

Supplementary List of Software for Bioinformatics and Comparative Genomics

We provide brief details and a link to bioinformatics software used on the 2009 summer school, plus some other useful bioinformatics software. Licensing varies, but certainly, all software below is freely available for non-profit academic use. For authorship, advice on citation, and details of the license, please see information on the software's own Web site.

A lot of the software requires a Unix-like operating system. Often you will be able to access a server running Unix or Linux at your university or institute. However, you can also easily get yourself a friendly Linux distribution from the internet, and install it as a second OS, or install it on a spare computer: have a look at **Ubuntu** @ <http://www.ubuntu.com>. Or, if you want to run Linux programs on your Windows PC, try installing **Cygwin** @ <http://www.cygwin.com>. Cygwin provides a "Bash shell" under windows, and emulates a Linux environment on your MS Windows computer. A "shell" is a command language interpreter, which is simply a macro processor that executes commands. The default Linux shell is called the "Bash" shell. If using Cygwin, make sure to do a full installation, so that you obtain a compiler and other necessary tools. Another option is to install "**Bio-Linux**" as an operating system on a computer, or to even run it as a "live" image. The advantage of Bio-Linux is that it comes with most common bioinformatics software pre-installed. Check out the web page at <http://nebc.nox.ac.uk/tools/bio-linux/bio-linux-5.0> on how to use Bio-Linux as an installation or live image on a USB dongle. For Apple computers, the usual operating system, **Mac OS X**, is already Unix. You can start a command-line by launching Macintosh HD/Applications/Utilities/Terminal. Some relevant software (X11, Developer Tools) is not installed by default but is included on the OS X install DVD as optional packages.

If you do not install your own Linux OS, you will most likely need a terminal emulator and SSH client for connecting to a server. PuTTY is a good solution:

<http://www.chiark.greenend.org.uk/~sgtatham/putty/>

All Linux and Unix systems, including Cygwin and Mac OS X, will have **Perl** pre-installed. You can also get Perl for Windows separately: <http://www.perl.org/>. Whatever you do, make sure that you **do not** install both Cygwin with Perl and Perl for MS Windows on the same machine; the programs will conflict.

1. General Tools

1.1. **EMBOSS**. “The European Molecular Biology Open Software Suite”. EMBOSS is a software analysis package specially developed for the needs of the molecular biology user community. It contains many programs that can deal with most routine bioinformatics tasks.

Available at: <http://emboss.sourceforge.net/>

2. Sequence Alignment and Genome Assembly

2.1. **SSAHA2** (Sequence Search and Alignment by Hashing Algorithm) is a pairwise sequence alignment program designed for the efficient mapping of sequencing reads onto genomic reference sequences. Reads of most sequencing platforms (ABI-Sanger, Roche 454, Illumina-Solexa) and a range of output formats (SAM, CIGAR, PSL etc.) are supported. A pile-up pipeline for analysis and genotype calling is available as a separate package.

Available at: <http://www.sanger.ac.uk/Software/analysis/SSAHA2/>

2.2. **Staden Package**: A fully developed set of DNA sequence assembly (Gap5), editing and analysis tools (Spin) for Unix, Linux, MacOSX and MS Windows. This was used for the visualisations of assemblies on the course.

Available at: <http://staden.sourceforge.net/> See the Gap5 link bottom left hand side of ‘Downloads’.

2.3. **BLAST**: The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Available at: <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/> OR http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download

databases: <ftp://ftp.ncbi.nlm.nih.gov/blast/db/>

2.4. **MAFFT** is a multiple sequence alignment program for unix-like operating systems. It offers a range of multiple alignment methods, L-INS-i (accurate; for alignment of <~200 sequences), FFT-NS-2 (fast; for alignment of <~10,000 sequences), *etc.*

Available at: <http://align.bmr.kyushu-u.ac.jp/mafft/software/>

Supplementary material <http://www.genome-bioinformatics.org>

2.5. **Jalview** is a multiple alignment editor written in Java. It is used widely in a variety of web pages (e.g. the EBI Clustalw server and the Pfam protein domain database) but is available as a general purpose alignment editor. Jalview may be used for viewing, editing, analysis, annotation and publishing.

Available at: <http://www.jalview.org/>

3. Hidden Markov Models

3.1. **HMMER**: Profile hidden Markov models (profile HMMs) can be used to do sensitive database searching using statistical descriptions of a sequence family's consensus. HMMER is a freely distributable implementation of profile HMM software for protein sequence analysis.

