Bioinformatics

Protein information resources

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

Protein information resources

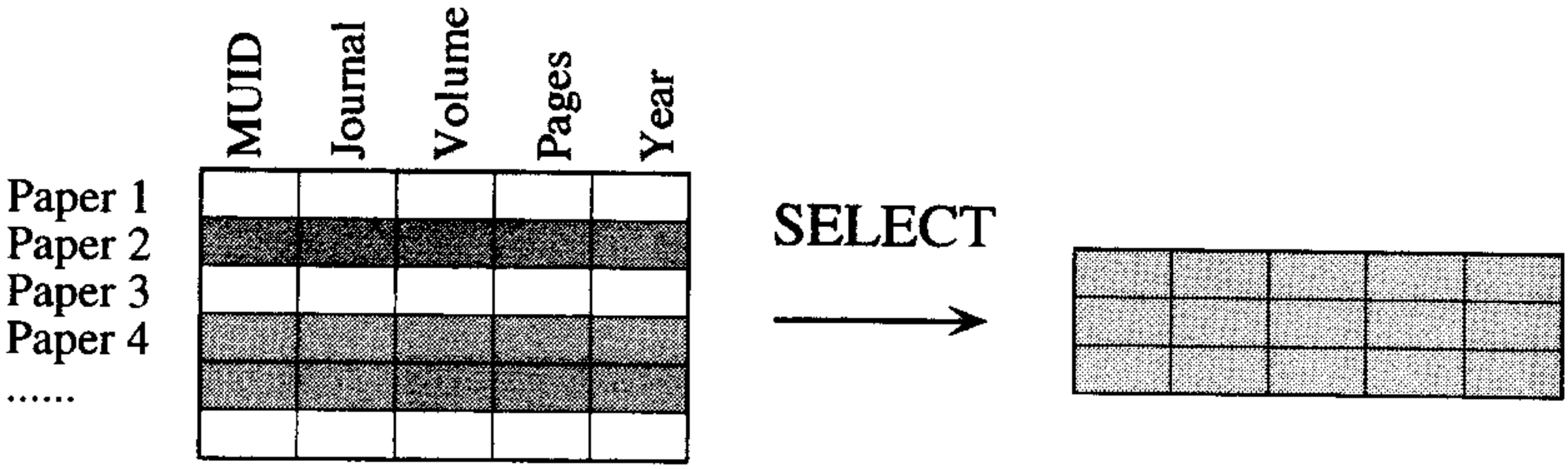
- biological databases - introduction
- primary protein sequence databases
- composite protein sequence databases
- secondary databases
- composite secondary databases
- protein structure databases
- protein structure classification databases

