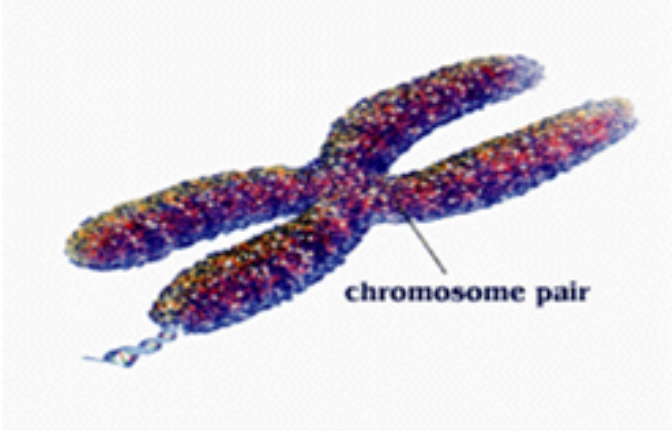
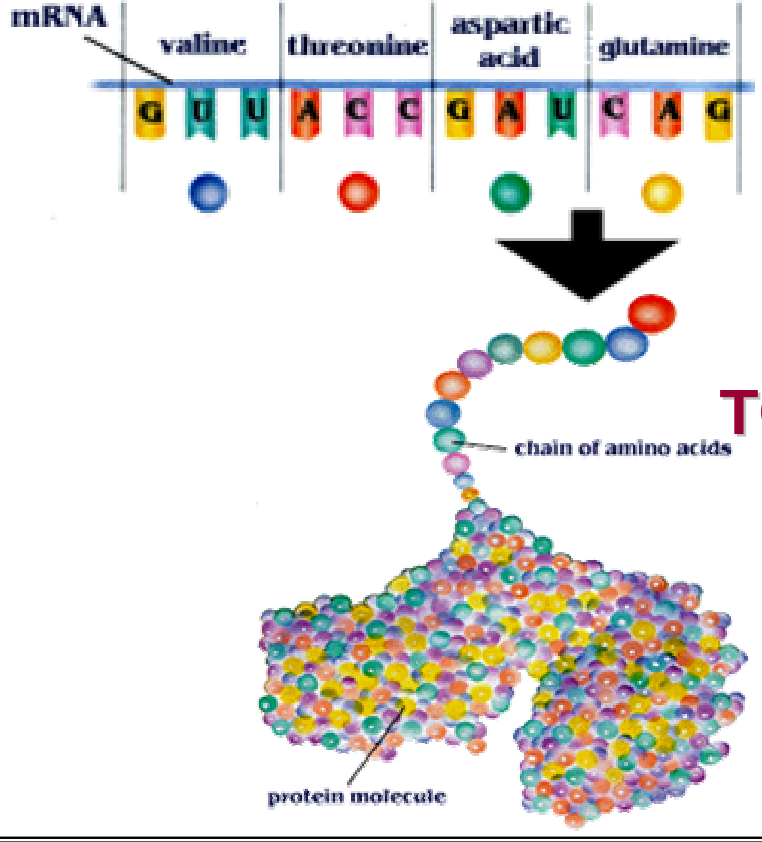


Lecture 92-09

PROTEIN SEQUENCE ANALYSIS





PROTEIN DATABASES

TOOLS

- PROTEIN SEQUENCE
- PROPERTIES
- MOTIF/DOMAIN
- FOLDINDING

Protein Sequence Databases



PIR - International Protein Sequence Database
http://pir.georgetown.edu



Swiss Institute of Bioinformatics
Institut Suisse de Bioinformatique
Schweizerisches Institut für Bioinformatik



http://www.isb-sib.ch/



Protein Data Bank
http://www.rcsb.org

Protein Sequence Analysis Tools



ExPASy Molecular Biology Server *http://expasy.nhri.org.tw*
The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE.

