  169   46 0.00017
SW:PRXC_PSEFL O31158 NON-HEME CHLOROPEROXIDASE (E  ( 273)  168   45 0.00019
SW:PRXC_STRLI P49323 NON-HEME CHLOROPEROXIDASE (E  ( 275)  140   39   0.012
SW:PIP_BACCO P46541 PROLINE IMINOPEPTIDASE (EC 3.  ( 288)  140   39   0.013
SW:PRXC_SYNY3 Q55921 PUTATIVE NON-HEME CHLOROPERO  ( 276)  125   36    0.11
SW:PIP_NEIGO P42786 PROLINE IMINOPEPTIDASE (EC 3.  ( 310)  122   35     0.2


>>SW:LINB_PSEPA P51698 1,3,4,6-TETRACHLORO-1,4-CYCLOHEXA  (296 aa)
 initn: 2041 init1: 2041 opt: 2041  Z-score: 2372.6  bits: 447.0 E(): 2.4e-125
Smith-Waterman score: 2041;  100.000% identity (100.000% ungapped) in 296 aa overlap
(1-296:1-296)

               10        20        30        40        50        60
EMBOSS MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SW:LIN MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIA
               10        20        30        40        50        60
```

Pairwise database searching

- BLAST
  - Basic Local Alignment Search Tool
  - algorithm by Altschul *et al.* (1990)
  - identifies short ungapped sub-sequences (segment pairs) of the same length
  - sub-sequences are extended using dynamic programming to obtain local alignments - high scoring pairs (HSPs)
  - improved algorithm by Altschul *et al.* (1997) - produces gapped alignments
  - algorithm very fast - most commonly used for databases searching

Output from BLAST search

```
BLASTP 2.0.14 [Jun-29-2000]


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

Query= /net/nfs0/vol1/production/w3nobody/tmp/918495.5350-
80758.blastall.a  [Unknown form], 297 bases, 818F03BD checksum.
         (296 letters)

Database: swissprot
           101,247 sequences; 37,135,523 total letters

Searching.....................................................done

                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

SW:LINB_PSEPA P51698 1,3,4,6-TETRACHLORO-1,4-CYCLOHEXADIENE ...   616  e-176
SW:YF79_MYCTU Q50642 HYPOTHETICAL 33.7 KDA PROTEIN RV2579.        450  e-126
SW:LUCI_RENRE P27652 RENILLA-LUCIFERIN 2-MONOOXYGENASE (EC 1...   218  2e-56
SW:DMPD_PSESP P19076 2-HYDROXYMUCONIC SEMIALDEHYDE HYDROLASE...    50  9e-06
SW:PRXC_PSEFL O31158 NON-HEME CHLOROPEROXIDASE (EC 1.11.1.10...    45  2e-04
SW:BPA2_STRAU P29715 NON-HAEM BROMOPEROXIDASE BPO-A2 (EC 1.1...    39  0.011
SW:PIP_BACCO P46541 PROLINE IMINOPEPTIDASE (EC 3.4.11.5) (PI...    39  0.014
SW:PIP_NEIMB Q9JZR6 PROLINE IMINOPEPTIDASE (EC 3.4.11.5) (PI...    36  0.16

>SW:LINB_PSEPA P51698 1,3,4,6-TETRACHLORO-1,4-CYCLOHEXADIENE HYDROLASE (EC
          3.8.1.-) (1,4- TCDN CHLOROHYDROLASE).
         Length = 296

 Score =  616 bits (1572), Expect = e-176
 Identities = 296/296 (100%), Positives = 296/296 (100%)