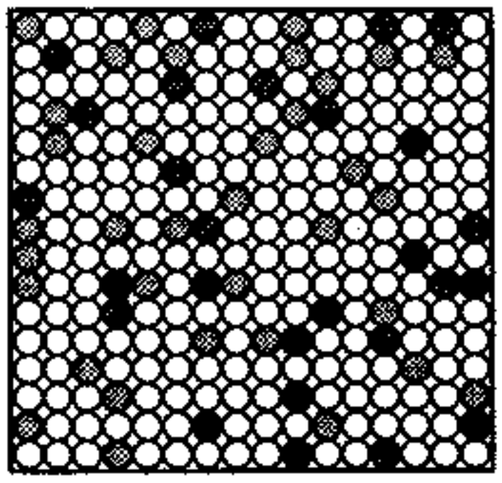
Biological databases - introduction

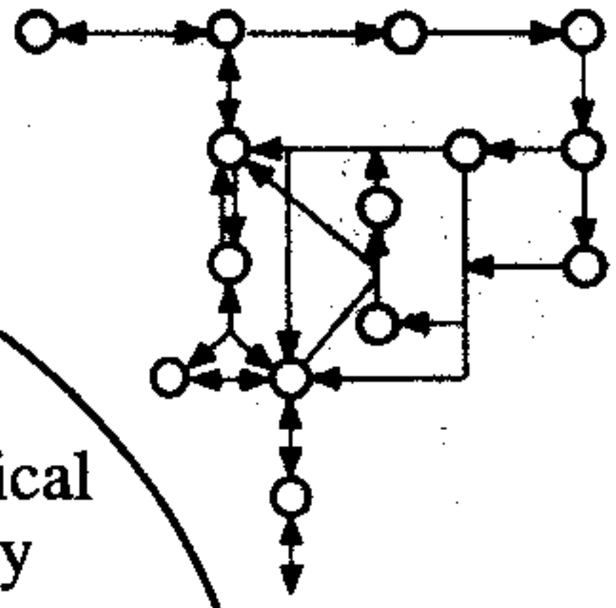
- Vast amounts of data produced - databases must be established for storage of the data.
- Databases must be maintained and disseminated together with the analysis tools.
- Classification of databases
  - flat files
  - relational
  - object-oriented

  - primary
  - secondary
  - composite

```
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 ctttcgagga aaattctcgg acccccatat acaaatatcg
      121 gaaaaagtat cgaacagttt cgcgacgcga agcgttaaga tcgccaaaag atctccgtgc
      181 ggaaacaaag aaattgaggc actattaaga gattgttgtt gtgcgcgagt gtgtgtcttc
      241 agctgggtgt gtggaatgtc aactgacggg ttgtaaaggg aaaccctgaa atccgaacgg
      301 ccagccaaag caaataaagc tgtgaatacg aattaagtac aacaaacagt tactgaaaca
      361 gatacagatt cggattcgaa tagagaaaca gatactggag atgcccccag aaacaattca
      421 attgcaaata tagtgcgttg cgcgagtgcc agtggaaaaa tatgtggatt acctgcgaac
      481 cgtccgccca aggagccgcc gggtgacagg tgtatccccc aggataccaa cccgagccca
      541 gaccgagatc cacatccaga tcccgaccgc agggtgccag tgtgtcatgt gccgcggcat
      601 accgaccgca gccacatcta ccgaccaggt gcgcctcgaa tgcggcaaca caattttcaa
          ..........
     3841 aactgtataa acaaaacgta tgccctataa atatatgaat aactatctac atcgttatgc
     3901 gttctaagct aagctcgaat aaatccgtac acgttaatta atctagaatc gtaagaccta
     3961 acgcgtaagc tcagcatgtt ggataaatta atagaaacga g
//
```


MUID
Journal
Volume
Pages
Year
Paper 1
Paper 2
Paper 3
Paper 4
.....
SELECT
PROJECT
JOIN
MUID
Author
Author 1-1
Author 1-2
Author 2-1
Author 2-2
Author 2-3
Author 3-1
.....
MUID
Journal
Volume
Pages
Year
Author





Gene expression profile

Biochemical pathway

```
atggcgacccgcagc
cctggcgtcgtggtg
agcagctcggcctgc
cggccctggccggtt
cagg......
```

Nucleotide sequence

Gene

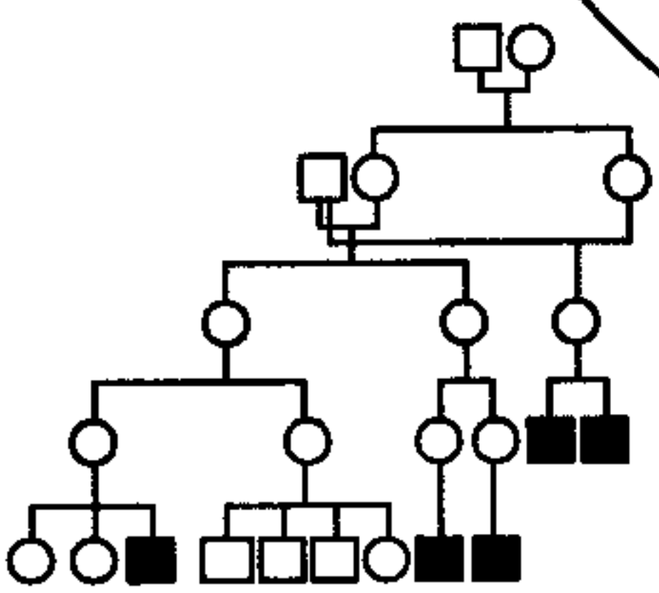
Amino acid sequence

```
ATRSPGVVISDDEPG
YDLDLFCIPNHYAED
LERVFIPHGLIMDRT
ERLARDVMKEMGGHH
IVAL......
```

