# Bioinformatics Applications in Proteomics

Chiquito Crasto
ccrasto@genetics.uab.edu
January 29, 2008

## GenBank

- http://www.ncbi.nlm.nih.gov/
  - Gene information
  - Protein Information
  - Literature
  - Links to Microarray Data, if any

BLAST

# SWISSPROT

- http://www.expasy.ch/
  - Prosite
  - SWISS-2DPage
  - SWISS-Model Repository (Homology Modeling)
  - Other Tools

# UNIPROT

- http://www.pir.uniprot.org/
  - **Interoperability** between
    - European Bioinformatics Institute (EBI)
    - EXPASY
    - Protein Information Resource (PIR)
    - Enzyme Nomenclature
      - http://www.chem.qmul.ac.uk/iubmb/enzyme/

## KEGG (Kyoto Encyclopedia of Genes and Genomes)

- Database
- Atlas
- Pathway
- BRITE
- Genes
- SSDB (Sequence Similarities DB)
- Ligand
- DBGet

## MASCOT—Protein Indentification from Mass Spectroscopy Data

- Peptide Mass Fingerprinting
- Sequence Query
- MS/MS Ion Search

- Alternatives
- Phenyx--http://phenyx.vital-it.ch/pwi/login/login.jsp
- Aldente--http://www.genebio.com/aldente/index.html

# Protein Data Bank-PDB

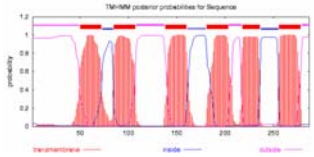
- http://www.rcsb.org/pdb/home/home.do
- Front Page for each entry
- File page and formats
- Molecular Viewer

# Bioinformatics as a Drive or Discovery

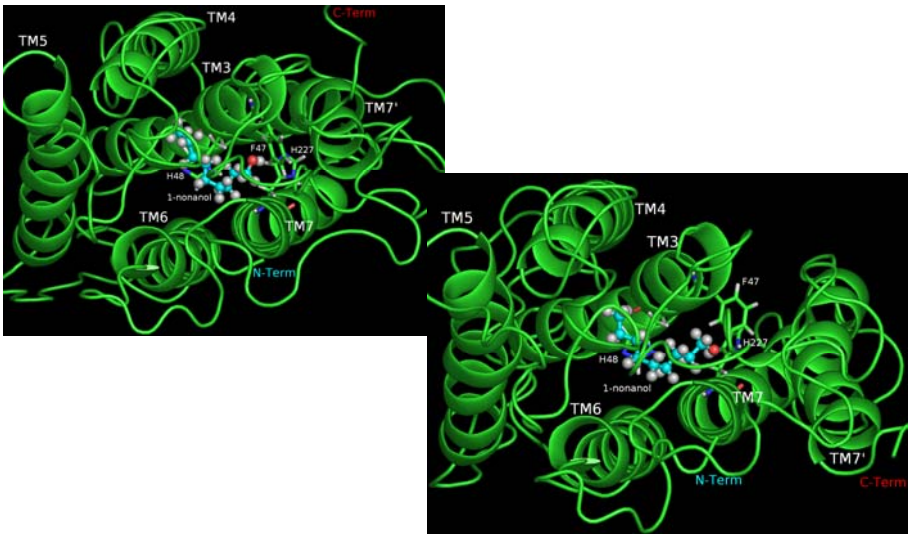
- **Recognition of the mechanism of olfaction depends upon understanding the sequence-structure-function relationships of olfactory receptors. We used bioinformatics methods to identify a new structural sub-class of olfactory receptors and GPCRs. We used available statistical methods to predict trans-membrane helical domains in olfactory receptor hOR17-210, a receptor that has been shown to be variably functional and pseudogenic in humans.**
- **TM domain identification was undertaken as a prelude to modeling this olfactory receptor in order to understand its interaction with ligands that have been experimentally shown to bind to this receptor. Our analyses revealed that there are only five typically observed TM regions in this protein with an additional orphan TM. The C-terminus is extra-cellular. This reversed polarity in the termini does not disrupt the positions of typical OR-motifs that initiate the signal transduction process at the membrane.**
- **Our observations are contrary to conventional structural knowledge about ORs and GPCRs. Preliminary sequence analysis studies have shown that such a structure is observed in a limited number of olfactory receptors distributed across different mammalian species.**
- **We also present a methodology for modeling such olfactory receptors. When combined with experimental data, we believe that this information will further our understanding of olfaction.**