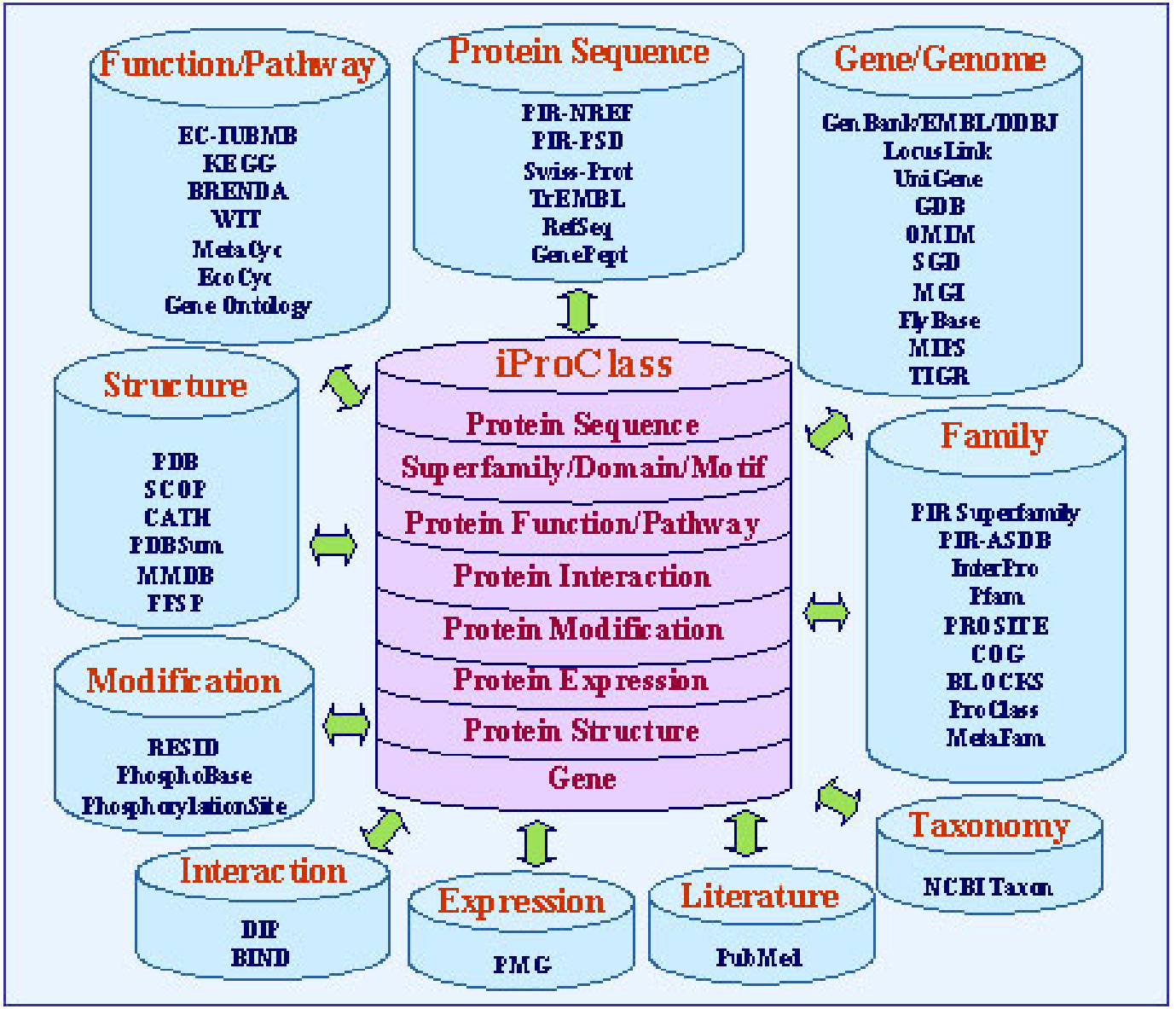


PIR-International Protein Sequence Database

Protein Sequence Database (PSD)
(http://pir.georgetown.edu/pirwww/search/textpsd.shtml) of functionally annotated protein sequences, which grew out of the *Atlas of Protein Sequence and Structure* (1965-1978) edited by Margaret Dayhoff and has been incorporated into an integrated knowledge base system of value-added databases and analytical tools.

***i*ProClass**, (*http://pir.georgetown.edu/iproclass*) a central point for exploration of protein information, provides summary descriptions of protein family, function and structure for PIR-PSD, Swiss-Prot, and TrEMBL sequences, with links to over 50 biological databases

PIR-NREF, *(http://pir.georgetown.edu/pirwww/search/pirnref.shtml)* a comprehensive database for sequence searching and protein identification, contains non-redundant protein sequences from PIR-PSD, Swiss-Prot, TrEMBL, RefSeq, GenPept, and PDB.


