Introduction to Bioinformatics

EBI-EMBL Training Course Overview

March 10th & 11th, 2005

The UK Bioinformatics Forum together with the European Bioinformatics Institute (EBI) and Oxford University Continuing Professional Development Center (OUDCE) organized the *Introduction to Bioinformatics* course held in Oxford on the 10th and 11th of March 2005. The course was well attended by academics from Oxford University and Oxford Brookes University, and representatives from industry.

The first day of the course was taught by Lisa Mullan, Scientific Training Officer at the EBI, and introduced the essential bioinformatics tools and methods available through the EBI-EMBL website. The EBI offers around 200 databases and the course introduced a text-based method and the Sequence Retrieval System (SRS), similar to the Enterez from NCBI, as a query tool to access this data. The databases can also be searched using fast searching algorithms such as BLAST, if only a query nucleotide or protein sequence is available. In addition, comparative analysis of sequences was also looked at using pairwise and multiple sequence alignment algorithms.

The second day introduced Ensembl, a collaborative project between the Sanger Institute and the EBI to allow free access to genomic information. Ensemble automatically curates genomic data, and uniquely has its own gene discovery pipeline, which uses Genscan and GeneWise gene prediction programs, and can be browsed via Ensmart, a genome data-mining tool. This session was taught by Michael Schuster, who currently works on the Ensembl helpdesk. The second half of the day introduced the EBI Macromolecular Structure Database (MSD). The MSD is the European initiative for the management and distribution of data on macromolecular structure by integrating current protein and structure databases and works in collaboration with the Research Collaboratory for Structural Bioinformatics (RCSB) who maintain the PDB in the States. New structures submitted by the AutoDep tool (http://autodep.ebi.ac.uk/autodep-basepage.shtml) are manually curated by the EBI MSD group and then passed on to the RCSB.

Databases

Genome:

EMBL is Europe's primary resource for nucleotide sequence database and is produced in collaboration with GenBank (USA) and DDBJ (Japan). New submissions are made independently to each of the three databases and all new or modified entries are exchanged on a daily basis between each of these databases to ensure a fully comprehensive database search. EMBL provides access to both completed phase 3 HTG (nr) or partially completed genomic sequence data through its Genomes Pages (www.ebi.ac.uk/genomes/). There are currently about 18,324,138 entries in the latest build. These entries are referenced with a unique reference number ([aA-zZ]5digits) or ([aA-zZ] [aA-zZ]6digits) to identify the type of entry. For example:

NCxxxxxx → Completed prokaryotic genome or eukaryotic chromosome
NGxxxxxx → Homo Sapiens Genomic Region
NMxxxxxx → mRNA of multiple genomes
RefSeq → Reference Sequence

Ensembl (www.ensembl.org) is a joint venture between the Wellcome Trust Sanger Institute and the EBI and provides access to non-vertebrate and vertebrate genomes. Ensembl uses assemblies of sequences generated from the original consortium responsible for sequencing the genome. In the case of humans for instance, Ensembl collaborates with the American NCBI and UCSC consortiums to access the human assemblies and then the Genome Browsers at Ensembl, NCBI & UCSC allow us to access the similar shared data. Ensembl does not do manual curation of genomic sequences, which can be very labor intensive; instead uses database comparisons to annotate its genomic entries and gene predictions (using GenScan or GenWise). The Genome Browser **Ensmart** (http://www.ensembl.org/Multi/martview) at Ensembl allows us the ability to do complex queries and allows users to start at a chromosome level and zoom down to the nucleotide level while accessing relevant information such as SNPs etc. One can also view pre-annotated sequences in Ensembl's pre-build site (http://pre.ensembl.org/) that provides early information on new genome assemblies currently in the process of being "build".

Also available at the EBI is **Genome Reviews** (www.ebi.ac.uk/GenomeReviews/), which contains a comprehensive collected of sequenced prokaryotic genomes. The data in Genome Reviews is cross-referenced to other databases providing coding sequences, domain information and protein processing and function and is updated every two weeks. The data in Genome Reviews is also incorporated into Integr8 (www.ebi.ac.uk/integr8) which is an integrated database offering comprehensive statistical analyses of data in GenomeReviews and UniProt proteome. It provides information on species descriptions, summary information on complete proteome and integrates data from InterPro, ClusTr and GO and information can be queries using the BioMart/Ensmart query interface.