Linkage

3-D Structure

Chromosome map



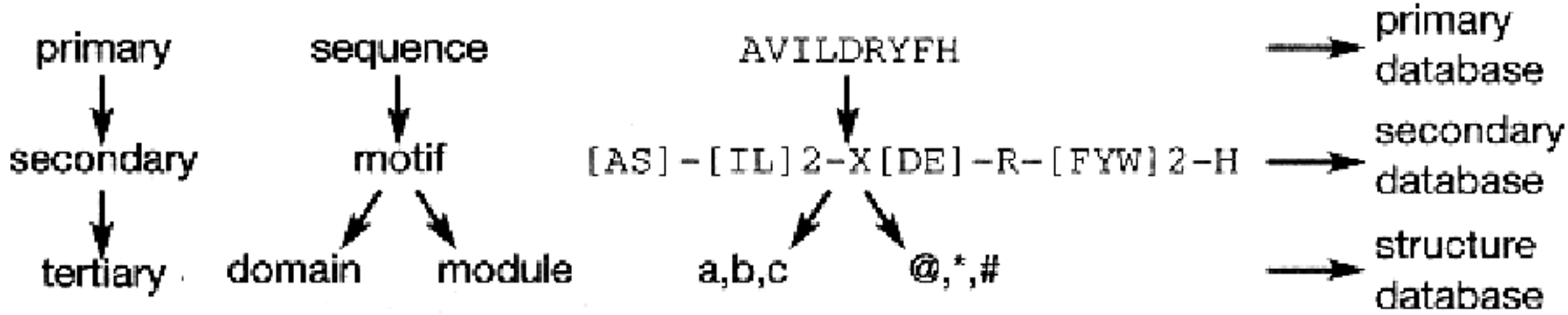
Similarity(X)