Function/Pathway
EC-IUBMB
KEGG
BRENDA
WIT
MetaCyc
EcoCyc
Gene Ontology
Protein Sequence
PIR-NREF
PIR-PSD
Sviss-Prot
TrEMBL
RefSeq
GenePept
Gene/Genome
GenBank/EMBL/DDBJ
LocusLink
UniGene
GDB
OMIM
SGD
MGI
FlyBase
MIPS
TIGR
Structure
PDB
SCOP
CATH
PDBSum
MMDB
FFSP
iProClass
Protein Sequence
Superfamily/Domain/Motif
Protein Function/Pathway
Protein Interaction
Protein Modification
Protein Expression
Protein Structure
Gene
Family
PIR Superfamily
PIR-ASDB
InterPro
Pfam
PROSITE
COG
BLOCKS
ProClass
MetaFam
Modification
RESID
PhosphoBase
PhosphorylationSite
Taxonomy
NCBI Taxon
Interaction
DIP
BIND
Expression
PMG
Literature
PubMed



PIR-International Protein Sequence Database

PIR is, in part, a redundant database. Sequences are made public as soon as the database curators receive them, even before annotation or classification is verified. Redundancy has it's disadvantages, most notably the repetition of sequences in different entries may include discrepencies. The redundancy at PIR can be advantages, as sequences are made public very quickly. The database is updated weekly.

The PIR-International protein sequence database is partitioned into four sections: PIR1-PIR4. There is no clear cut difference between the entries in PIR1 and PIR2.

PIR1

Classified, annotated, verified and non-redundant with respect to other PIR1 entries.

PIR2

Essentially indistinguishable from PIR1. Classification may not be quite so extensive as in PIR1.

PIR3

Not classified, annotated or verified. No attempts have been made to reduce redundancy.

PIR4

Unencoded or untranslated



Swiss Institute of Bioinformatics



Institut Suisse de Bioinformatique
Schweizerisches Institut für Bioinformatik

http://www.isb-sib.ch/

SWISS-PROT (established 1986) is a protein sequence database, accessible from the Swiss EMBL Outstation, EXPASY.

SWISS-PROT excels in annotation, exhibits very little redundancy and is thoroughly integrated with other databases. The extensive annotation and exhaustive to reduce redundancy mean that entries can take time before they are made available, but when they are, they are a complete and thorough resource. Annotation is updated with information from published review articles, and by external expert referees.

The entries are similar in layout to EMBL entries, with similar two letter codes defining the contents of each line. These include **CC** (comment), **FT** (feature table) and **KW** (keywords). Annotation includes information about the protein's function, post-translational modifications, disease associated deficiency, domains, structure and more. Where applicable, SWISS-PROT entries are cross referenced with PDB, a database of experimentally determined protein structure. Three dimensional (3D) models can be viewed with most web browsers, or files can be downloaded for local viewing.

View a SWISS-PROT Report | **NICE View a SWISS-PROT Report**

TrEMBL is a supplement to SWISS-PROT that contains computer annotated translations of EMBL. TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL Nucleotide Sequence Database, which are not yet integrated into SWISS-PROT. When entry annotation and verification is complete, it is moved from TrEMBL to SWISS-PROT (assuming the entry does not already exist, in which case they will be merged). Since preparing entries for SWISS-PROT is so time consuming, TrEMBL basically attempts to bridge the gap, and provide a redundant database of (less extensively) annotated translations of coding sequences (CDS) that are not listed in SWISS-PROT.

TrEMBL has two main sections.

SW-TrEMBL (SWISS-PROT TrEMBL), which contains sequences that are en route to SWISS-PROT.

REM-TrEMBL stores the remaining entries. This includes entries specifically excluded from SWISS-PROT, such as the many variations of immunoglobulins and T-cell receptors, synthetics sequences, fragments of less than eight amino acids, CDS from patent applications and EMBL CDS translations where the curators have strong evidence that the nucleotide does not code for real proteins.

PROTEIN DATA BANK http://www.rcsb.org

The **Brookhaven Protein Data Bank** (PDB) is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the National Institute of Standards and Technology -- three members of the Research Collaboratory for Structural Bioinformatics (RCSB). The PDB is supported by funds from the National Science Foundation, the Department of Energy, and two units of the National Institutes of Health: the National Institute of General Medical Sciences and the National Library of Medicine.

This database contains entries for molecular sequences, whose structure has been experimentally determined by X-ray crystallography or nucleic magnetic resonance imaging (NMR, MRI). The images presented have been experimentally acquired, and are not theoretical.