Query: 1   MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIA 60
           MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIA
Sbjct: 1   MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIA 60
```

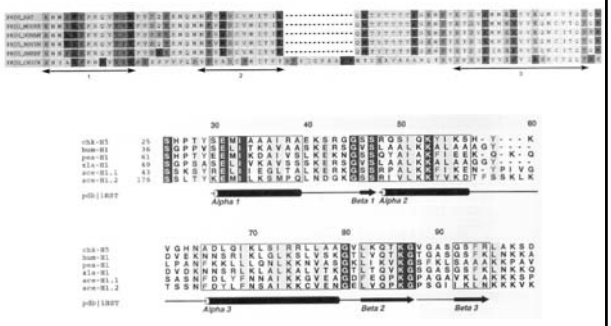
Multiple sequence alignment

- multiple sequence alignment
- consensus sequence
- manual methods
- simultaneous and progressive methods
- databases of multiple sequence alignments
- hybrid approach for database searching

Multiple sequence alignment

- Multiple sequence alignment is a 2D table in which the rows represent individual sequences and the columns the residue positions.
- Multiple sequence alignments are essential for analysis of sets of gene families.
- Sequence-based multiple sequence alignments - constructed according to similar strings of amino acid residues.
- Structure-based multiple sequence alignments - constructed according to structural evidence.

Colour-coded multiple sequence alignments



Multiple sequence alignment

- Construction of a multiple sequence alignment:
  - positioning of residues within any sequence is preserved (absolute positions)
  - similar residues in all sequences are brought into vertical register (relative positions)
- All residues in any single column of an alignment will have the same relative position but different absolute position (unless the sequences are identical).

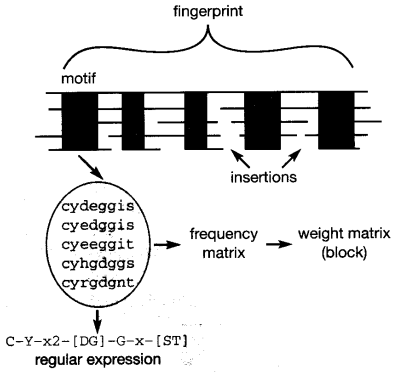
Consensus sequence

- The alignment table can be summarised by:
  - a single line: pseudo-sequence
  - unweighted matrix: fingerprint
  - ungapped block of residues (weighted): block
  - weighted matrix: profile

Multiple alignment and the consensus sequence

	1	2	3	4	5	6	7	8	9	10
I	Y	D	G	G	A	V	–	E	A	L
II	Y	D	G	G	–	–	–	E	A	L
III	F	E	G	G	I	L	V	E	A	L
IV	F	D	–	G	I	L	V	Q	A	V
V	Y	E	G	G	A	V	V	Q	A	L
	y	d	G	G	A/I	V/L	V	e	A	l

Multiple alignment and the profile, block and fingerprint



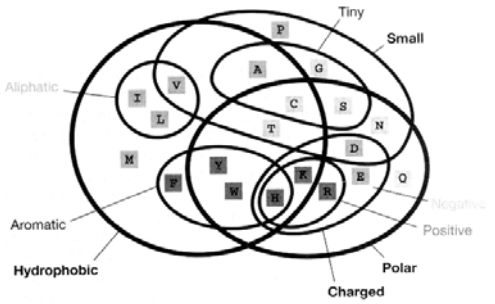
Manual methods

- Manual methods are subjective however they enable to incorporate experimental evidences (e.g., mutagenesis data, structural knowledge) into the multiple alignment.
- Manual modification of the multiple alignments from automatic methods is the best approach.
- Intuitive colouring schemes assist the eye in spotting similarities.
- Quantitative evaluation of relatedness through calculation of residue identities/similarities.

Amino acid property groupings and colouring

Residue	*Property*	*Colour*
Asp, Glu	Acidic	red
His, Arg, Lys	Basic	blue
Ser, Thr, Asn, Gln	Polar neutral	green
Ala, Val, Leu, Ile, Met	Hydrophobic aliphatic	white
Phe, Try, Trp	Hydrophobic aromatic	purple
Pro, Gly	Special structural properties	brown
Cys	Disulphide bond former	yellow

Venn diagram grouping properties of the amino acids



Simultaneous methods

- Simultaneous methods align all sequences in a given set at once, rather than aligning pairs of sequences or building sequence clusters.
- Extension of 2D dynamic programming matrix to more dimensions.
- Number of dimensions = number of sequences.
- Suitable only for small sets of short sequences.

Progressive methods

- Multi-dimensional programming matrix is not applicable to realistic problems - larger sets of longer sequences.
- CLUSTAL
  - 1. construction of evolutionary tree
  - 2. pairwise alignment of two the most closely related sequences, addition of less related sequences
  - 3. final alignment, final evolutionary tree
- CLUSTALW
  - positioning of gaps in closely related sequences according to their variability

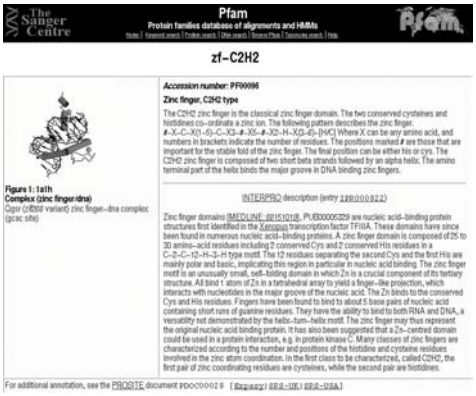
Databases of multiple alignments

- Multiple alignments bring together sequences from different species. This important evolutionary information can enhance sensitivity of database searches.
- Various abstractions (regular expressions, profiles, blocks, fingerprints or HMMs) can be searched against sequence databases. More information used in a query - higher sensitivity.
- Results of the searches using the multiple alignments are more difficult to interpret.

Databases of multiple alignments