- **Sequence Features for OR17-210**
- **This protein sequence for olfactory receptor OR17-210 appears as a pseudogene in the HORDE1 database**:
  - ATGATGAAGAAGAACCAAACCATGATCTCAGAGTTCCTGCTCCTGGGCCTTCCATCCAACCTGAGCAGCAGAATCTGTTCTATGCCT TGTTCTTGGCCGTGTATCTTACCACCCTCCTGGGGAACCTCCTCGTCATTGTCCTCATTCGACTGGACTCCCACCTCCACATGCCTA TGTATTTGTGTCTCAGCAACTTGTCCTTCTCTGACCTCTGCTTTTCCTCGGTCACAATGCCCAAATTGCTGCAGAACATGCAGAGCCA AAACCCATCCATCCCCTTTGCGGACTGCCTGGCTCAGATGTACTTTCATCTGTTTTATGGAGTTCTGGAGAGCTTCCTCCTTGTGGT CATGGCTTATCACTGCTATGTGGCTATTTGCTTTCCTCTGCACTACACCACTATCATGAGCCCCAAGTGTTGCCTTGGTCTGCTGACA CTCTCCTGGCTGTTGACCACTGCCCATGCCACGTTGCACACCTTGCTTATGGCCAGGCTGTCCTTTTGTGCTGAGAATGTGATTCCT CACTTTTTCTGTGATACATCTACCTTGTTGAAGCTGGCCTGCTCCAACACGCAAGTCAATGGGTGGGTGATGTTTTTCATGGGCGGG CTCATCCTTGTCATCCCATTCCTACTCCTCATCATGTCCTGTGCAAGAATCGTCTCCACCATCCTCAGGGTCCCTTCCACTGGGGGC ATCCAGAAGGCTTTCTCCACCTGTGGCCCCCACCTCTCTGTGGTGTCTCTCTTCTATGGGACAATTATTGGTCTCTACTTGTGCCCA TTGACGAATCATAACACTGTGAAGGACACTGTCATGGCTGTGATGTACACTGGGGTGACCCACATGCTGAACCCCTTCATCTACAGC CTGAGGAACAGAGACATGAGGGGGAACCCTGGGCAGAGTCTTCAGCACAAAGAAAATTTTTTTTGTCTTTAAAATAGTAATAGTTGGC ATTTTACCGTTATTGAAT
- **Intuitively Translated as**:
  - MMKKNQTMISEFLLLGL/PIQPEQQNLFYALFLAVYLTTLLGNLLVIVLIRLDSHLHMPMYLCLSNLSFSDLCFSSVTMPKLLQNMQSQNPSIP FADCLAQMYFHLFYGVLESFLLVVMAYHCYVAICFPLHYTTIMSPKCCLGLLTLSWLLTTAHATLHTLLMARLSFCAENVIPHFFCDTSTLLK LACSNTQVNGWVMFFMGGLILVIPFLLLIMSCARIVSTILRVPSTGGIQKAFSTCGPHLSVVSLFYGTIIGLYLCPLTNHNTVKDTVMAVMYT GVTHMLNPFIYSLRNRDMRGNPGQSLQHKENFFVFKIVIVGILPLLN
- **A two nucleotide frame shift however results in a functional protein with the following sequence**:
  - MPMYLCLSNLSFSDLCFSSVTMPKLLQNMQSQNPSIPFADCLAQMYFHLFYGVLESFLLVVMAYHCYVAICFPLHYTTIMSPKCCLGLLTL SWLLTTAHATLHTLLMARLSFCAENVIPHFFCDTSTLLKLACSNTQVNGWVMFFMGGLILVIPFLLLIMSCARIVSTILRVPSTGGIQKAFSTC GPHLSVVSLFYGTIIGLYLCPLTNHNTVKDTVMAVMYTGVTHMLNPFIYSLRNRDMRGNPGQSLQHKENFFVFKIVIVGILPLLN

- **OR17-210 is an Atypical Olfactory Receptor**
- **OR17-210 begins with MPMY---. This sequence PMY is strongly conserved in most ORs. This sequence typically marks the beginning of the second transmembrane region. Hidden Markov Models2 have predicted that in OR17-210, this region is not a TM3. Furthermore, an HA-epitope tag experiment revealed this region of the protein to be extra-cellular.**