View a PDB Report

Secondarly Protein Sequence Databases

InterPro provides an Integrated resource of Protein Families, Domains and Sites of the commonly used signature databases, and has an intuitive interface for text- and sequence-based searches.



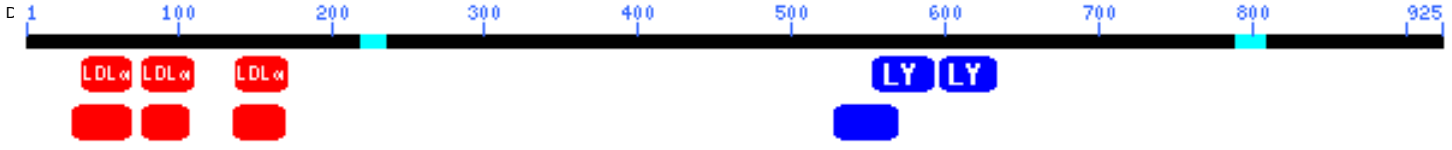
Bioinformatics infrastructural activities are crucial to modern biological research. Complete and up-to-date databases of biological knowledge are vital for the increasingly information-dependent biological and biotechnological research. Secondary protein databases on functional sites and domains like PROSITE, PRINTS, SMART, Pfam, ProDom, etc. are vital resources for identifying distant relationships in novel sequences, and hence for predicting protein function and structure. Unfortunately, these signature databases do not share the same formats and nomenclature, and each database has is own strengths and weaknesses.

To capitalise on these, the following partners: EBI, SIB, University of Manchester, Sanger Institute, GENE-IT, CNRS/INRA, LION bioscience AG and University of Bergen unified PROSITE, PRINTS, ProDom and Pfam into InterPro (Integrated resource of Protein Families, Domains and Sites). The latest databases to join the project were SMART, and more recently, TIGRFAMs.

Secondarly Protein Sequence Databases

NCBI Conserved Domain Search

NCBI will perform Conserved Domain Search when using blastp



Protein Sequence Analysis Tools



ExPASy Molecular Biology Server *http://expasy.nhri.org.tw*
The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE.

DATABASES
+
TOOLS

Protein Sequence Analysis in GCG

1.* ***CoilScan*** - Locates coiled-coil segments in protein sequences.
2.* ***HelicalWheel*** - plots a peptide sequence as a helical wheel.
3. ***HTHScan*** - Scans for helix-turn-helix motif.
4.* Isoelectric - Plots the charge as a function of pH for peptide sequence.
5.* ***Moment*** - Plots the helical hydrophobic moment of a peptide sequence.
6. ***Motifs*** - searching through proteins for the patterns defined in the PROSITE.
7. ***PepPlot*** - plots protein secondary structure and hydrophobicity in panels.
8. PeptideMap - Creates a map of an amino acid sequence.
9. ***PeptideSort*** - Shows fragments from a digest of an amino acid sequence.
10. ***PeptideStructure*** - Makes secondary structure predictions for a peptide sequence.
* 11. PlotStructure - Plots secondary structure from PeptideStructure output.
12. ***ProfileScan*** - Uses a database of profiles to find motifs in protein sequences.
13. Seg - Replaces low complexity regions in protein sequences with X characters.
14. ***SPScan*** - Scans protein sequences for secretory signal peptides
15. Xnu - Replaces tandem repeats in protein sequences with X characters.

Protein Sequence Analysis in SeqWEB

HmmerPfam

Compares one or more sequences to a database of profile hidden Markov models, such as the Pfam library, in order to identify known domains within the sequences.

PeptideStructure

Makes secondary structure predictions for a peptide sequence.

These predictions include (in addition to alpha, beta, coil, and turn) measures for antigenicity, flexibility, hydrophobicity, and surface probability. The predictions are displayed graphically.

CoilScan

Locate coiled-coil segments in protein sequences. .

HTHScan

Locate helix-turn-helix motifs in protein sequences.

SPScan

Locate secretory signal peptides in protein sequences.

PeptideSort

Shows the peptide fragments from a digest of an amino acid sequence.

It sorts the peptides by weight, position, and HPLC relative retention, and shows the composition of each peptide. It also prints a summary of the composition of the whole protein.

PepPlot

Plots predicted protein secondary structure and hydropathy plot. .

Moment

Makes a contour plot of the helical hydrophobic moment of a peptide sequence.

HelicalWheel

Plots a peptide sequence as a helical wheel to help you recognize amphiphilic regions or beta sheets.