- Multiple alignments databases available via Web are produced automatically (e.g., PFAM) or manually (e.g., PRINTS).
- Iterative automatic methods may include false-positive sequences in the alignment which will corrupt it by insertion of many unrealistic gaps.

Example entry from PFAM database

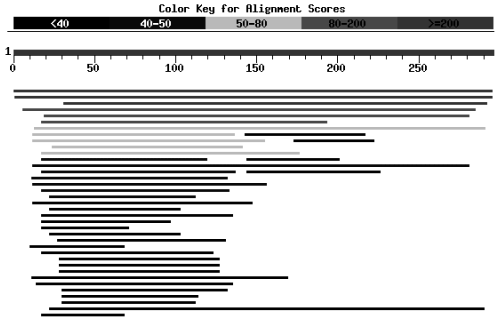


Example alignment from PFAM database

Hybrid approach for database searching

- PSI-BLAST
  - Position-Specific Iterated - BLAST
  - algorithm by Altschul *et al.* (1997)
  - incorporates elements of both pairwise and multiple sequence alignment methods
  - procedure: initial search - creation of position specific profiles from the hits - new search … in iterations
  - advantage: detects even very weak similarities
  - disadvantages: the profile can be diluted if low-complexity regions are not masked; inclusion of single false-positive sequence into the profile leads to bias towards unrelated sequences

Graphic hit list from a database search using PSI-BLAST



Secondary database searching

- why to search secondary databases?
- secondary databases
- regular expressions
- fingerprints
- blocks
- profiles
- Hidden Markov Models

Why to search secondary databases?

- Interpretation of the results from primary database searches is sometimes difficult:
  - X.000.000 sequences from XX.000 organisms
  - complex and redundant search outputs
  - irrelevant matches of low-complexity sequences, repetitive sequences, modular sequences
  - local regions of similarity in multi-domain proteins
  - truncated description lines
- Secondary database searches enable to identify both homology and more exacting orthology.

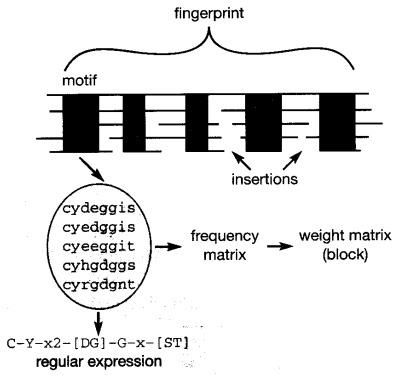
Secondary databases

- Contains information derived from primary sequence data, typically in the form of abstractions: regular expressions, fingerprints, blocks, profiles or Hidden Markov Models.
- These abstractions represent distillations of the most conserved features of multiple alignments.
- The abstractions are useful for discrimination of family membership for newly determined sequences.

Secondary databases

- PROSITE - regular expressions
- PRINTS - fingerprints
- BLOCKS - blocks
- Profiles - profiles
- Pfam - Hidden Markov Models
- IDENTIFY - fuzzy regular expressions

Terms used in sequence analysis methods



Regular expressions

- Regular expression reduces the sequence data to the most conserved residue information.

Multiple alignment	Regular expression
ADLGAVFALCDRYFQ	[AS]-D-[IVL]-G-X5-C-[DE]-R-[FY]2-Q
SDVGPRSCFCERFYQ	
ADLGRTQNRCDRYYQ	
ADIGQPHSLCERYFQ	

- Limitations:
  - stringent pattern - retrieves only identical matches and can miss remote relatives
  - fuzzier pattern - better chance to detect remote relatives, but results in more noisy output
  - single motif may not be sufficient to infer the function

Regular expressions

- Regular expressions works most effectively when a particular protein family can be characterised by a highly conserved motif (10-20 residues).
- Limitation: short patterns (3-4 residues) are not sufficiently discriminative.

Asp-Ala-Val-Ile-Asp (DAVID)	71 exact matches in OWL29.6
Asp-Ala-Val-Glu (DAVE)	1088 exact matches in OWL29.6

Regular expressions

- Rules - short patterns that can be used to provide a guide to possible existence of functional sites:

Functional site	Regular expression
N-glycosilation site	N-{P}-[ST]-{P}
Protein kinase C phosphorylation site	[ST]-X-[RK]
Casein kinase II phosphorylation site	[ST]-X(2)-[DE]
Asp adn Asn hydroxylation site	C-X-[DN]-X(4)-[FY]-X-C

Regular expressions

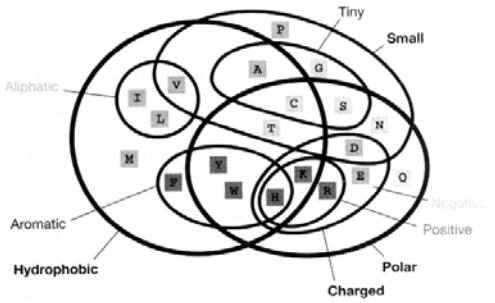
- Fuzzy regular expressions - regular expressions with introduced fuzziness into patterns using groups of amino acids with similar biochemical properties (FYW - aromatic, HKR - basic, etc.).

Multiple alignment	Fuzzy regular expression
ADLGAVFALCDRYFQ	[ASGPT]-D-[IVLM]-G-X5-C-[DENQ]-R-[FYW]2-Q
SDVGPRSCFCERFYQ	
ADLGRTQNRCDRYYQ	
ADIGQPHSLCERYFQ	

Amino acid property groupings and colouring

Residue	*Property*	*Colour*
Asp, Glu	Acidic	red
His, Arg, Lys	Basic	blue
Ser, Thr, Asn, Gln	Polar neutral	green
Ala, Val, Leu, Ile, Met	Hydrophobic aliphatic	white