- **What is typically helix 3 in ORs is helix 1 in OR17-210. This TM has the MAYD(E)RY motif, which marks the intracellular side and (part of intracellular loop 2) of TM3 The directionality of this TM1 is extracellular to intracellular. This correctly positions the DRY region of the TM intracellulary—where structural changes following activation may be necessary for signal transduction in GPCRs4**
- **This allows only five typically observed in TMs in OR17-210. HMM strongly predicts that the cDNA sequence has an additional TM helix in the long C-terminus following what would be the seventh TM in most OR sequences. We call this the 7' TM. OR17-210 has a homolog in chimpanzee with greater than 95% sequence similarity. A BLAST search of the 7' sequence, "FVFKI VIVGILPLLN LVGVVKLI" does not return any matches in other ORs, GPCRs or any other protein sequence in GENBANK.**
- **TM 7' can then occupy either the position of the missing TM1 or TM2 in order to maintain the TM scaffold and protect the ligand and the binding pocket from the surrounding lipid layer**
- **If one follows the progression of N-terminus-TM1-IC1-TM2-EC1-TM3 .. etc, the C-terminus of this receptor is extra-cellular**




## Frame shifts, stop and start codons

- http://ca.expasy.org/tools/dna.html
- http://bioportal.weizmann.ac.il/HORDE/search_horde.html
- http://bip.weizmann.ac.il/cgi-bin/HORDE/showGene.pl?key=symbol&value=OR1E3P
- TMHMM-- http://www.cbs.dtu.dk/services/TMHMM/

## Problems during Protein Identification

- No sequence in database --- nothing to correlate with
- Problems with entries in database: human errors in entering information (typographical errors and curation); sequencing errors; errors during transcription
- Modifications in large proteins: degradation, oxidation of methionine, deamidation of N and Q, remember glycosylations, phosphorylations, and acetylations ….

  http://www.unimod.org/ lists the possible modifications that can occur

## Sources for Protein Identification

- EBI, European Bioinformatics Institute, EBI Download site "The EBI is a centre for research and services in bioinformatics. The Institute manages databases of biological data including nucleic acid, protein sequences and macromolecular structures."
  - Expasy "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 and Software Packages.
  - INFOBIOGEN, DBCAT, The Public Catalog of Databases
  - NCBI, The National Center for Biotechnology Information "provides an integrated approach to the use of gene and protein sequence information" Databases and Tools
  - PIR Protein Information Resource "An integrated public resource of protein informatics to support genomic and proteomic research and scientific discovery." Located at Georgetown University.
  - Plant Genome Database (PlantGDB) Resource for Plant Comparative Genomics
  -

# Sources contd….

- SIB, Swiss Institute of Bioinformatics "The SIB is an academic not-for-profit foundation established on March 30, 1998 whose mission is to promote research, the development of databanks and computer technologies, teaching and service activities in the field of bioinformatics, in Switzerland with international collaborations"
  - RESID Database at the EBI "The RESID Database of Protein Modifications is a comprehensive collection of annotations and structures for protein modifications including amino-terminal, carboxyl-terminal and peptide chain cross-link post-translational modifications." (quote form the RESID site)
  - UNIMOD "Protein Modifications For Mass Spectrometry", A list of potential amino acid modifications and mass shifts.
  - UniProt The Universal Protein Resource " is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR."
  - What is FASTA format? Protein ID programs use sequence databases (flat files) that are formatted in FASTA format. Other Formats

Tools you should consider using throughout the course:
Source: http://ca.expasy.org/tools/
Protein Identification and Characterization

- AACompIdent - Identify a protein by its amino acid composition
- AACompSim - Compare the amino acid composition of a UniProtKB/Swiss-Prot entry with all other entries
- Aldente - Identify proteins with peptide mass fingerprinting data. A new, fast and powerful tool that takes advantage of Hough transformation for spectra recalibration and outlier exclusion
- FindMod - Predict potential protein post-translational modifications and potential single amino acid substitutions in peptides. Experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified Swiss-Prot entry or from a user-entered sequence, and mass differences are used to better characterize the protein of interest.
- FindPept - Identify peptides that result from unspecific cleavage of proteins from their experimental masses, taking into account artefactual chemical modifications, post-translational modifications (PTM) and protease autolytic cleavage
- MultiIdent - Identify proteins with *pI*, *Mw*, amino acid composition, sequence tag and peptide mass fingerprinting data
- PeptideCutter - Predicts potential protease and cleavage sites and sites cleaved by chemicals in a given protein sequence
- PeptideMass - Calculate masses of peptides and their post-translational modifications for a UniProtKB/Swiss-Prot or UniProtKB/TrEMBL entry or for a user sequence
- TagIdent - Identify proteins with *pI*, *Mw* and sequence tag, or generate a list of proteins close to a given *pI* and *Mw*