Transcriptomes

EBI hosts **ArrayExpress** (www.ebi.ac.uk/arrayexpress), which is the database containing information from microarray experiments. The data contained within corresponds with the standardizations set by MIAME (Minimum Information About a Microarray Experiment) devised by the MGED (Microarray Gene Expression Data Society) (www.mged.org). The data submitted in this database comes from a relatively small number of high throughput transcriptomics laboratories. Many journals require their authors of micro-array data-based papers to submit their data to a MAIME-compliant database. Other MAIME compliant databases include Gene Expression Omnibus (**GEO**) (http://www.ncbi.nlm.nih.gov/geo/) and Center for Information Biology gene Expression database (**CIBEX**) (http://cibex.nig.ac.jp/index.jsp)

Proteomes

Protein sequence databases contain translated sequences from EMBL as well as from Protein Identification Resource (PIR) and those extracted from literature or directly submitted by researchers. The annotation is of high quality and the data is extensively cross-referenced to other databases. Two primary resources exist to make protein information publicly available.

1. Manually curated **Swiss-Prot** (http://us.expasy.org/sprot/), an EBI & Swiss Inst of Bioinformatics collaboration:
   Contains ~ 115105 entries curated manually and highly integrated with other databases. There is a high level of annotation including function, domain information, post-translational modifications etc. Entry names consist of the name of the gene followed by the species. Accession numbers are of the following format:
   [O, P, Q] [0-9] [A-Z, 0-9] [A-Z, 0-9] [A-Z, 0-9] [0-9] → such as P26367 (PAX6_HUMAN)

And it's complementary/supplementary database **TrEMBL** (translated EMBL sequences) where curation is done computationally allowing for a much larger database. TrEMBL entries are manually annotated before being entered into SwissProt. Currently containing ~ 632013 entriesSwissProt TrEMBL (SpTrEMBL) contains entries, which will eventually be integrated into the SwissProt database and SwissProt accession numbers have been assignedRemaining TrEMBL (RemTrEMBL) contains entries that will never be incorporated into SwissProt. These would include immunoglobulins; T-cell receptors; small fragments; synthetic sequences; CDS not coding for real proteins; patent application sequences.

2. American Protein Identification Resource (**PIR**) (http://pir.georgetown.edu/home.shtml), which is the world's first repository of fully, annotated proteins. The PIR is a computer system offering both peptide and nucleotide sequences designed to aid protein identification and although most of the PIR sequences have been incorporated into the SwissProt, there might still be a few rogue sequences. There are approximately 283175 entries in the PIR.The RefSeqP database **RefSeqP** provides a protein reference standard. It is used, as is RefSeq, to provide a foundation for the functional annotation of the human genome. The current

release contains 402006 entries and the Accession numbers for all proteins are of the format: NP_123456

The latest attempts to provide a combined comprehensive single global resource for protein information has led to the creation of **UniProt** (www.uniprot.org), which combines data from Swiss-Prot, TrEMBL & PIR. UniProt consists of the *UniProt Archive* & *UniParc* containing open access to non-redundant protein sequence database providing relevant cross-references to the original sequence source and annotation. The *UniProt Knowledgbase* contains the amino acid sequence, protein description, taxonomy and citation information and information on protein function, post-translational modifications, functional domains, active sites, subunit structure, subcellular localization, and disease associated mutations and variant sequence information.

The **UniNRef N**on-redundant **REF**erence database (http://www.ebi.ac.uk/uniref/) aims to facilitate sequence merging in UniProt and allows for faster and more informative sequence similarity searches. UniNRef clusters (like Unigene clusters) consist of closely related sequences based on sequence identity cutoffs. UniRef90 and UniRef50 databases consists of sequences where no pair of sequences have greater than 90% or 50% sequence identity respectively. The purpose of clustering is to reduce a long list of similar or identical alignments when doing a search, which might not allow us to see novel matches in the output. This also allows for faster database searching. The UniRef100 database presents identical sequences and sub-fragments as a single entry with protein IDs, sequences, bibliography, and links to protein databases.