Phe, Try, Trp	Hydrophobic aromatic	purple
Pro, Gly	Special structural properties	brown
Cys	Disulphide bond former	yellow

Venn diagram grouping properties of the amino acids



Regular expressions

- Introduction fuzziness into regular expressions increases the number of matches retrieved from the sequence database:

Regular expression	No. of exact matches (OWL29.6)
D-A-V-I-D	71
D-A-V-I-[DENQ]	252
[DENQ]-A-V-I-[DENQ]	925
[DENQ]-A-[VLI]-I-[DENQ]	2739
[DENQ]-[AQ]-[VLI]2-[DENQ]	51506

Fingerprints

- Motivation: there are often more than one conserved region present in multiple alignment.
- Groups of motifs excised from the sequence and converted into matrices populated by the residue frequencies observed at each position.
- Unweighted scoring system - no additional mutation or substitution matrices are employed.
- Weighted scoring system - additional matrices are employed resulting in less sparse matrix, but poor signal-to-noise performance.

Example of (a) ungapped aligned motif and (b) its corresponding frequency matrix

(a)

YVTVQHKKLRTPL
YVTVQHKKLRTPL
YVTVQHKKLRTPL
AATMKFKKLRHPL
AATMKFKKLSHPL
YIFATTKSLRTPA
VATLRYKKLRQPL
YIPQTTKSLRTPA
WFSAAKSLRTPS
WIFSTSKSLRTPS
YLFSKTKSLQTPA
YLFTKTKSLQTPA

(b)

T	C	A	G	N	S	P	F	L	Y	H	Q	V	K	D	E	I	W	R	M	B	X	Z
0	0	2	0	0	0	0	0	0	7	0	0	1	0	0	0	0	2	0	0	0	0	0
0	0	3	0	0	0	0	0	2	0	0	0	4	0	0	0	3	0	0	0	0	0	0
0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	3	0	0	1	0	0	0	3	0	0	0	0	0	0	2	0	0	0
0	0	1	1	0	0	0	0	0	0	0	3	0	4	0	0	0	0	1	0	0	0	0
0	0	1	0	0	1	0	2	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	6	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	10	0	0	0	0
0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	4	0	0	2	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Example of frequency matrix derived from (a) initial unweighted motif and (b) PAM-weighted matrix

(a)

T	C	A	G	N	S	P	F	L	Y	H	Q	V	K	D	E	I	W	R	M	B	X	Z
0	0	4	0	0	0	0	8	4	34	0	0	15	0	0	0	1	7	0	0	0	0	0
0	4	15	0	0	0	0	0	7	0	0	0	37	0	0	0	10	0	0	0	0	0	0
50	0	0	0	0	3	0	18	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
3	0	12	2	1	8	0	3	6	0	0	0	14	0	0	0	15	2	0	7	0	0	0
9	2	2	2	1	1	0	0	0	0	1	25	0	20	0	6	0	0	4	0	0	0	0
14	0	2	0	0	4	0	14	0	8	31	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	70	0	0	0	0	2	0	0	0	0
0	0	2	1	0	17	0	0	0	0	0	0	0	52	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	73	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	68	0	0	0	0
44	0	0	0	0	6	0	0	0	0	12	11	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	69	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
2	0	11	0	0	7	0	0	53	0	0	0	0	0	0	0	0	0	0	0	0	0	0

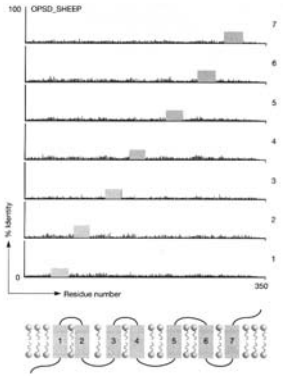
(b)

T	C	A	G	N	S	P	F	L	Y	H	Q	V	K	D	E	I	W	R	M
-29	-22	-29	-48	-24	-24	-46	40	-13	82	-10	-40	-22	-35	-44	-44	-15	15	-30	-22
-1	-32	-1	-18	-20	-10	-13	-9	20	-22	-21	-18	32	-23	-22	-20	32	-51	-26	19
0	-36	-18	-30	-24	-12	-30	36	0	24	-18	-36	-6	-30	-36	-30	6	-30	-30	-6
3	-29	3	-4	-30	-1	-7	-22	3	-33	-19	-15	14	-12	-15	-13	11	-52	-15	11
3	-48	-1	-8	7	1	-4	-54	-31	-46	6	14	-17	23	6	5	-20	-48	14	-9
2	-27	-7	-19	-3	-5	-13	0	-16	6	5	-10	-11	-15	-13	-11	-7	-37	-12	-15
0	-60	-12	-24	12	0	-12	-60	-36	-48	0	12	-24	60	0	0	-24	-36	36	0
6	-30	0	-6	12	12	0	-48	-36	-42	-6	0	-18	30	0	0	-18	-30	18	-12
-24	-72	-24	-48	-36	-36	-36	**24**	**72**	-12	-24	-24	**24**	-36	-48	-36	**24**	-24	-36	**48**
-12	-50	-20	-32	2	-2	0	-50	-34	-48	26	18	-24	32	-6	-6	-24	10	62	-2
24	-29	7	-5	5	6	0	-36	-24	-31	5	1	-5	1	4	4	-6	-56	-4	-14
0	-36	12	-12	-12	12	72	-60	-36	-60	0	0	-12	-12	-12	-12	-24	-72	0	-24
-6	-44	-2	-18	-16	-10	-12	-10	22	-24	-18	-14	10	-22	-24	-18	6	-40	-26	16

Blocks

- Conserved motifs are located by a first motif-finding algorithm: search for the spaced residue triplets (e.g., Ala-X-X-Val-X-Trp); a block score is weighted using BLOSUM 62 substitution matrix.
- Validation of blocks by a second motif-finding algorithm: search for the highest-scoring set of blocks in the correct order without overlapping.
- Sequences are clustered to avoid a bias due to identical sequences.

Visualisation of protein fingerprints and blocks



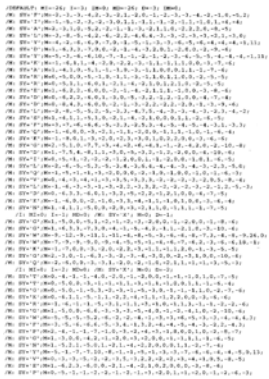
Block with clustered sequences and weighted scores