Isoelectric

Plots the charge as a function of pH for a peptide sequence.

TransMem

Scans for likely transmembrane helices in a peptide sequence.

OTHERS

Motifs

Looks for sequence motifs by searching through proteins for the patterns defined in the PROSITE Dictionary of Protein Sites and Patterns. Motifs ca display an abstract of the current literature on eac of the motifs it finds.

Practical: *Gene*; *RNA*; *Protein*

U62639 (Gene)

```
aaaaatgtat gtctgatttt gaaatgctca tttcctttga ggtttccatt tttgagttgc 61   ccgtaatttg tatttttctg aagatgagca attcaatttt taaattgccc gcacctctac 121
cgtttccatc gtgtattttg ttaaaatatt cacagattaa cccatttacc gtttcatcca 181  cctgtttttc ctcgaaaaga ttccaatgtt ctataattct acaaaacttc ccacgcgaga 241
aacaactgta ataaactgaa tatattatct atcgcatcgt tttcaaccag aattaagcaa 301  gaggttccac aactttaaac accaacaacg caatcctaaa tcatttgcaa gattttattt 361
cagatgctac actttctgcc tgaaaaaaat tctgaaaagc cgaacaataa ttcatggtaa 421  caatgaatgg cagatacatc aaagttttag atgaacaatt tttatgtatt aaatgtacat 481
ttaaaaacaa attgcacaac gattctacta ctgtcgcact aattttacgt atgtctgtac 541  ttgaagattt cgaattaatt tgttcaatat tgtgttaaaa tgtttgattt atacactcaa 601
atctttaaaa gatttattgg aaaagataaa tggttaattt aaaccaaaaa tttccatcaa 661  gccttttctg aaaacactaa aattattttc gtggtgggac caggcgcgcg cgtcccatga 721
tgttccttta atcaaaatgc atttctgtcc cggcgggaga aattgaattt tgattttaag 781  gcgcgaattt ttgcctaaaa acgatgccat tctttcattc ttttcataat ctcactcacc 841
atgagaacca tgcgccttgc ttggttgctc ccacttttta ttcacatact aatcaaggta 901  atttccccgt ttttctagtt ttttcaatgt attttcatgt ttcagaacac agctcaagct 961
ccggctgtca acaactcgac atgcgatcaa gcaaaggaat ttgattgcgg gaacgggaga 1021 ctccgatgca ttcccgcgga gtggcaatgc gacaacgtag cggactgcga caaaggaaga 1081
gacgaatcgg gctgctcata tgcgcatcat tgttcgacaa gcttcatgtt atgcaagaat 1141 ggactgtgtg tcgcaaatga gttcaaatgc gacggcgaag acgactgccg cgatggaagc 1201
gatgagcagc attgcgagta caatatcctg aagtctcgct tcgatggttc caatccttcg 1261 gctcctacca ctttcgttgg tcacaatggc ccagaatgcc atcctcctcg tttacgatgc 1321
cgatcaggac aatgtattca accagatctc gtttgtgatg gacatcagga ttgttctgga 1381 ggagatgatg aggtcaactg caccagaagg ggacatgaaa atatgcagtc ctcgactgat 1441
tttcacgatg atgttcatct tgtcgatcca acctttttcg ctaatgaaga caataaggta 1501 attgtttaat gtttattaat ccgttttaac ttttatttt cagtgtcgga gtggatacac 1561
aatgtgccat agcggagacg tctgcatacc tgacagtttt ctttgtgacg gcgatctaga 1621 ttgtgatgat gcttcggacg agaaaaactg ccaaactaat gctccaagcg aagaagaata 1681
tctttctggg caagccgatc acatgcattc gtgctcagca gcaggaatgt attcttgtgg 1741 aacaaaagga tccgaaattg gcgtttgtat tccgatgaat gccacgtgta atgggatcaa 1801