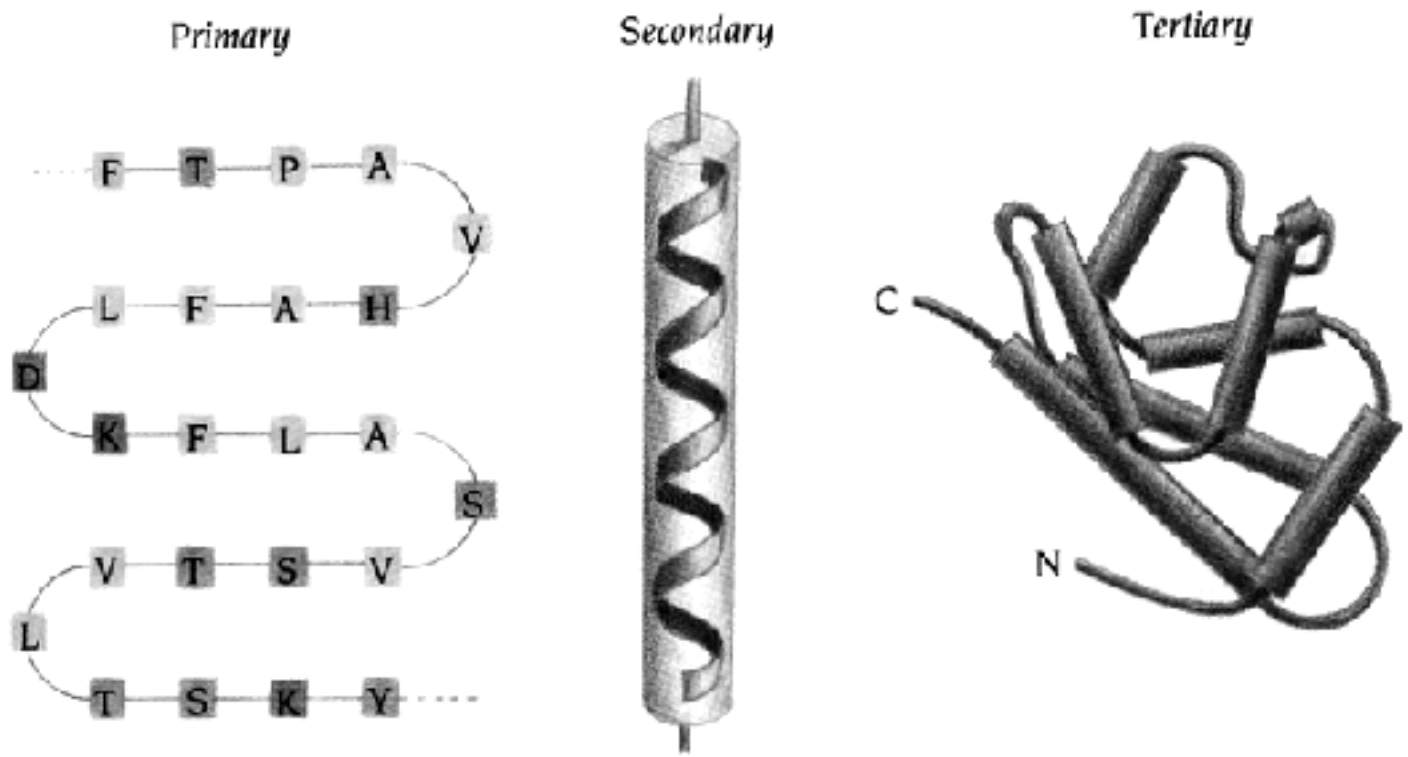
object X

message

class sequence
Similarity
.....
class structure
Similarity
.....
class expression
Similarity
.....
class pathway
Similarity
.....