```
CCKR_HUMAN ( 362) SSCVNPIIYCFMHKKFK 3
CCKR_RAT ( 378) SSCVNPIIYCFMHKKFK 3

FML2_HUMAN ( 294) NSCLNFMLVVFVGQDFK 4
FMLR_HUMAN ( 293) NSCLNFMLVVFMGQDFK 4
FMLR_MOUSE ( 304) NSCLNFMLVVFMGQDFK 4
FMLR_RABIT ( 295) NSCLNFMLVVFMGQDFK 4

GASR_CANFA ( 388) SACVNPLVYCFMSKKFR 5
GASR_HUMAN ( 382) SACVNPLVYCFMHKKFR 5
GASR_PANHA ( 385) SACVNPLVYCFMHKKFR 5
GASR_RABIT ( 387) SACVNPLVYCFMHKKFR 5
  GASR_RAT ( 387) SACVNPLVYCFMHKKFR 5

ETIR_BOVIN ( 361) NSCINPIALVFVSKKFK 9
  ETIR_RAT ( 361) NSCINPIALVFVSKKFK 9
ETHR_BOVIN ( 377) NSCINPIALVLVSKKFK 9
ETHR_HUMAN ( 378) NSCINPIALYLVSKKFK 9
  ETHR_PIG ( 379) NSCINPIALYLVSKKFK 9
  ETHR_RAT ( 378) NSCINPIALVLVSKKFK 9

OPSD_LOLFO ( 307) SAIMNPMIYSVSNPKFR 12
OPSD_OCTDO ( 308) SAIMNPIVYSVSNPKFR 12
OPSD_TODPA ( 306) SAIMNPMIYSVSNPKFR 12

P2UR_HUMAN ( 296) NSCLDPVLYFLAGQRLV 13
P2UR_MOUSE ( 296) NSCLDPVLYFLAGQRLV 13
  P2UR_RAT ( 297) NSCLDPVLYFLAGQRLV 13

   5HE_RAT ( 312) NSTMNPIIYFLFMSDFK 16

EDG1_HUMAN ( 302) NSGTNPIIYTLTNKEMR 21

EBI2_HUMAN ( 300) NCCMDFFIYFFACKGYK 23

OXYR_HUMAN ( 321) NSCCNPWIYMLFTGHLF 24
  OXYR_PIG ( 323) NSCCNPWIYMLFTGHLF 24
V1AR_HUMAN ( 340) NSCCNPWIYMFFSGHLL 18
  V1AR_RAT ( 346) NSCCNPWIYMFFSGHLL 18