ggagtgtcca ctaggagatg acgagtcaaa acattgctcc gaatgtgcca gaaagcgatg 1861 tgaccacaca tgtatgaaca ctccacacgg ggctcgctgc atttgtcaag aaggatataa 1921
gcttgccgat gacggactca cttgcgagga tgaagatgag tgtgcaactc atgggcactt 1981 gtgccagcat ttctgtgaag atcgtttggg ttcctttgca tgcaaatgtg ccaacggtta 2041
tgagcttgaa acggatgggc attcttgtaa atacgaggca accactacgc cagaaggata 2101 tttgttcatc agtcttggtg gagaagttcg acagatgcca ttggcagatt tcaccgatgg 2161
ttcaaattac tcggcgattc aaaagtttgc tggccacgga accatcagat cgatcgactt 2221 catgcatcgc aacaacaaaa tgttcatgtc aatttctgat gagcacggtg atccaactgg 2281
cgaattgtca gtgtccgaca atggattgat gagagttctt cgagaaaatg tcattggagt 2341 gagcaacgtg gcagtcgact ggattggtgg aaacgttttc ttcacacaaa aatgtatgtt 2401
tatctaatgt ttaaattttt catttgtgat tcttacagct ccatctccaa gcgctgggat 2461 ttccatctgc acaatgagcg gaatgttctg tcgccgagtt atcgaaggca aagaacaagg 2521
acaatcctat cgtggtcttg ttgttcaccc gatgcgcggt ctcatcatct ggatcgattc 2581 ttatcagaaa tatcatcgca tcatgatggc taatatggat gggtctcagg tgagtcgatc 2641
gagtcgatct gatttagttc atttctaaat aaatttcagg tcagaatcct tctcgacaac 2701 aagttggaag ttccatcagc tcttgccatc gactacatcc gccacgatgt ctattttgga 2761
gatgttgaac gtcagttgat cgaaagagtc aatatcgaca cgaaagagcg ccgcgtagtg 2821 atttcgaacg gagttcatca tccgtatgac atggcttact tcaatggttt cctatactgg 2881
gcagattggt aagacatctt atctaattta tattttcaaa tttatttttc aggggaagcg 2941 agtcattaaa ggttcaagag atgacccatc atcattcgag tcctcaagtc atccatactt 3001
tcaatcgtta tccatatggt attgctgtca atcactcact ctaccagact ggtcctccat 3061 caaacccatg ccttgaactc gagtgcccat ggctctgcgt tattgtgcca aagagcgatt 3121
tcattatgac tgccaagtgt gtctgcccag acggatacac tcattccgtc actgaaaact 3181 cttgcatccc gcctgtgacg attgaggacg aggagaacct tgagaagctt tcccacattg 3241
gatctgcttt gatggccgaa tactgcgaag ctggtgtcgc gtgtatgaat ggaggagcct 3301 gccgtgaact acaaaatgag cacggaagag ctcatcgcat cgtttgtgat tgtgagggtc 3361
catatgacgg gcaatactgc gaacggctca atccagagaa gttctccgca atggaagagg 3421 aagattcgtc cttatggctt atcgttctgc ttctcatttt tctcatcatc gttgcggtag 3481
tcggaattat tgccttcctt tggttttctc aacaagagca tatgaaagat gtgatttcca 3541 ctgcccgtgt ccgtgttgat aacatggcta gaaaagcgga agatgctgca gctccaattg 3601
tcgagaagtt ccgcaaggtc actgataagc agaggagcac gcctcctaga gaaggttgtc 3661 aaacggcaac aaacgttgac ttcgtttcct acgagacaaa tgctgagaaa agaattcgga 3721
tggactcttc gccgacgtca tacggaaacc ccatgtacga tgaagttcct gaatcgtcaa 3781 ctggtttcgt cagatcggct tccgcaccat tcgctggagt cattcgattt gagaacgaca 3841
gcttgttgtg aattctacta caaaattact aaatcagatg tctgtaaagt atatctattt 3901 ttgcctattt attgcatgaa agttgataat gtcta
```