Integr8 (www.ebi.ac.uk/integr8), the integrated database from EBI, allows us to do a more comprehensive search on a particular proteome, which provides more coverage then the individual proteomics database. Integr8 provides InterPro access for each proteome. **InterPro** (http://www.ebi.ac.uk/interpro/) is a database of protein families, domains and functional sites and is used to study features in unknown proteins by comparing with known protein information presented in this database. It allows:

- Comparisons with other proteomes
- Generation of hierarchical clusters of proteins based on sequences in order to study orthologes, paralogs and singletons
- Functional classifications using the Gene Ontology (GO) (www.geneontology.org) and
- Links to secondary (HSSP) (http://www.cmbi.kun.nl/gv/hssp/) and tertiary (PDB) structures.
- Ability to download full proteome sets
- Able to use EnsMart to extract information from several databases in a single query

Structures

World Wide PDB (**wwPDB**) (http://www.wwpdb.org/index.html) is a collaboration of MSD-EBI, RCSB and PDBj in order to manage and maintain a single freely available Protein Data Bank of macromolecular structural data. Data into this archive is deposited from all three organizations while each maintains its own view of the data contained

within with their individual tools. The archives are maintained and hosted by the RCSB and a PDB 4-letter code [0-9][aA-zZ, 0-9] [aA-zZ, 0-9] [aA-zZ, 0-9] is assigned to each submission.

The **MSD** (http://www.ebi.ac.uk/msd/index.html) is a member of the wwPDB and is a relational database (standardized format of tables that can be queries using SQL etc.). It holds a collection of 3D coordinates of each atom in a protein, allowing the structure to be displayed by viewing software such as Protein Explorer, Rasmol, Gromacs, MolScript, Astexviewer etc. Protein structures are submitted by individual researchers and have been determined by x-ray diffraction, NMR or 3D Electron Microscopy. Structures can be submitted to the MSD via the AutoDep tool (http://www.ebi.ac.uk/msd-srv/autodep4after which they are manually curated before being sent to the RCSB for central deposition. The MSD curation process involves authentication of source and structure and validation of the methodology used. The structures are checked for errors and consistencies to PDB standards before being sent to the RCSB. Regio and Stereoisomers are differentiated as well.

The MSD database consists of a deposition database (normalized) with thousands of relationships linking ~ 400 tables and a simpler query database (denormalized) with duplicated items and aggregated into 40 wider tables making it easier to searching and data retrieval. They allow ftp downloads as well as various Application Program Interface (API) access using Perl etc. The MSD has a very powerful query interface catered for the novice to expert users of the structure database. *MSDbar* caters to novices or new users and can be used simply to search for data in MSD, RCSB, PDBj & OCA. *MSDlite* is designed for more intermediate users who wish to conduct a more refined search while *MSDPro* uses a powerful & unique drag and drop java servlet to allow users to build complex linked queries graphically instead of a text based approach as in Ensmart. MSD also contains a protein quaternary structure database (**PQS**) (http://www.ebi.ac.uk/msd-srv/pqs) which uses information derived from the PDB entries by applying crystal symmetry matches to protein structures. Upon transferring data from the deposition database to the search database, additional information is built into the entries, such as characterization of ligand binding sites, derivation of secondary structure as well as cross-referencing to other structural databases such as **SCOP** (http://scop.mrc-lmb.cam.ac.uk/scop/) which contains protein folds and family information and **CATH** (http://www.biochem.ucl.ac.uk/bsm/cath_new/) which contains protein classification, architecture, topology and homology information.
A list of specific search utilities in MSD include:

- *MSDlite* – web form application
- *MSDpro* – java applet
- *MSDchem* – complete collection of chemical species and small molecules in the PDB
- *Emsearch* – Electron Microscopy search tool
- *MSDfold* – Secondary structure matching tool for protein structure comparison
- *MSDsite* – active site database search
- *MSDtarget* – tools for searching and tracking structural genomic targets
- *Biobar* – Mozilla and Netscape toolbar application for searching