# Tools …

- Phenyx - Protein and peptide identification/characterization from PMF and MS/MS data from GeneBio, Switzerland
- IsotopIdent - Predicts the theoretical isotopic distribution of a peptide, protein, polynucleotide or chemical compound
- Mascot - Peptide mass fingerprint, sequence query and MS/MS ion search from Matrix Science Ltd., London
- OMSSA - MS/MS peptide spectra identification by searching libraries of known protein sequences
- PepMAPPER - Peptide mass fingerprinting tool from UMIST, UK
- PepSea - Protein identification by peptide mapping or peptide sequencing from Protana, Denmark
- PFMUTS - Shows the possible single and double mutations of a peptide fragment from MALDI peptide mass fingerprinting
- ProteinProspector - A variety of tools from UCSF (MS-Fit, MS-Tag, MS-Digest, etc.) for mining sequence databases in conjunction with mass spectrometry experiments [Mirrors at UCL-Ludwig, UK / Ludwig Institute Melbourne (Australia)]
- PROWL - Protein chemistry and mass spectrometry resource from Rockefeller and NY Universities [or from Genomic Solutions]
- SearchXLinks - Analysis of mass spectra of modified, cross-linked, and digested proteins whose amino acid sequence is known, from Caesar, Germany

# Tools …
# DNA -> Protein

- Translate - Translates a nucleotide sequence to a protein sequence
- Transeq - Nucleotide to protein translation from the EMBOSS package
- Graphical Codon Usage Analyser - Displays the codon bias in a graphical manner
- BCM search launcher - Six frame translation of nucleotide sequence(s)
- Backtranslation - Translates a protein sequence back to a nucleotide sequence
- Reverse Translate - Translates a protein sequence back to a nucleotide sequence
- Genewise - Compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors
- FSED - Frameshift error detection
- LabOnWeb - Elongation, expression profiles and sequence analysis of ESTs using Compugen LEADS clusters

## Tools …
## Similarity Searches

- BLAST and WU-BLAST - Interfaces to various versions of the **B**asic **L**ocal **A**lignment **S**earch **T**ool
- BLAST  Network Service on ExPASy
- BLAST at EMBnet-CH/SIB (Switzerland)
- BLAST at NCBI
- WU-BLAST at Bork's group in EMBL (Heidelberg)
- WU-BLAST and BLAST at the EBI (Hinxton)
- BLAST at PBIL (Lyon)
- **Bic** ultra-fast rigorous (Smith/Waterman) similarity searches using the Bioccelerator [At DKFZ or at Weizmann]
- DeCypher - Smith/Waterman or FrameSearch search using the DeCypher hardware accelerator
- Fasta3 - FASTA version 3 at the EBI
- FDF - Smith/Waterman type searches on Paracel's Fast Data Finder (FDF) at EMBnet-CH
- MPsrch - Smith/Waterman sequence comparison at EBI
- PropSearch - Structural homolog search using a 'properties' approach at Montpellier
- SAMBA - Systolic Accelerator for Molecular Biological Applications
- SAWTED - Structure Assignment With Text Description
- Scanps - Similarity searches using Barton's algorithm
- SEQUEROME - BLAST similarity search and sequence profiling at Georgetown University
- SHOPS - Analysis of the genomic operon context for any group of proteins