Practical: *Gene*; *RNA*; *Protein*

U62639 (mRNA)

```
atgagaacca tgcgccttgc ttggttgctc ccacttttta ttcacatact aatcaagaac 61   acagctcaag ctccggctgt caacaactcg acatgcgatc aagcaaagga atttgattgc 121
gggaacggga gactccgatg cattcccgcg gagtggcaat gcgacaacgt agcggactgc 181  gacaaaggaa gagacgaatc gggctgctca tatgcgcatc attgttcgac aagcttcatg 241
ttatgcaaga atggactgtg tgtcgcaaat gagttcaaat gcgacggcga agacgactgc 301  cgcgatggaa gcgatgagca gcattgcgag tacaatatcc tgaagtctcg cttcgatggt 361
tccaatcctt cggctcctac cactttcgtt ggtcacaatg gcccagaatg ccatcctcct 421  cgtttacgat gccgatcagg acaatgtatt caaccagatc tcgtttgtga tggacatcag 481
gattgttctg gaggagatga tgaggtcaac tgcaccagaa ggggacatga aaatatgcag 541  tcctcgactg attttcacga tgatgttcat cttgtcgatc caacctttt cgctaatgaa 601
gacaataagt gtcggagtgg atacacaatg tgccatagcg gagacgtctg catacctgac 661  agttttcttt gtgacggcga tctagattgt gatgatgctt cggacgagaa aaactgccaa 721
actaatgctc caagcgaaga agaatatctt tctgggcaag ccgatcacat gcattcgtgc 781  tcagcagcag gaatgtattc ttgtggaaca aaaggatccg aaattggcgt ttgtattccg 841
atgaatgcca cgtgtaatgg gatcaaggag tgtccactag gagatgacga gtcaaaacat 901  tgctccgaat gtgccagaaa gcgatgtgac cacacatgta tgaacactcc acacggggct 961
cgctgcattt gtcaagaagg atataagctt gccgatgacg gactcacttg cgaggatgaa 1021 gatgagtgtg caactcatgg gcacttgtgc cagcatttct gtgaagatcg tttgggttcc 1081
tttgcatgca aatgtgccaa cggttatgag cttgaaacgg atgggcattc ttgtaaatac 1141 gaggcaacca ctacgccaga aggatatttg ttcatcagtc ttggtggaga agttcgacag 1201
atgccattgg cagatttcac cgatggttca aattactcgg cgattcaaaa gtttgctggc 1261 cacggaacca tcagatcgat cgacttcatg catcgcaaca acaaaatgtt catgtcaatt 1321
tctgatgagc acggtgatcc aactggcgaa ttgtcagtgt ccgacaatgg attgatgaga 1381 gttcttcgag aaaatgtcat tggagtgagc aacgtggcag tcgactggat tggtggaaac 1441
gttttcttca cacaaaaatc tccatctcca agcgctggga tttccatctg cacaatgagc 1501 ggaatgttct gtcgccgagt tatcgaaggc aaagaacaag gacaatccta tcgtggtctt 1561
gttgttcacc cgatgcgcgg tctcatcatc tggatcgatt cttatcagaa atatcatcgc 1621 atcatgatgg ctaatatgga tgggtctcag gtcagaatcc ttctcgacaa caagttggaa 1681
gttccatcag ctcttgccat cgactacatc cgccacgatg tctattttgg agatgttgaa 1741 cgtcagttga tcgaaagagt caatatcgac acgaaagagc gccgcgtagt gatttcgaac 1801
ggagttcatc atccgtatga catggcttac ttcaatggtt tcctatactg ggcagattgg 1861 ggaagcgagt cattaaaggt tcaagagatg acccatcatc attcgagtcc tcaagtcatc 1921
catactttca atcgttatcc atatggtatt gctgtcaatc actcactcta ccagactggt 1981 cctccatcaa acccatgcct tgaactcgag tgcccatggc tctgcgttat tgtgccaaag 2041
agcgatttca ttatgactgc caagtgtgtc tgcccagacg gatacactca ttccgtcact 2101 gaaaactctt gcatcccgcc tgtgacgatt gaggacgagg agaaccttga gaagctttcc 2161
cacattggat ctgctttgat ggccgaatac tgcgaagctg gtgtcgcgtg tatgaatgga 2221 ggagcctgcc gtgaactaca aaatgagcac ggaagagctc atcgcatcgt ttgtgattgt 2281
gagggtccat atgacgggca atactgcgaa cggctcaatc cagagaagtt ctccgcaatg 2341 gaagaggaag attcgtcctt atggcttatc gttctgcttc tcatttttct catcatcgtt 2401
gcggtagtcg gaattattgc cttcctttgg ttttctcaac aagagcatat gaaagatgtg 2461 atttccactg cccgtgtccg tgttgataac atggctagaa aagcggaaga tgctgcagct 2521
ccaattgtcg agaagttccg caaggtcact gataagcaga ggagcacgcc tcctagagaa 2581 ggttgtcaaa cggcaacaaa cgttgacttc gtttcctacg agacaaatgc tgagaaaaga 2641
attcggatgg actcttcgcc gacgtcatac ggaaacccca tgtacgatga agttcctgaa 2701 tcgtcaactg gtttcgtcag atcggcttcc gcaccattcg ctggagtcat tcgatttgag 2761
aacgacagct tgttgtga
```