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

```
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 SHHKKSHHHR 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

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

Composite protein sequence databases

- NRDB - Non-Redundant DataBase
  - developed and maintained by NCBCI
  - 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, GenPept, 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

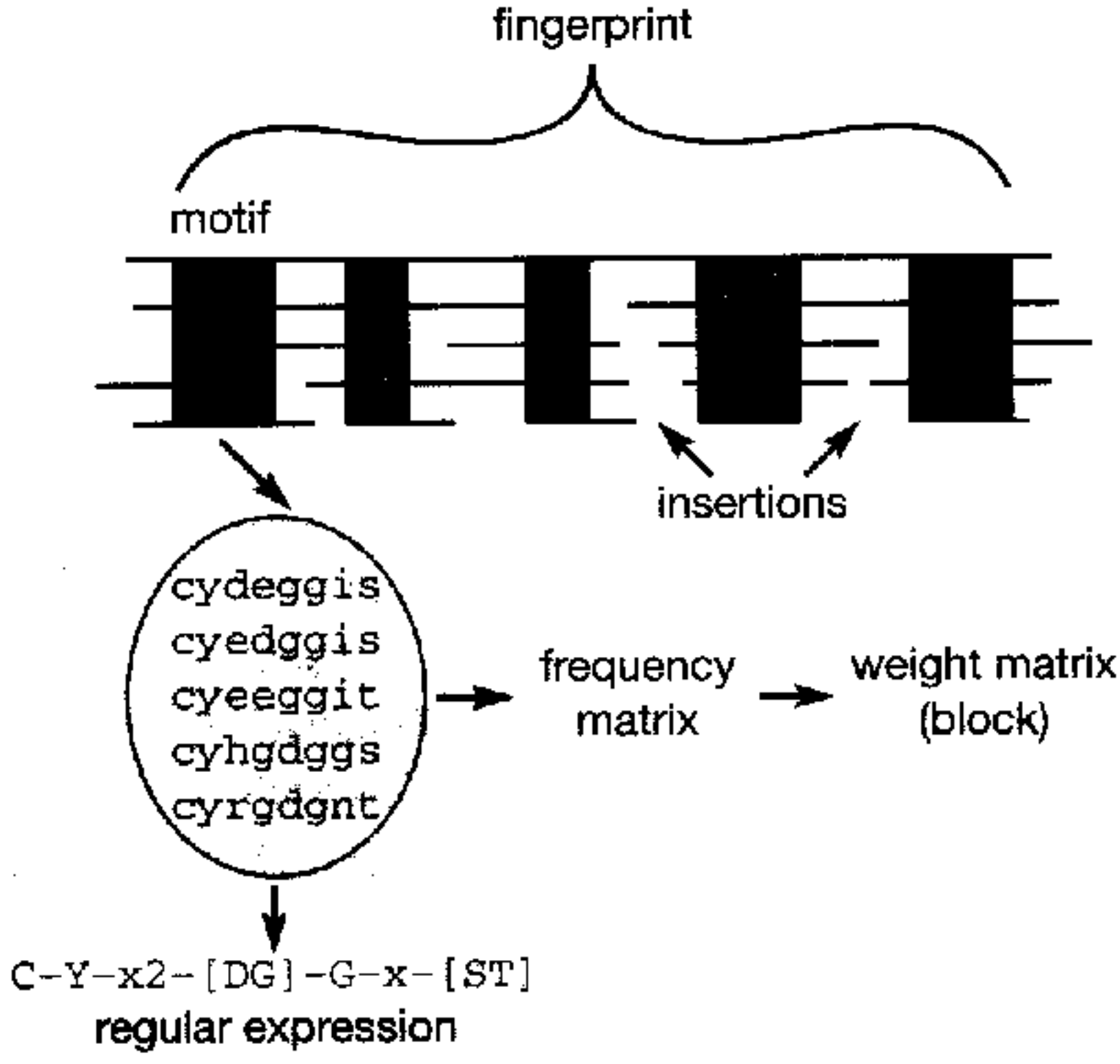
- SWISS-PROT+TrEMBL
  - developed and maintained by EBI
  - composite: SWISS-PROT, TrEMBL
  - advantages: comprehensive, minimally redundant, fewer errors
  - disadvantages: not as up-to-date as NRDB
  - available from SRS of EBI

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	

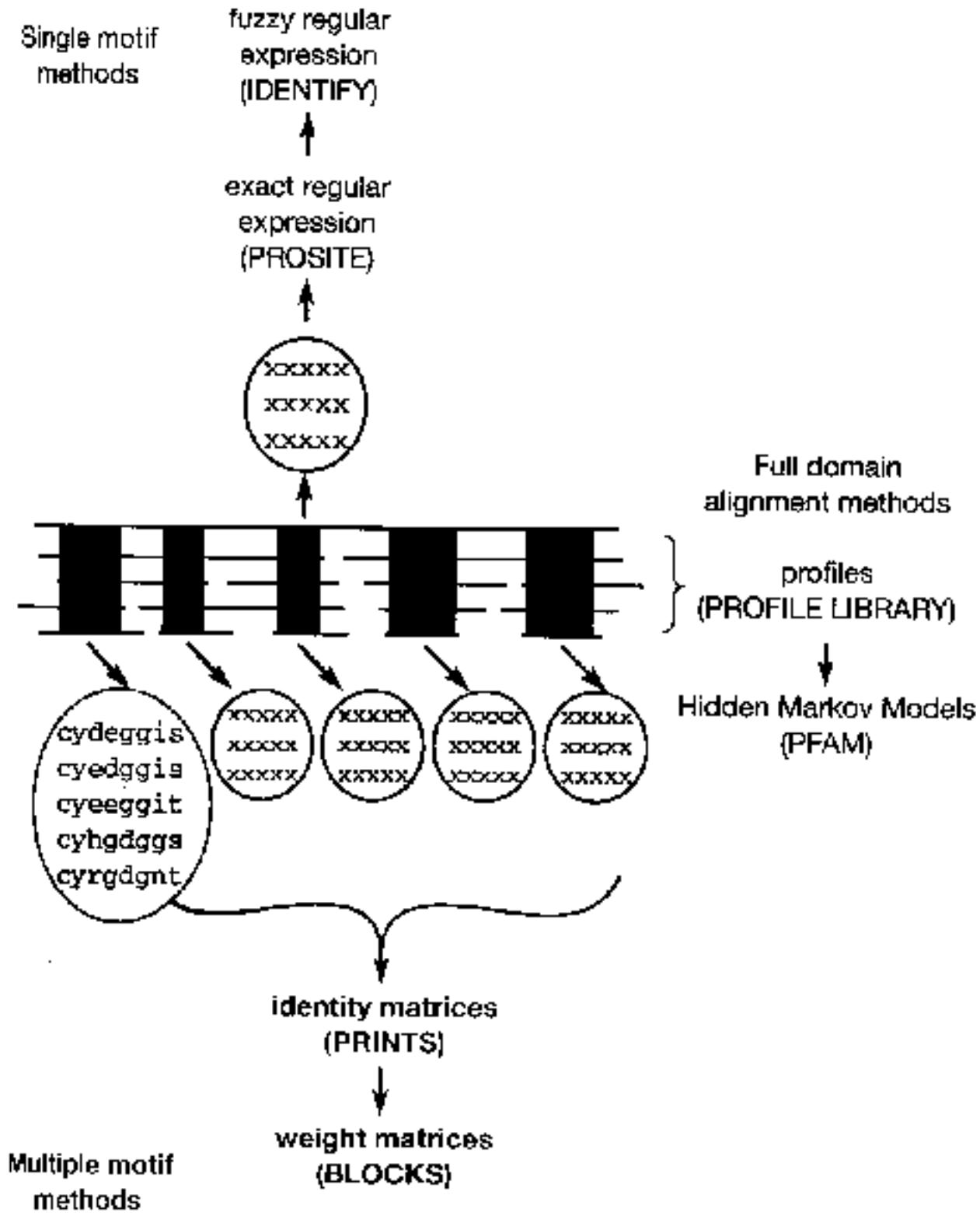
Secondary databases

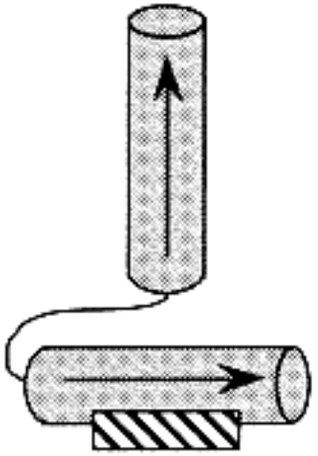
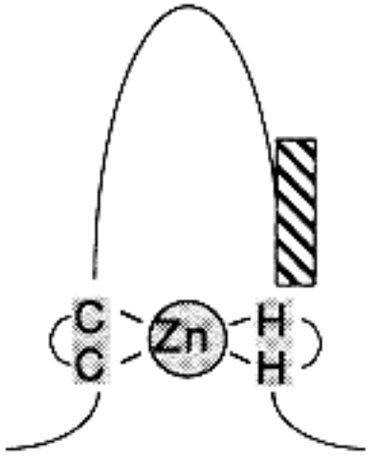