Protein Interactions & Pathways

Protein-Protein interactions are cataloged by the Human Proteome Organization (**HUPO**) (http://211.32.65.137/) which works in collaboration with the EBI. **IntAct** (www.ebi.ac.uk/intact/) is a HUPO developed central repository for storing and accessing protein-protein interactions containing both experimental and curated literature data. IntAct allows us to search proteins of interest interacting with other proteins, graphically display interaction networks, analyze interaction networks using GO terms, visualize minimal connecting networks for protein sets, download data in PSI-MI format. IntAct is a member of the International Molecular Interaction Exchange (**IMEx**) consortium and includes other protein interaction sites such as The Biomolecular Interaction Network Database (**BIND**) (http://bind.ca/index.jsp?pg=0), Database of Interacting Proteins (**DIP**) (http://dip.doe-mbi.ucla.edu/), Molecular Interactions Database (**MINT**) (http://mint.bio.uniroma2.it/mint/), and Munich Information Centre for Protein Sequences (**MIPS**). The EBI also curates another database containing Chemical Entities of Biological Interest (**ChEBI**) (http://www.ebi.ac.uk/chebi/). The entries within constitute atoms, molecules, ions, ion-pair, radicals etc, each identified by a separate distinguishable identifier conforming to the IUPAC and NC-IUBMB standards. This excludes genomic molecules. The data in ChEBI was obtained from **IntEnz** (Integrated relational Enzyme Database) (http://www.ebi.ac.uk/intenz/index.html) at the EBI, **KEGG LIGAND** (Kyoto Encyclopaedia of Genes and Genomes) composite database.

The record for molecular interaction pathways is kept at **Reactome** (www.reactome.org), a collaboration between CSHL, EBI and GO consortium. Reactome catalogs records of human and other vertebrate metabolic regulatory pathways. Each pathway is represented as a series of events and sub-events with defined inputs and outputs and cross-linked to UniProt, Ensembl and LocusLink.

Sequence Alignments & Search

The open source bioinformatics suite **EMBOSS** (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/) contains over 200 applications for bioinformatics analysis and can be either utilized directly via the EBI website using a web-based GUI or installed independently on user machines. We can launch these EMBOSS applications via the EBI-SRS tools tab (http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz).

There are three ways of observing sequence similarity between two or more sequences. The first **segment method** of **Dotplots**, such the *DottupN* application allows us to do a simple dotplot search between two identical sequences (eg. cDNA against a gDNA entry). If, however, we know that our sequences are not exactly alike but still quite similar, the second algorithm *Dotmatcher* utilizes a "sliding windows" approach where two sequences are compared using a threshold score value to reflect the degree of similarity of sequences required. The default dotplots matrices are **EDNAFULL** for nucleotides and **EBLOSUM62** for proteins.

EDNAFULL

	A	T	G	C	S	W	R	Y	K	M	B	V	H	D	N	U
A	5	4	4	4	4	1	1	4	4	1	4	1	1	1	2	4
T	4	5	4	4	4	1	4	1	1	4	1	4	1	1	2	5
G	4	4	5	4	1	4	1	4	1	4	1	1	4	1	2	4
C	4	4	4	5	1	4	4	1	4	1	1	1	1	4	2	4
S	4	4	1	1	1	4	2	2	2	2	1	1	3	3	1	4
W	1	1	4	4	4	1	2	2	2	2	3	3	1	1	1	1
R	1	4	1	4	2	2	1	4	2	2	3	1	3	1	1	4
Y	4	1	4	1	2	2	4	1	2	2	1	3	1	3	1	1
K	4	1	1	4	2	2	2	2	1	4	1	3	3	1	1	1
M	1	4	4	1	2	2	2	2	4	1	3	1	1	3	1	4
B	4	1	1	1	1	3	3	1	1	3	1	2	2	2	1	1
V	1	4	1	1	1	3	1	3	3	1	2	1	2	2	1	4
H	1	1	1	4	3	1	1	3	1	3	2	2	2	1	1	1
D	1	1	1	4	3	1	1	3	1	3	2	2	2	1	1	1
N	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	2
U	4	5	4	4	4	1	4	1	1	4	1	4	1	1	2	5

	A	T	G	C
A	5	-4	-4	-4
T	-4	5	-4	-4
G	-4	-4	5	-4
C	-4	-4	-4	5

We can see from the above matrix that if a window size of 4 was selected, the maximal score for perfect alignment would be 5 x 4 = 20. Likewise, if the window size was selected at 10, the maximal score would be 50. If the threshold value was 10, then a dot

would be placed wherever there was an exact match for sequence with a total score greater than or equal to 10. Similarly, if the threshold was set at 23 (default) this would allow 3 mismatches out of ten nucleotides (we lose 3 x 5 points =15 and gain 3 x (- 4) points = -12 which means 50 –15 = 35 + (-12) = 23), we would get a much darker dotplot as we would expect a lot of random hits. Increasing the window size would eliminate the problems of getting non-specific hits, but would come at the cost of losing out on information about the presence of small repeats or inverted repeats (which might be useful for stem-loop analysis). Likewise, keeping the threshold value equal to the maximal score for a selected window size would give the most specific match (Raising the threshold value over the maximal score might actually reduce the number of matches returned). A dotplot will provide information on the presence of insertions or deletions shown by a shift in the diagonal, and we could find internal repeats, repeated domains or regions of low complexity in a sequence. However, it will not give any more detailed information on the actual sequence or residue and more importantly will not incorporate gaps to create an "optimal alignment" and is thus limited in its use.