## Tools …
## Pattern and Profile Searches

- InterPro Scan - Integrated search in PROSITE, Pfam, PRINTS and other family and domain databases
- ScanProsite  - Scans a sequence against PROSITE or a pattern against the UniProt Knowledgebase (Swiss-Prot and TrEMBL)
- MotifScan - Scans a sequence against protein profile databases (including PROSITE)
- **Pfam HMM search**; scans a sequence against the Pfam protein families db [At Washington University or at Sanger Centre]
- FingerPRINTScan - Scans a protein sequence against the PRINTS Protein Fingerprint Database
- FPAT - Regular expression searches in protein databases
- ELM - Eukaryotic Linear Motif resource for functional sites in proteins
- **PRATT** - Interactively generates conserved patterns from a series of unaligned proteins; [at EBI / ExPASy  ]
- PPSEARCH - Scans a sequence against PROSITE (allows a graphical output); at EBI
- PROSITE scan - Scans a sequence against PROSITE (allows mismatches); at PBIL
- PATTINPROT - Scans a protein sequence or a protein database for one or several pattern(s); at PBIL
- SMART - Simple Modular Architecture Research Tool; at EMBL
- TEIRESIAS - Generate patterns from a collection of unaligned protein or DNA sequences; at IBM
- Hits - Relationships between protein sequences and motifs

## Tools …
## Topology Prediction

- PSORT - Prediction of protein subcellular localization
- TargetP - Prediction of subcellular location
- DAS - Prediction of transmembrane regions in prokaryotes using the Dense Alignment Surface method (Stockholm University)
- HMMTOP - Prediction of transmembrane helices and topology of proteins (Hungarian Academy of Sciences)
- PredictProtein - Prediction of transmembrane helix location and topology (Columbia University)
- SOSUI - Prediction of transmembrane regions (TUAT; Tokyo Univ. of Agriculture & Technology)
- TMAP - Transmembrane detection based on multiple sequence alignment (Karolinska Institut; Sweden)
- TMHMM - Prediction of transmembrane helices in proteins (CBS; Denmark)
- TMpred - Prediction of transmembrane regions and protein orientation (EMBnet-CH)
- TopPred - Topology prediction of membrane proteins (France)

## Tools … Primary Structure Analysis

- ProtParam - Physico-chemical parameters of a protein sequence (amino-acid and atomic compositions, *pI*, extinction coefficient, etc.)
- Compute pI/Mw - Compute the theoretical *pI* and *Mw* from a UniProt Knowledgebase entry or for a user sequence
- ScanSite pI/Mw - Compute the theoretical *pI* and *Mw*, and multiple phosphorylation states
- MW, pI, Titration curve - Computes *pI*, composition and allows to see a titration curve
- Radar - De novo repeat detection in protein sequences
- REP - Searches a protein sequence for repeats
- REPRO - De novo repeat detection in protein sequences
- TRUST - De novo repeat detection in protein sequences
- SAPS - Statistical analysis of protein sequences at EMBnet-CH [Also available at EBI]
- Coils - Prediction of coiled coil regions in proteins (Lupas's method) at EMBnet-CH [Also available at PBIL]
- Paircoil - Prediction of coiled coil regions in proteins (Berger's method)
- Multicoil - Prediction of two- and three-stranded coiled coils
- 2ZIP - Prediction of Leucine Zippers
- PESTfind - Identification of PEST regions at EMBnet Austria
- HLA_Bind - Prediction of MHC type I (HLA) peptide binding
- PEPVAC - Prediction of supertypic MHC binders
- RANKPEP - Prediction of peptide MHC binding
- SYFPEITHI - Prediction of MHC type I and II peptide binding
- ProtScale - Amino acid scale representation (Hydrophobicity, other conformational parameters, etc.)
- Drawhca - Draw an HCA (Hydrophobic Cluster Analysis) plot of a protein sequence
- Protein Colourer - Tool for coloring your amino acid sequence
- Three To One - Tool to convert a three-letter coded amino acid sequence to single letter code
- Colorseq - Tool to highlight (in red) a selected set of residues in a protein sequence
- HelixWheel / HelixDraw - Representations of a protein fragment as a helical wheel
- RandSeq - Random protein sequence generator

## Tools … Secondary Structure Preduitction

- AGADIR - An algorithm to predict the helical content of peptides
- APSSP - Advanced Protein Secondary Structure Prediction Server
- GOR - Garnier et al, 1996
- HNN - Hierarchical Neural Network method (Guermeur, 1997)
- Jpred - A consensus method for protein secondary structure prediction at University of Dundee
- JUFO - Protein secondary structure prediction from sequence (neural network)
- nnPredict - University of California at San Francisco (UCSF)
- Porter - University College Dublin
- PredictProtein - PHDsec, PHDacc, PHDhtm, PHDtopology, PHDthreader, MaxHom, EvalSec from Columbia University
- Prof - Cascaded Multiple Classifiers for Secondary Structure Prediction
- PSA - BioMolecular Engineering Research Center (BMERC) / Boston
- PSIpred - Various protein structure prediction methods at Brunel University
- SOPMA - Geourjon and Del�age, 1995
- SSpro - Secondary structure prediction using bidirectional recurrent neural networks at University of California
- DLP - Domain linker prediction at RIKEN