PER3_BOVIN ( 337) NQILDPWVYLLLRKILL 35
PER3_HUMAN ( 338) NQILDPWVYLLLRKILL 35

YNB4_CAEEL ( 331) SCVAYPLIPTLLNRGIR 100
```

Profiles

- Based on entire sequences.
- Profiles define which residues are allowed at given positions, which positions are highly conserved and which degenerate, which positions can tolerate insertions.
- The scoring system may include evolutionary weights and results from structural analysis.

Example of PROSITE profile

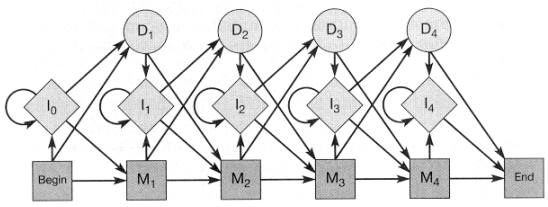


Hidden Markov Models

- Based on entire sequences.
- HMMs are probabilistic models consisting of a number of interconnecting states - linear chains of match, delete or insert states.
- Each position in the multiple alignment is assigned to either match, insert or delete state.
- Construction: seed alignment, iterative sequence gathering, final alignment (all automatic).

Scheme of the linear Hidden Markov Model

(M) match, (I) insertion, (D) deletion



Analysis packages

- commercial databases
- commercial software
- comprehensive packages
- packages for DNA analysis
- intranet packages
- Internet packages

Comprehensive packages

- GCG
- EGCG/EMBOSS
- Staden
- Lasergene

Packages for DNA analysis

- Sequencher
- VectorNTI
- MacVector

Intranet packages

- SYNERGY
- GeneMill, GeneWorld, GeneThesaurus

Internet packages

- CINEMA
- EGCG/EMBOSS
- Alfresco

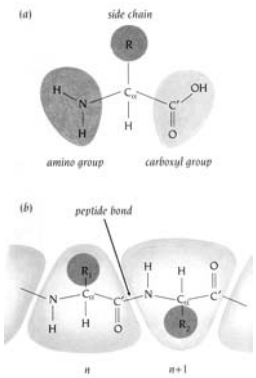
Protein structure modelling

- protein structure
- protein structure databases
- prediction of secondary structure
- prediction of protein fold
- prediction of tertiary structure
- modelling of protein-ligand complexes

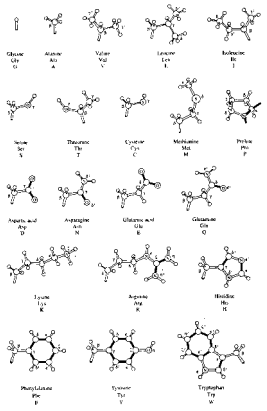
Protein structure

- Proteins are build up by amino acids that are linked by peptide bonds. The 20 different amino acids occur naturally in proteins.
- Protein structure can be experimentally determined by X-ray crystallography, nuclear magnetic resonance (NMR) or by electron crystallography.
- Levels of protein structure:
  - primary structure
  - secondary structure
  - supersecondary structure
  - tertiary structure
  - quaternary structure

Scheme of an amino acids (a) and polypeptide chain (b)



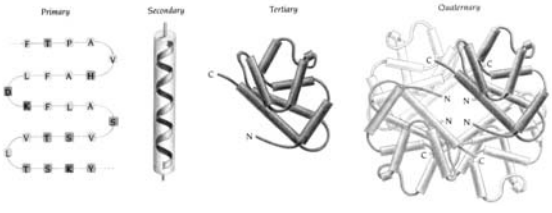
Side chains of 20 different amino acids that occur in proteins



Levels of protein structure

Primary structure:	the linear sequence of amino acids in a protein molecule
Secondary structure:	regions of local regularity within a protein fold (e.g., α-helices, β-turns, β-strands)
Super-secondary structure:	the arrangement of α-helices and/or β-strands into discrete folding units (e.g., β-barrels, βαβ-units, Greek keys, etc.)
Tertiary structure:	the overall fold of a protein sequence, formed by the packing of its secondary and/or super-secondary structure elements
Quaternary structure:	the arrangement of separate protein chains in a protein molecule with more than one subunit
Quinternary structure:	the arrangement of separate molecules, such as in protein–protein or protein–nucleic acid interactions

Levels of protein structure



Synchrotron radiation facility

European Synchrotron Radiation Facility at Grenoble, France

Protein structure databases

- PDB
- PDBsum

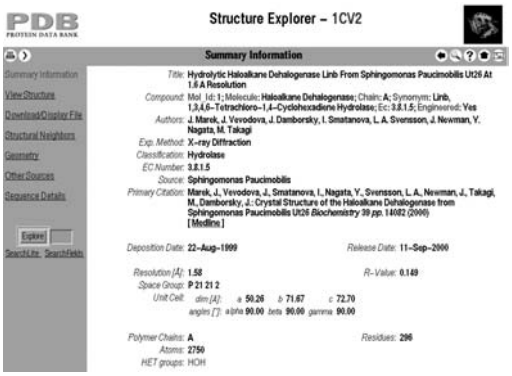
Protein structure classification databases

- SCOP
- CATCH

Protein structure databases

- PDB - Protein Data Bank
  - developed at Brookhaven National Laboratory
  - currently maintained by Research Collaboratory for Structural Bioinformatics (RCSB)
  - world repository of three-dimensional protein structures
  - entries from crystallographic analysis (80%), nuclear magnetic resonance (16%) and modelling (2%)
  - entries stored as flat files composed of section for information records and section for co-ordinates
  - entries identified by unique PDB-ID code (e.g., 1EDE)
  - searchable by keywords
  - interactive visualization of structures

Information on entry from the PDB database



PDB
PROTEIN DATA BANK

Structure Explorer – 1CV2

Summary Information

Summary Information
View Structure
Download/Display File
Structural Neighbors
Geometry
Other Sources
Sequence Details
Explore
Search by: Search Fields

Title: Hydrolytic Haloalkane Dehalogenase Linb From Sphingomonas Paucimobilis Ut26 At 1.6 A Resolution
Compound: Mol. Id: 1; Molecule: Haloalkane Dehalogenase; Chain: A; Synonym: Linb, 1,3,4,6-Tetrachloro-1,4-Cyclohexadiene Hydrolase; Ec: 3.8.1.5; Engineered: Yes
Authors: J. Marek, J. Vevodova, J. Damborsky, I. Smatanova, L. A. Svensson, J. Newman, Y. Nagata, M. Takagi
Exp. Method: X-ray Diffraction
Classification: Hydrolase
EC Number: 3.8.1.5
Source: Sphingomonas Paucimobilis
Primary Citation: Marek, J., Vevodova, J., Smatanova, I., Nagata, Y., Svensson, L. A., Newman, J., Takagi, M., Damborsky, J.: Crystal Structure of the Haloalkane Dehalogenase from Sphingomonas Paucimobilis Ut26 *Biochemistry* 39 pp. 14082 (2000) [Medline]

Deposition Date: 22-Aug-1999 Release Date: 11-Sep-2000

Resolution [Å]: 1.58 R-Value: 0.149
Space Group: P 21 21 2
Unit Cell: dim [Å]: a 50.26 b 71.67 c 72.70
angles [°]: alpha 90.00 beta 90.00 gamma 90.00

Polymer Chains: A Residues: 296
Atoms: 2750
HET groups: HOH

Entry from the PDB database (header)