Practical: *Gene*; *RNA*; *Protein*

AAD09364 (Protein)

```
  1 MRTMRLAWLL PLFIHILIKN TAQAPAVNNS TCDQAKEFDC GNGRLRCIPA EWQCDNVADC
 61 DKGRDESGCS YAHHCSTSFM LCKNGLCVAN EFKCDGEDDC RDGSDEQHCE YNILKSRFDG
121 SNPSAPTTFV GHNGPECHPP RLRCRSGQCI QPDLVCDGHQ DCSGGDDEVN CTRRGHENMQ
181 SSTDFHDDVH LVDPTFFANE DNKCRSGYTM CHSGDVCIPD SFLCDGDLDC DDASDEKNCQ
241 TNAPSEEEYL SGQADHMHSC SAAGMYSCGT KGSEIGVCIP MNATCNGIKE CPLGDDESKH
301 CSECARKRCD HTCMNTPHGA RCICQEGYKL ADDGLTCEDE DECATHGHLC QHFCEDRLGS
361 FACKCANGYE LETDGHSCKY EATTTPEGYL FISLGGEVRQ MPLADFTDGS NYSAIQKFAG
421 HGTIRSIDFM HRNNKMFMSI SDEHGDPTGE LSVSDNGLMR VLRENVIGVS NVAVDWIGGN
481 VFFTQKSPSP SAGISICTMS GMFCRRVIEG KEQGQSYRGL VVHPMRGLII WIDSYQKYHR
541 IMMANMDGSQ VRILLDNKLE VPSALAIDYI RHDVYFGDVE RQLIERVNID TKERRVVISN
601 GVHHPYDMAY FNGFLYWADW GSESLKVQEM THHHSSPQVI HTFNRYPYGI AVNHSLYQTG
661 PPSNPCLELE CPWLCVIVPK SDFIMTAKCV CPDGYTHSVT ENSCIPPVTI EDEENLEKLS
721 HIGSALMAEY CEAGVACMNG GACRELQNEH GRAHRIVCDC EGPYDGQYCE RLNPEKFSAM
781 EEEDSSLWLI VLLLIFLIIV AVVGIIAFLW FSQQEHMKDV ISTARVRVDN MARKAEDAAA
841 PIVEKFRKVT DKQRSTPPRE GCQTATNVDF VSYETNAEKR IRMDSSPTSY GNPMYDEVPE
901 SSTGFVRSAS APFAGVIRFE NDSLL
```

Practical: *Gene; RNA; Protein*

1. Download the sequences Gene, RNA and Protein
2. Upload to SeqWEB

ANALYSIS:

1. Exon/intron organization.
   Use (1) BESTFIT & GAP (“gene” vs “rna”)
   (2) Genome Blastn
2. Opening Reading Frame
   Use MAP to find the ORF
   Use TRANSLATE to write the ORF
   Compare your ORF with “protein”
3. Protein Domain Search (NCBI CD Search, Interpro)
4. Protein Sequence Analysis
   see next page

Protein Sequence Analysis in SeqWEB

DO all the REDS

HmmerPfam

Compares one or more sequences to a database of profile hidden Markov models, such as the Pfam library, in order to identify known domains within the sequences.

PeptideStructure

Makes secondary structure predictions for a peptide sequence.
These predictions include (in addition to alpha, beta, coil, and turn) measures for antigenicity, flexibility, hydrophobicity, and surface probability. The predictions are displayed graphically.

CoilScan

Locate coiled-coil segments in protein sequences. .

HTHScan

Locate helix-turn-helix motifs in protein sequences.

SPScan

Locate secretory signal peptides in protein sequences.

PeptideSort

Shows the peptide fragments from a digest of an amino acid sequence.
It sorts the peptides by weight, position, and HPLC relative retention, and shows the composition of each peptide. It also prints a summary of the composition of the whole protein.

PepPlot

Plots predicted protein secondary structure and hydropathy plot. .

Moment

Makes a contour plot of the helical hydrophobic moment of a peptide sequence.

HelicalWheel

Plots a peptide sequence as a helical wheel to help you recognize amphiphilic regions or beta sheets.

Isoelectric

Plots the charge as a function of pH for a peptide sequence.

TransMem

Scans for likely transmembrane helices in a peptide sequence.

OTHERS

Motifs

Looks for sequence motifs by searching through proteins for the patterns defined in the PROSITE Dictionary of Protein Sites and Patterns. Motifs ca display an abstract of the current literature on eac of the motifs it finds.