Available at: <http://hmmer.janelia.org/>

3.2. **GeneWise**, which predicts gene structure using similar protein sequences, and **Genomewise**, which provides a gene structure final parse across cDNA- and EST-defined spliced structure. Both algorithms are heavily used by the Ensembl annotation system. The GeneWise algorithm was developed from a principled combination of hidden Markov models (HMMs). Both algorithms are highly accurate and can provide both accurate and complete gene structures when used with the correct evidence.

Available at: http://www.ebi.ac.uk/Tools/Wise2/doc_wise2.html which links to <ftp://ftp.ebi.ac.uk/pub/software/unix/wise2/>

3.3. **SNAP**. SNAP is a general purpose gene finding program suitable for both eukaryotic and prokaryotic genomes. SNAP is an acronym for Semi-HMM-based Nucleic Acid Parser.

Available at: <http://homepage.mac.com/iankorf/>

4. Comparative Genomics

4.1. **Artemis**: a free genome viewer and annotation tool that allows visualization of sequence features and the results of analyses within the context of the sequence, and its six-frame translation. Artemis is written in Java, and is available for UNIX, Macintosh and Windows systems. It can read EMBL and GENBANK database entries or sequence in FASTA or raw format. Extra sequence features can be in EMBL, GENBANK or GFF format.

Available at: <http://www.sanger.ac.uk/Software/Artemis/>

Supplementary material <http://www.genome-bioinformatics.org>

4.2.**ACT** (Artemis Comparison Tool) is a DNA sequence comparison viewer based on Artemis. In common with Artemis, ACT is written in Java and runs on UNIX, GNU/Linux, Macintosh and MS Windows systems. It can read complete EMBL and GENBANK entries or sequence in FASTA or raw format. Extra sequence features can be in EMBL, GENBANK or GFF format. The sequence comparison displayed by ACT is usually the result of running a blastn or tblastx search.

Available at: <http://www.sanger.ac.uk/Software/ACT/>

4.3.**MUMmer**: a system for rapidly aligning entire genomes

Available at: <http://mummer.sourceforge.net/>

5. Phylogeny

5.1.**Phylogeny Programs**. A list of phylogeny programs, compiled by Joe Felsenstein. It is an attempt to be completely comprehensive.

Available at: <http://evolution.genetics.washington.edu/phylip/software.html>

5.2.**Modelgenerator** is a model selection program that selects optimal amino acid and nucleotide substitution models from Fasta or Phylip alignments. ModelGenerator supports 56 nucleotide and 96 amino acid substitution models.

Available at: <http://bioinf.may.ie/software/modelgenerator/>

5.3.**PHYML**. Fast, accurate estimation of large PHYlogenies by Maximum Likelihood.

Available at: <http://www.atgc-montpellier.fr/phyml/>

5.4.**Dendroscope**. An interactive viewer for large phylogenetic trees and networks.

Available at: <http://www-ab.informatik.uni-tuebingen.de/software/dendroscope>

5.5.**PAML**. Phylogenetic Analysis by Maximum Likelihood. A package of programs for phylogenetic analyses of DNA or protein sequences using maximum likelihood, including: Comparison and tests of phylogenetic trees; estimation of parameters in sophisticated substitution models; likelihood ratio tests of hypotheses through comparison of implemented models; estimation of divergence times under global and local clock models; likelihood (Empirical Bayes) reconstruction of ancestral sequences using nucleotide, amino acid and codon models; generation of datasets of nucleotide, codon, and amino acid sequence by Monte Carlo simulation; estimation of synonymous and nonsynonymous substitution rates and detection of positive selection in protein-coding DNA sequences; and Bayesian estimation of species divergence times incorporating uncertainties in fossil calibrations.

Available at: <http://abacus.gene.ucl.ac.uk/software/paml.html>

Supplementary material <http://www.genome-bioinformatics.org>

5.6.**Notung**. A software tool that offers a unified framework for incorporating duplication/loss parsimony into phylogenetic tasks, including: **reconciling** a gene tree with a species tree and estimating upper and lower bounds on the time of duplication; **rooting** an unrooted gene tree by minimizing duplication and loss events; **rearranging** a rooted gene tree in areas of weak sequence support to minimize the number of duplications and losses; and **resolving** a non-binary gene tree by fitting it to a binary species tree.

Available at: <http://www.cs.cmu.edu/~durand/Notung/>