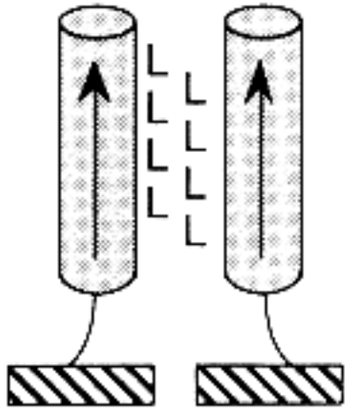
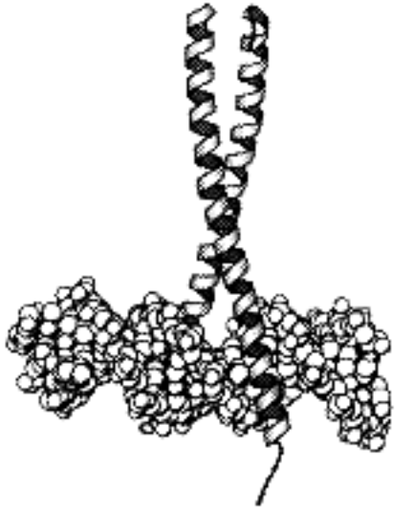
- Contains information derived from primary sequence data, typically in the form of abstractions: regular expressions, fingerprints, blocks, profiles or Hidden Markov Models.
- These abstractions represent destillations of the most conserved features of multiple alignments.
- The abstractions are useful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

Terms used in sequence analysis methods



Three principal methods for building secondary databases



Name	Helix-loop-helix (Myc type)	Cys-His zinc finger	Leucine zipper
Sequence	[DENSTAP]-K-[LIVMWAGN]-{FYWCPHKR}-[LIVT]-[LIV]-x(2)-[STAV]-[LIVMSTAC]-x-[VMFYH]-[LIVMTA]-{P}-{P}-[LIVMSR]	C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H	L-x(6)-L-x(6)-L-x(6)-L
Structure			
Function	DNA Binding	DNA Binding	DNA Binding
Example			
	3CRO	2DRP	1YSA

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