```
HEADER    HYDROLASE                               22-AUG-99   1CV2
TITLE     HYDROLYTIC HALOALKANE DEHALOGENASE LINB FROM SPHINGOMONAS
TITLE    2 PAUCIMOBILIS UT26 AT 1.6 A RESOLUTION
COMPND    MOL_ID: 1;
COMPND    2 MOLECULE: HALOALKANE DEHALOGENASE;
COMPND    3 CHAIN: A;
COMPND    4 SYNONYM: LINB, 1,3,4,6-TETRACHLORO-1,4-CYCLOHEXADIENE
COMPND    5 HYDROLASE;
COMPND    6 EC: 3.8.1.5;
COMPND    7 ENGINEERED: YES;
COMPND    8 BIOLOGICAL_UNIT: MONOMER
SOURCE    MOL_ID: 1;
SOURCE    2 ORGANISM_SCIENTIFIC: SPHINGOMONAS PAUCIMOBILIS;
SOURCE    3 STRAIN: UT26;
SOURCE    4 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE    5 EXPRESSION_SYSTEM_STRAIN: HB101;
SOURCE    6 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE    7 EXPRESSION_SYSTEM_PLASMID: PHYLB1
KEYWDS    DEHALOGENASE, LINDANE, BIODEGRADATION, ALPHA/BETA-HYDROLASE
EXPDTA    X-RAY DIFFRACTION
AUTHOR    J.MAREK,J.VEVODOVA,J.DAMBORSKY,I.SMATANOVA,L.A.SVENSSON,
AUTHOR   2 J.NEWMAN,Y.NAGATA,M.TAKAGI
REMARK    1 REFERENCE 1
REMARK    1  AUTH   I.SMATANOVA,Y.NAGATA,L.A.SVENSSON,M.TAKAGI,J.MAREK
REMARK    1  TITL   CRYSTALLIZATION AND PRELIMINARY X-RAY DIFFRACTION
REMARK    1  TITL 2 ANALYSIS OF HALOALKANE DEHALOGENASE LINB FROM
REMARK    1  TITL 3 SPHINGOMONAS PAUCIMOBILIS UT26
REMARK    1  REF    ACTA CRYST. D                 V. D53  1231 1999
REMARK    1  REFN                   DK ISSN 0907-4449
REMARK    2
REMARK    2 RESOLUTION. 1.58 ANGSTROMS.
REMARK    3
REMARK    3 REFINEMENT.
REMARK    3   PROGRAM     : SHELXL-97
REMARK    3   AUTHORS     : G.M.SHELDRICK
REMARK    3
```

Entry from the PDB database (crystallographic info)

```
REMARK   3  DATA USED IN REFINEMENT.
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 1.58
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 20.0
REMARK   3   DATA CUTOFF            (SIGMA(F)) : 0.000
REMARK   3   COMPLETENESS FOR RANGE        (%) : 94.2
REMARK   3   CROSS-VALIDATION METHOD           : THROUGHOUT
REMARK   3   FREE R VALUE TEST SET SELECTION   : RANDOM
REMARK
REMARK 290 CRYSTALLOGRAPHIC SYMMETRY
REMARK 290 SYMMETRY OPERATORS FOR SPACE GROUP: P 21 21 2
REMARK 290
REMARK 290      SYMOP   SYMMETRY
REMARK 290     NNNMMM   OPERATOR
REMARK 290       1555   X,Y,Z
REMARK 290       2555   -X,-Y,Z
REMARK 290       3555   1/2-X,1/2+Y,-Z
REMARK 290       4555   1/2+X,1/2-Y,-Z
REMARK 290
REMARK 290      WHERE NNN -> OPERATOR NUMBER
REMARK 290            MMM -> TRANSLATION VECTOR
REMARK 290
REMARK 290 CRYSTALLOGRAPHIC SYMMETRY TRANSFORMATIONS
REMARK 290 THE FOLLOWING TRANSFORMATIONS OPERATE ON THE ATOM/HETATM
REMARK 290 RECORDS IN THIS ENTRY TO PRODUCE CRYSTALLOGRAPHICALLY
REMARK 290 RELATED MOLECULES.
REMARK 290   SMTRY1   1  1.000000  0.000000  0.000000        0.00000
REMARK 290   SMTRY2   1  0.000000  1.000000  0.000000        0.00000
REMARK 290   SMTRY3   1  0.000000  0.000000  1.000000        0.00000
REMARK 290   SMTRY1   2 -1.000000  0.000000  0.000000        0.00000
REMARK 290   SMTRY2   2  0.000000 -1.000000  0.000000        0.00000
REMARK 290   SMTRY3   2  0.000000  0.000000  1.000000        0.00000
```

Entry from the PDB database (sequence, sec. elements)

```
DBREF  1CV2 A    1   296  DBJ    BAA03443 BAA03443         1    296
SEQRES   1 A  296  MET SER LEU GLY ALA LYS PRO PHE GLY GLU LYS LYS PHE
SEQRES   2 A  296  ILE GLU ILE LYS GLY ARG ARG MET ALA TYR ILE ASP GLU
SEQRES   3 A  296  GLY THR GLY ASP PRO ILE LEU PHE GLN HIS GLY ASN PRO
SEQRES   4 A  296  THR SER SER TYR LEU TRP ARG ASN ILE MET PRO HIS CYS
SEQRES   5 A  296  ALA GLY LEU GLY ARG LEU ILE ALA CYS ASP LEU ILE GLY
SEQRES   6 A  296  MET GLY ASP SER ASP LYS LEU ASP PRO SER GLY PRO GLU
SEQRES
HELIX    1   1 SER A   42  ALA A   53
HELIX    2   2 TYR A   82  LEU A   96
HELIX    3   3 TRP A  109  ARG A  120
HELIX    4   4 GLU A  145  ARG A  155
HELIX    5   5 GLY A  159  LEU A  164
HELIX    6   6 VAL A  168  LEU A  177
HELIX    7   7 GLU A  184  GLU A  192
HELIX    8   8 ARG A  202  ILE A  211
HELIX    9   9 ALA A  218  SER A  234
HELIX   10  10 THR A  250  ARG A  258
HELIX   11  11 ILE A  274  ASP A  277
HELIX   12  12 SER A  278  LEU A  293
SHEET    1   S1 8 LYS A  12  ILE A  14  0
SHEET    2   S1 8 MET A  21  GLU A  26 -1  N  MET A  21   O  ILE A  14
SHEET    3   S1 8 ARG A  57  ASP A  62 -1  N  ALA A  60   O  ILE A  24
SHEET    4   S1 8 ASP A  30  HIS A  36  1  N  ILE A  32   O  ARG A  57
SHEET    5   S1 8 VAL A 102  HIS A 107  1  N  VAL A 103   O  PRO A  31
SHEET    6   S1 8 VAL A 125  MET A 131  1  N  ALA A 129   O  LEU A 104
SHEET    7   S1 8 LYS A 238  PRO A 245  1  N  ILE A 241   O  TYR A 130
SHEET    8   S1 8 GLN A 263  GLY A 270  1  N  THR A 264   O  LYS A 238
CISPEP   1 ASN A   38    PRO A   39          0        -2.50
CISPEP   2 ASP A   73    PRO A   74          0        -2.40
CISPEP   3 THR A  216    PRO A  217          0        -3.84
CISPEP   4 GLU A  244    PRO A  245          0         3.01
CISPEP   5 PRO A  295    ALA A  296          0        20.14
```

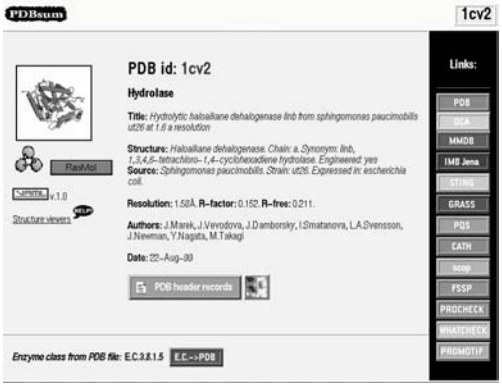
Entry from the PDB database (co-ordinates)