The second method of **optimal global alignment** allows us to compare two sequences over their whole length allowing for the best overall score for the comparison of two sequences to be obtained. The EBI global alignment tool can be accessed via this link (http://www.ebi.ac.uk/emboss/align/index.html). The best scores are based on maximizing regions of similarity and minimizing gaps using a given scoring matrix and gap penalty and extension input. The scoring matrix is designed to cope with mutational data and is designed to reflect similarity between the different amino acids rather than just identity. The *Needleman Wunsch* algorithm uses the PAM or BLOSUM matrices for global alignments. It should be noted however, that the results obtained could be heavily skewed by the parameters chosen for gap opening and extension penalty. It is possible for the computer to actually sacrifice a proper alignment a long way down the sequence string because of a high cost of gap opening or extension penalty, in favor of the minor costs of an imperfect alignment that would give a higher overall score. This would be very relevant if we were searching for the exons of a gene in genomic DNA. An exon separated by a very long intron (>5-6 kb) would not be found by the program because here the gap extension penalty would extend up to 6kb and that would prove to be very costly to the overall score of the alignment (at a gap penalty of 0.5), so the computer favors doing a local alignment where the score is far better. In order to make sure that we do not miss out on this, we would have to adjust the gap penalty to 0.1. While, increasing the gap penalty might give us more stringent search, it needs to be carefully selected in a global alignment situation where it might be necessary to have long gaps between relevant sequences.

The third method of **optimal local alignment** algorithm is very similar to the global alignment algorithm, except it allows for local similarity searches between two sequences so the alignment may be over a short sequence span. This would be useful if there are a lot of repeats in the sequence or when comparing proteins with multiple repeated domains. The EMBOSS program **water** is the local alignment tool based on the *Smith Waterman* algorithm. Both the global and local alignment algorithms work on a similar principle of **dynamic programming,** however, with the local alignment, the alignment is discarded once the score reaches below zero. We will generally find the overall identity

and similarity scores to go up in a local alignment. The overall scores are, therefore, not a very good judge of predicting the quality of the alignment and are dependent on the length & span of the query sequence within the subject sequence. A dotplot will generally give us an idea on the existence of multiple domains within a query sequence; however, water is only designed to give a single alignment with the best score. If our sequence contains more than one match, we can use the program **MatcherN**, also available through the SRS interface where we can specify the "number of alternative matches" (alignments). The cut-off threshold for an alignment is higher than the water program; therefore the alignments will be much cleaner. In addition, EMBOSS contains another pairwise alignment program **Stretcher** which is a global alignment tool, however, is less rigorous than needle and, thus takes up less computational time and is useful for database searching. Another program **Supermatcher** is designed for local alignment for very large sequences and is even less rigorous. **Est2Genome** is also very useful for aligning cDNA sequences to a genomic sequence and contains information on splice sites and coding regions. It aligns ESTs, cDNAs or mRNAs to an unspliced genomic DNA sequence and will insert intronic gaps of arbitrary length and ensure that these introns start and stops occurs at the splice consensus dinucleotides GT (start) and AG (end). The program first aligns both strands of the spliced sequence against the forward strand of the gDNA, assuming splice consensus GT/AG in the forward direction and then the maximum scoring orientation is re-aligned in the reverse strand assuming the CT/AC splice consensus. It outputs the maximally scoring alignment along with a list of introns and exons.

A list of all EMBOSS applications can be found at:
www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/

Scoring Matrices

The simplest method of comparing two sequences is to use a dotplot using a pairwise scoring matrix. This identity matrix would simply match one sequence against another and assign a score without taking into account gap penalties. A simple identity matrix is the EDNAFULL matrix discussed previously. This matrix would also cover comparisons for nucleotide ambiguity codes. However, a simple alignment is not necessarily the optimal alignment and can be misleading in terms of the evolutionary divergence between the two sequences. In order to determine protein homology, we need to align protein sequences such that we can get information on conserved regions or residues, which will assist in deciphering protein structure and function. A scoring scheme would give us an "optimal alignment" and measure any evolutionary change between the two sequences. The most widely used matrices for protein alignment are PAM and BLOSUM. Both of these matrices calculate substitution frequencies between amino acids and are derived from known protein alignments.