```
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)-[DENF]-[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  P08099, 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_SPHSP, 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_POMMI, T;
DR  P02700, OPSD_SHEEP, T; P29403, OPSD_XENLA, T; P22671, OPSD_LAMJA, T;
DR  P31355, OPSD_RANPI, T; P24603, OPSD_LOLFO, T; P09241, OPSD_OCTDO, T;
DR  P35356, OPSD_PROCL, T; P31356, OPSD_TODPA, T; P35360, OPS1_LIMPO, T;
DR  P35361, OPS2_LIMPO, T; P32310, OPSB_CARAU, T; P28682, OPSB_CHICK, T;
DR  P35357, OPSB_GECGE, T; P03999, OPSB_HUMAN, T; P28684, OPSV_CHICK, T;
DR  P22330, OPSG_ASTFA, T; P22331, OPSH_ASTFA, T; P32311, OPSG_CARAU, T;
DR  P32312, OPSH_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;
```

{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-retinal. 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 pigments (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 (R1-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)-[DENF]-[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] Fryxell K.J., Meyerowitz E.M.
J. Mol. Evol. 33:367-378(1991).
[3] Shen D., Jiang M., Hao W., Tao L., Salazar M., Fong H.K.W.
Biochemistry 33:13117-13125(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

(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
```