```
ATOM      1  N   GLY A   4       7.096   3.531   6.684  1.00 20.14           N
ATOM      2  CA  GLY A   4       7.885   3.530   5.461  1.00 31.48           C
ATOM      3  C   GLY A   4       9.300   4.043   5.633  1.00 20.62           C
ATOM      4  O   GLY A   4       9.599   4.865   6.501  1.00 12.02           O
ATOM      5  N   ALA A   5      10.240   3.571   4.814  1.00 15.21           N
ATOM      6  CA  ALA A   5      11.609   4.057   4.935  1.00 10.23           C
ATOM      7  C   ALA A   5      11.883   5.182   3.955  1.00 11.96           C
ATOM      8  O   ALA A   5      12.950   5.809   3.978  1.00 13.69           O
ATOM      9  CB  ALA A   5      12.621   2.943   4.674  1.00 10.47           C
ATOM     10  N   LYS A   6      10.929   5.437   3.056  1.00 12.31           N
ATOM     11  CA  LYS A   6      11.251   6.452   2.053  1.00 17.87           C
ATOM     12  C   LYS A   6      11.223   7.850   2.660  1.00  9.09           C
ATOM     13  O   LYS A   6      10.310   8.161   3.422  1.00 10.53           O
ATOM     14  CB  LYS A   6      10.274   6.419   0.870  1.00 20.08           C
ATOM     15  CG  LYS A   6      10.901   6.898  -0.436  1.00 47.79           C
ATOM     16  CD  LYS A   6      10.695   8.377  -0.703  1.00 63.02           C
ATOM     17  CE  LYS A   6      11.654   8.950  -1.734  1.00 62.33           C
ATOM     18  NZ  LYS A   6      11.574  10.435  -1.832  1.00 50.12           N
ATOM     19  N   PRO A   7      12.171   8.696   2.307  1.00 12.48           N
ATOM     20  CA  PRO A   7      12.108  10.087   2.748  1.00 15.50           C
ATOM     21  C   PRO A   7      10.895  10.808   2.144  1.00 16.27           C
ATOM     22  O   PRO A   7      10.244  10.396   1.170  1.00 15.29           O
ATOM     23  CB  PRO A   7      13.394  10.717   2.217  1.00 12.40           C
ATOM     24  CG  PRO A   7      13.877   9.803   1.151  1.00 23.02           C
ATOM     25  CD  PRO A   7      13.347   8.427   1.456  1.00 21.52           C
ATOM     26  N   PHE A   8      10.608  11.942   2.751  1.00 10.86           N
ATOM     27  CA  PHE A   8       9.557  12.848   2.302  1.00  6.69           C
ATOM     28  C   PHE A   8      10.134  13.914   1.384  1.00 17.38           C
ATOM     29  O   PHE A   8      11.121  14.590   1.716  1.00 17.05           O
ATOM     30  CB  PHE A   8       8.912  13.490   3.531  1.00  6.78           C
ATOM     31  CG  PHE A   8       7.776  14.444   3.183  1.00 12.53           C
ATOM     32  CD1 PHE A   8       6.526  13.921   2.874  1.00 18.59           C
ATOM     33  CD2 PHE A   8       7.984  15.811   3.166  1.00 15.33           C
ATOM     34  CE1 PHE A   8       5.494  14.797   2.547  1.00 19.62           C
ATOM     35  CE2 PHE A   8       6.961  16.701   2.851  1.00 15.73           C
ATOM     36  CZ  PHE A   8       5.718  16.160   2.537  1.00 22.66           C
ATOM     37  N   GLY A   9       9.544  14.110   0.215  1.00 18.38           N
```

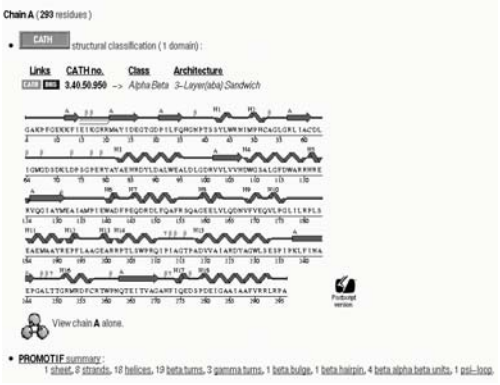
Protein structure databases

- PDBsum
  - developed at University College London
  - summaries and analyses of protein structures (secondary database derived from PDB)
  - summary of PDB entries: resolution, R-factor, # protein chains, topology, ligands, metal ions, etc.
  - analysis of PDB entries: protein-metal and protein-ligand interactions, protein validation
  - provides links to many related databases

Information on entry from the PDBsum database



Entry from the PDBsum database (secondary elements)



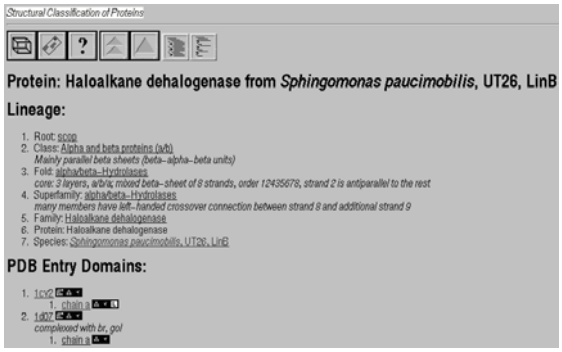
Protein structure classification databases

- Classification attempts to capture the structural similarities among proteins.
- The structural similarities relate to the evolution.
- The structural similarities may imply function.
- The classification scheme is dependent on the underlying philosophy.

Protein structure classification databases

- SCOP - Structural Classification of Proteins
  - developed at MRC Laboratory of Molecular Biology
  - construction: combination of manual and automatic methods (complicated by multidomain proteins)
  - fold = same secondary elements in same arrangement, independently of common evolutionary origin
  - superfamily = low identity but common evolutionary origin implied from common structure and function
  - family = sequence identity >30%

Information on entry from the SCOP database



Protein structure classification databases

- CATCH - Class, Architecture, Topology, Homology
  - developed at University College London
  - construction: mostly automatic
  - unique numbering scheme analogous to Enzyme Classification (E.C.) scheme
  - class = gross secondary structure content
  - architecture = gross secondary structure arrangement
  - topology = shape and connectivity of secondary structures (60% of larger protein matches smaller one)
  - homology = sequence identity >35%, common ancestry
  - sequence = clustering based on sequence identity

Information on entry from the CATH database

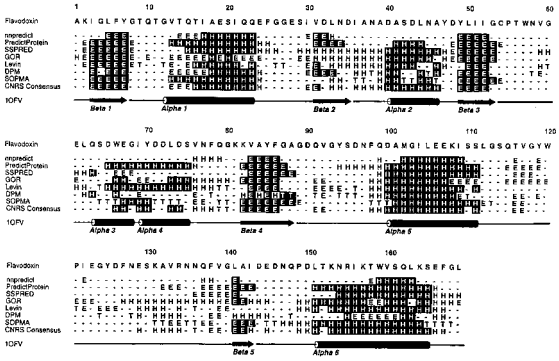
Prediction of secondary structure

- Algorithms assign probability for occurrence of α-helix, β-strand, turn and random coil at particular position in the sequence.
- Methods: statistical, stereochemical and homology/neural networks based. All methods rely on information derived from known 3D structures. Most recent methods use the information from multiple alignments.
- Reliability of the best current methods is >70%.

Prediction of secondary structure

- Chou-Fasman and GOR
  - statistical - amino acids show preference for particular secondary structure elements
- PHD and NNPredict
  - neural networks - the rules for prediction are not defined in advance, they are created by training
- NNSSP and PREDATOR
  - nearest neighbour approach
- JPRED
  - consensus approach - utilises multiple alignments and state-of-art method - makes consensus

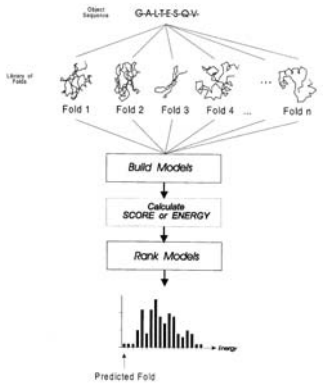
Comparison of secondary structure predictions



Prediction of protein fold

- Threading
  - threading = protein fold assignment or fold recognition
  - target sequence is searched against database of folds (3D profiles) and threaded models are constructed
  - 3D profile - each residue in 3D structure is assigned environmental variables (buried area, fraction of side chain covered polar atoms, secondary structure, etc.). Assumption - environment of the residue should be more conserved that the residue itself.
  - residue can be also described by its interactions
  - match of target sequence with 3D profile (quality of threaded models) is quantified by Z-score or energy
  - limitation: can not handle multi-domain proteins

Fold recognition by threading



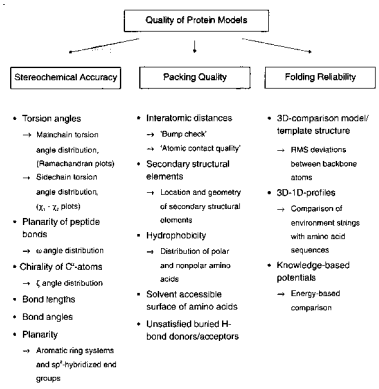
Prediction of protein fold

- Bioinbgu
  - consensus method utilising predictions from five different algorithms
- 3D-PSSM
  - scoring functions: 1D-PSSMs (sequence profiles built from relatively close homologues), 3D-PSSMs (more general profiles containing more remote homologues), matching of secondary structure elements, and propensities of the residues for solvent accessibility
- GenThreader
  - hybrid method: profile-based alignment, evaluation of alignments by threading, evaluation of threaded models by neural network

Prediction of tertiary structure

- *Ab initio*
  - 3D structure of a protein is predicted from first principles (search for global minimum structure)
  - current algorithms are not very reliable
- Homology modelling
  - 1. alignment of modelled sequence against sequences of structurally similar proteins (templates)
  - 2. "extraction" of the backbone from template structure and positioning of side-chain
  - 3. modelling of loops
  - 4. structure refinement and validation

Validation of protein models



Prediction of tertiary structure

- SWISS-MODEL
  - fully automated modelling server
  - input = protein sequence; output = PDB file
  - 1. search of ExNRL-3D using BLASTP for potential templates; 2. select all templates with sequence identities above 25%; 3. Generate structures of 3D models; 4. energy minimise models using GROMOS 96
  - first approach and optimise mode (Swiss-PDBViewer)
- MODELLER
  - most widely used academic program for homology modelling (satisfaction of spatial restrains)

Modelling of protein-ligand complexes

- Docking
  - positioning of small organic molecules (ligands) inside the protein active site
  - different orientations and conformations of the ligand are evaluated using geometric or energetic scoring functions
  - Protein-ligand interaction energy = van der Waals term + electrostatic term + H-bond term + entropic term
  - flexible docking - considers different conformation of ligand; different rotamers of protein side chains
- Software: DOCK, AUTODOCK, FLEX