PAM is the unit of Point Accepted Mutations and is a measure of the evolutionary distance between the two sequences. A distance of 1 PAM would indicate an amount of evolution producing an average of one mutation per 100 amino acids. Therefore, a higher number on PAM matrix would represent a larger evolutionary distance.

Dayhoff PAM 250 Matrix

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

AMINO ACIDS	%
A	8.3
C	1.7
D	5.3
E	6.2
F	3.9
G	7.2
H	2.2
I	5.2
K	5.7
L	9.0
M	2.4
N	4.4
P	5.1
Q	4.0
R	5.7
S	6.9
T	5.8
V	6.6
W	1.3
Y	3.2

Therefore, in order to compare sequences with short evolutionary distance between them, especially for short and local alignments, we might chose PAM40, whereas PAM250 would be better suited to compare protein orthologs in different species. A high score in the matrix, for instance, Phe → Trp: 7, implies where this particular mutation is conserved. The substitution of Ala → Ala has a low score of 2 because of the high frequency of the residue Ala (8.3%) that occurs in nature, thus its increased likelihood of being present due to chance.

The BLOSUM (BLOcks Substituted Matrix) scoring matrix is also derived from an aligned family of proteins and is the default matrix for FASTA sequence analysis. Instead of comparing single residues, the Blosum matrix compares regions of high conservation in a selected family of aligned proteins and these regions are stored in the BLOCKS database (http://blocks.fhcrc.org/). The Blosum matrix is built by eliminating sequences

that are identical by more than a certain percentage (x%). This eliminates bias in favor of a certain protein. Thus a BLOSUM50 matrix is built of sequences that are no more than 50% identical. Likewise, a BLOSUM 62 matrix is built of sequences that are no more than 62% identical, but because these sequences are more similar, a higher BLOSUM matrix will allow us to compare sequences that are not too far apart in their evolutionary divergence, which is the opposite of the PAM matrices.

BLOSUM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

GAP PENALTIES

In addition, to taking into account evolutionary relationship between different amino acid residues or sequence clusters, scoring matrices also have to take into account the cost of opening and extending gap penalties. Gap penalties are necessary to ensure that the optimal alignment is reached for a given sequence pair. Generally, for two sequences as below:

GATCA &
GACTATC

The minimal alignment length L is 5:

GATCA--
GACTATC

The maximal alignment length is 12:

GATCA- - - - - - -
- - - - - GACTATC

Therefore, we could have 12!/(7!(5!)) = 792 different possible alignments which would take up too much computational time. Addition of gap penalties would assign a cost of opening up gaps and extending those gaps and thus eliminate all, but the optimally scoring alignment based on the selected gap parameters.

Matrices and their gap penalities

Matrix	Gap Opening Penalty	Gap Extension Penalty
BLOSUM45	15	2
BLOSUM62	11	1
BLOSUM80	10	1
PAM30	9	1
PAM70	10	1

There are two types of gap penalties:

Linear gap penalty (treats each opened gap as a novel gap):
$F(g) = -g\,d$
Where g is the gap and d is the gap opening score

Affine gap penalty (charges an opening gap penalty score followed by a cost associated with every succeeding extending gap):
$F(g) = -d\,(-e(g-1))$
Where $-d$ is the gap opening score and $-e\,(g-1)$ is the score associated with extending successive gaps

DYNAMIC PROGRAMMING

As mentioned earlier, both the local and global alignment algorithms work on the principle of dynamic programming using the PAM and BLOSUM matrices. This algorithm is designed to significantly reduce computational time from an $O(2^N)$ algorithm to an $O(N^2)$ algorithm.

Dynamic programming aims to find the best scoring alignment by calculating the scores of all the possible alignment in a matrix of two sequences and then tracing back the matrix to find the highest scoring alignment.

The first step involves creating a matrix with x+1 columns and y+1 rows where x and y are the sizes of two sequences to be aligned.