## Tools … Tertiary Structure Prediction

- iMolTalk - An Interactive Protein Structure Analysis Server
- MolTalk - A computational environment for structural bioinformatics
- Seq2Struct - A web resource for the identification of sequence-structure links
- STRAP - A structural alignment program for proteins
- TLSMD - TLS (Translation/Libration/Screw) Motion Determination

Tools … Computational Methods

- SWISS-MODEL - An automated knowledge-based protein modelling server
- 3Djigsaw - Three-dimensional models for proteins based on homologues of known structure
- CPHmodels - Automated neural-network based protein modelling server
- ESyPred3D - Automated homology modeling program using neural networks
- Geno3d - Automatic modelling of protein three-dimensional structure
- SDSC1 - Protein Structure Homology Modeling Server

# Tools …

- Threading
- 3D-PSSM - Protein fold recognition using 1D and 3D sequence profiles coupled with secondary structure information (Foldfit)
- Fugue - Sequence-structure homology recognition
- HHpred - Protein homology detection and structure prediction by HMM-HMM comparison
- Libellula - Neural network approach to evaluate fold recognition results
- LOOPP - Sequence to sequence, sequence to structure, and structure to structure alignment
- SAM-T02 - HMM-based Protein Structure Prediction
- Threader - Protein fold recognition
- ProSup - Protein structure superimposition
- SWEET - Constructing 3D models of saccharides from their sequences
- *Ab initio*
- HMMSTR/Rosetta - Prediction of protein structure from sequence

# Tools … (Graphics)

- Swiss-PdbViewer - A program to display, analyse and superimpose protein 3D structures
- Astex Viewer
- MolMol
- Rasmol
- VMD
- YASARA - Molecular graphics, modeling, simulations and eLearning
- Prediction of disordered regions
- DisEMBL - Protein disorder prediction
- GlobPlot - Protein disorder/order/globularity/domain predictor

## Tools … (Sequence Alignment)

- Binary
- SIM + LALNVIEW - Alignment of two protein sequences with SIM, results can be viewed with LALNVIEW
- LALIGN - Finds multiple matching subsegments in two sequences
- Dotlet - A Java applet for sequence comparisons using the dot matrix method
- Multiple
- Decrease redundancy - Reduce a set of sequences into a non-redundant set
- **CLUSTALW** [At EBI, PBIL or at EMBnet-CH]
- MAFFT - Multiple sequence alignment at GenomeNet
- Muscle - Multiple protein sequence alignment at Berkeley
- **T-Coffee** [At EMBnet Switzerland or at BioASP]
- MSA - at Genestream (IGH)
- DIALIGN - Multiple sequence alignment based on segment-to-segment comparison, at University of Bielefeld, Germany
- Match-Box - at University of Namur, Belgium - at Washington University
- **Multalin** [At INRA or at PBIL]
- MUSCA - Multiple sequence alignment using pattern discovery, at IBM
- Alignment analysis
- AMAS - Analyse Multiply Aligned Sequences
- Bork's alignment tools - Various tools to enhance the results of multiple alignments (including consensus building).
- CINEMA - Color Interactive Editor for multiple alignments
- ESPript - Tool to print a multiple alignment
- SVA - Sequence Variability Analyser for multiple alignments
- WebLogo - Sequence logos at Berkeley/USA
- plogo - Sequence logos at CBS/Denmark
- GENIO/logo - Sequence logos at Stuttgart/Germany
- SeqLogo - Sequence logos at MIF/USA
- WebLogo - Sequence logos at Cambridge/UK

# Tools … (Text and Data Mining)

- AcroMed - A computer generated database of biomedical acronyms and the associated long forms extracted from the recent Medline abstracts
- MedMiner - Extract and organize relevant sentences in the literature based on a gene, gene-gene or gene-drug query
- Protein Annotator's Assistant - A software system which assists protein annotators in the task of assigning functions to newly sequenced proteins
- XplorMed - Explore a set of abstracts derived from a bibliographic search in MEDLINE