(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
YIFATTKSLRTPA	OPS1_DROME	73	73
VATLRYKKLRQPL	OPSB_HUMAN	57	57
YIFGGTKSLRTPA	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

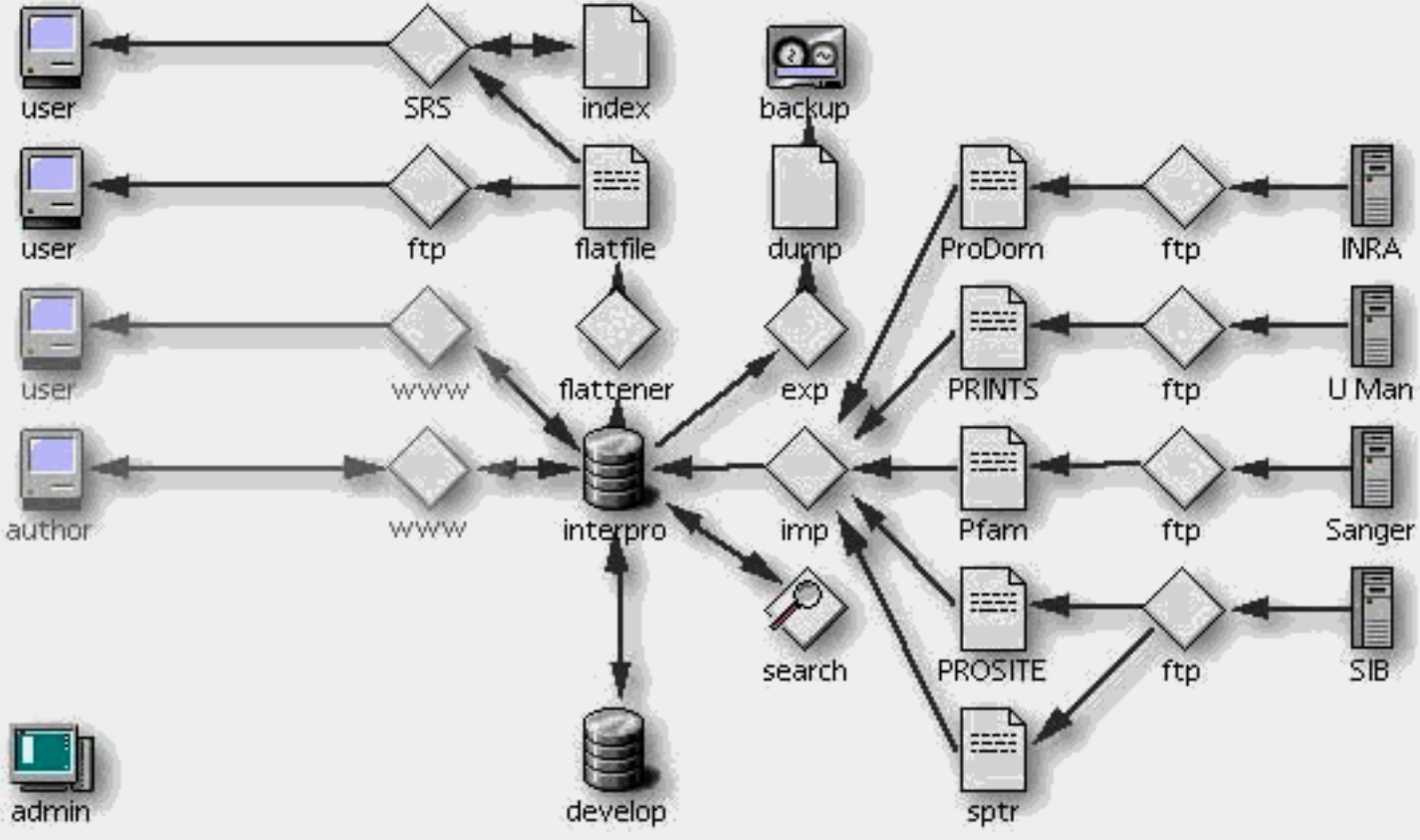
Properties of amino acids used in eMOTIF

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

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


user
SRS
index
backup
user
ftp
flatfile
dump
ProDom
ftp
INRA
user
www
flattener
exp
PRINTS
ftp
U Man
author
www
interpro
imp
Pfam
ftp
Sanger
search
PROSITE
ftp
SIB
admin
develop
sptr

Protein structure databases

- PDB
- PDBsum

Protein structure classification databases

- SCOP
- CATCH