Sequence 1: $x_1x_2\ldots x_n$
Sequence 2: $y_1y_2\ldots y_n$
Score F where,

F(i, j) = score of optimal path of subsequences $x_1 \ldots x_i$ and $y_1 \ldots y_i$

Assuming a linear gap penalty for a global alignment:

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(xi,yj) & (s = \text{match/mismatch score}) \\ F(i-1, j) - d & (\text{gap in y}) \\ F(i, j-1) - d & (\text{gap in x}) \end{cases}$$

With F(0,0) = 0, F(i,0) = -id, F(0.j) = -jd

Once the matrix has been filled using the above approach, the traceback step determines the actual alignment that would result in the maximal alignment score.

Sequence Based Data Mining

BLAST

BLAST (Basic Local Alignment Search Tool) is a database search tool developed and maintained by the NCBI (http://www.ncbi.nlm.nih.gov/BLAST/). BLAST uses a heuristic algorithm that compares a query and subject sequence in a given database and finds the highest scoring locally optimal alignment between the two sequences. BLAST works by allowing gaps in the alignments that it creates which means that the final alignment score is a more accurate reflection of the biological relationship between the two sequences.

BLAST uses the local alignment algorithm discussed above and creates all sequence matches above a given cut-off. BLAST scans a database for words (usually 3-mers for proteins and at least 7 for nucleotides) that score at least T (threshold value) when aligned with a word in the query sequences. This word alignment is called a hit. If another non-overlapping hit is found within a distance (A) of the first hit and on the same diagonal, then the first hit is extended between the database and query on both directions. This extension continues until the running score drops below the maximum score seen so far by a value of X, thus giving us a high scoring local alignment called an HSP (high-scoring segment pair) or MSP (maximum scoring segment pair). If the alignment score of the HSP exceeds a given value Sg (the gapped score), then a gapped extension of the HSP is initiated.

There are 5 search programs available through the BLAST family of programs and allow all combinations of DNA and protein query sequences to be searched against DNA or protein databases:

Program	Query Sequence	Query Database	Function	Utility
BlastN	DNA	DNA	Compares nucleotide sequence against a nucleotide database	Finds DNA sequences to match a query
BlastP	Protein	Protein	Compares an amino acid query against a protein database	To find homologous proteins

BlastX	DNA	Protein	Compares a nucleotide query translated in all reading frames against a protein database	To find protein sequence encoded by the query sequence
TBlastN	Protein	DNA	Compares protein query sequence against a nucleotide sequence dynamically translating in all reading frames	Allows us to find genes in unknown DNA sequences
TBlastX	DNA	DNA	Compares the six-frame translations of a nucleotide query against the six-frame translations of a nucleotide sequence database	To find degree of homology between the coding region of the query sequence against the known genes in the database

A list of few databases available for BLAST searches at NCBI	
nr	Non-redundant protein and nucleotide databases of all sequences (phase 3 finished sequences with or without annotation) excluding EST, STS, GSS and Phase 0, 1 or 2 HTG sequences
est	Expressed Sequence Tags - available for humans only (est_human), mouse only (est_mouse), all non-human & non-mouse ESTs (est_others)
gss	Genomic survey sequences (includes single-pass genomic data, exon trapped sequences and Alu PCR sequences
htgs	Unfinished high throughput genomic sequences in phase 0 (1-2 pass reads of a single clone), 1 (unfinished ordered or unordered contigs with gaps) and 2 (unfinished, ordered and oriented contigs with or without gaps)
pat	Patented GenBank Protein sequences
yeast	S. *cerevisiae* genome and protein sequences
mito	Mitochondrial sequence database
month	All new or revised nt or protein sequences added to nr in the last 30 days
pdb	Sequences obtained from the Brookhaven Protein Data Bank 3D protein structures
dbsts	Sequence Tagged Sites, roughly 200-500 bp length, unique in a genome and used to define specific positions on a physical map; from GenBank, EMBL or DDBJ
ecoli	E. *coli* genomic CDS translated sequences
drosophila	Drosophila protein sequences provided by Celera and Berkley Drosophila Genome Project (BDGP)

The statistical significance of the BLAST results is judged by the E value. It describes the random background noise that exists for matches between the sequences by describing the number of distinct alignments possible with score equivalent to or better than the one of interest, that are possible entirely by chance. A smaller E value indicates a more significant score. Generally, an E value of less than 0.0001 would indicate that the two sequences compared are homologous to each other and is thus a good way to identify sequence orthologs.

Generally, before initiating a query, a program is run on the query sequence to identify regions of low complexity and is marked off as NNNN (for nucleotide) or XXXX (for protein sequences). This prevents artefactual hits that could return a high score, which is not a true reflection of the similarity between the two sequences. This filtering also masks out repeated regions, which the program identifies by searching the query sequence for repeats as compared with a database of human, and rodent repeat sequences. The NCBI blast usually filters low complexity regions by default, whereas the EBI BLAST (http://www.ebi.ac.uk/Tools/similarity.html) does not have filtering turned on by default. There are also two different versions of BLAST algorithm. The NCBI blast was co-developed by Warren Gish at the NCBI who later moved to Washington University and developed the algorithm further to allow gapped BLAST searches as well as allowing additional functionalities for the advanced command line users. Another feature of the BLAST family is PSI-BLAST (http://www.ncbi.nlm.nih.gov/blast/index.htm) that allows us to search for weakly related homologs to a query protein sequence against the protein database. Position specific iterative (PSI) BLAST works by constructing a profile or a position specific scoring matrix (PSSM) from a multiple alignment of the highest scoring hits returned from an initial BLAST search of the query sequence. This PSSM would give a high score to a highly conserved residue in the initial alignment while weakly conserved residues would be assigned a score closer to zero. This profile is then used to do another BLAST search and the results of this second iteration are then used to further refine the PSSM giving it more sensitivity than straight BLAST. However, it should be noted that if the query sequence contains a strongly conserved domain then the profiles generated would be weighted towards this domain and away from the rest of the sequence as further iterations are performed. This would potentially miss weakly homologous areas. There is also the danger that unrelated sequences might occasionally be included during further iterations and if allowed to go unnoticed, further iterations might end up preferentially selecting for this sequence resulting in an entirely different end product. Inclusion of low complexity sequences such as Gln, Ser, Thr, and Pro-rich regions would result in the inclusion of unrelated sequences containing similarly regions of low complexity.

Reference

Protein Matrices

BLOSUM

1. Henikoff S, Henikoff JG., (1992) Amino acid substitution matrices from protein blocks *Proc Natl Acad Sci.* 89(22):10915-9

PAM

2. Dayhoff, M., Schwartz, R.M., and Orcutt, B.C. (1978) A model of evolutionary change in proteins, in *Atlas of Protein Sequence and Structure,*Vol. 5, pp. 345-352, National Biomedical Research Foundation, Silver Spring, MD

Dynamic Programming: Needleman-Wunsch Algorithm

3. Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.*48, 443-453

Global and Local Alignments

4. Smith, F.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*147, 195-197

Database Searches

FASTA

5. Lipman, D.J. and Pearson, W.R. (1985) Rapid and sensitive protein similarity search. *Science*227, 1435-1441

BLAST

6. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) A basic local alignment search tool. *J. Mol. Biol.*215, 403-410

7. Altschul S. F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs *Nucleic Acids Research,* 25(17) 3389–3402

Multiple Alignments

CLUSTAL Algorithm

8. Higgins, D.G. and Sharp, P.M. (1988) CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene*73, 237-244

9. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucl. Acids Res.*22, 4673-4680

10. Valdar W., (2002) Scoring Residue Conservation *PROTEINS: Structure, Function, and Genetics* 48:227–241

Dynamic Programming & Big-O Notation

11. Per Kraulis, *Stockholm Bioinformatics Center, Molecular Bioinformatics 2001 notes,* http://www.sbc.su.se/~per/molbioinfo2001/dynprog/dynamic.html

12. Rahul Simha, Department of Computer Science, George Washington University, *Algorithms and Data Structures II* http://www.seas.gwu.edu/~simhaweb/cs151/lectures/module12/align.html

13. Marina Alexandersson, *Fraunhofer-Chalmers Research Centre for Industrial Mathematics* http://www.fcc.chalmers.se/~marina/files/BioI_DynProg_2003.pdf

14. UC Berkeley Lecture Notes on Big-O Notation - http://www.me.berkeley.edu/~e77/lecnotes/ch20/ch20.htm

Others

15. Amaro R. et al (2004) Sequence Alignment Algorithms (www.ks.uiuc.edu/Training/Tutorials/) *University of Illinois at Urbana-Champaign Luthey-Schulten Group, Theoretical and Computation Biophysics Group*

16. Brooksbank, C., Cameron, G., Thornton, J., (2005) The European Bioinformatics Institute's data resources: towards systems biology *Nucl. Acids Res*. 33: D46-53