BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Current trends
Volume 9(13)

Workflow management systems for gene sequence analysis and evolutionary studies – A Review

Anu Sharma*, Anil Rai & SB Lal

Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, Library Avenue, Pusa, New Delhi - 110012; Anu Sharma – Email: anu@iasri.res.in; *Corresponding author

Received February 19, 2013; Accepted February 23, 2013; Published July 17, 2013

Abstract:
Post 'omic' era has resulted in the development of many primary, secondary and derived databases. Many analytical and visualization bioinformatics tools have been developed to manage and analyze the data available through large sequencing projects. Availability of heterogeneous databases and tools make it difficult for researchers to access information from varied sources and run different bioinformatics tools to get desired analysis done. Building integrated bioinformatics platforms is one of the most challenging tasks that bioinformatics community is facing. Integration of various databases, tools and algorithm is a challenging problem to deal with. This article describes the bioinformatics analysis workflow management systems that are developed in the area of gene sequence analysis and phylogeny. This article will be useful for biotechnologists, molecular biologists, computer scientists and statisticians engaged in computational biology and bioinformatics research.

Keywords: Analysis, bioinformatics, databases, phylogeny, integration, workflows.

Background:

MODERN biology is driven by large scale processing of heterogeneous data, which may come from diverse sources, such as sequences from GenBank, EMBL, PDB, DDJB, PROSITE, NGS and many other secondary databases. The interface which allows access to these different data sources vary widely. Therefore, in order to access these resources a researcher needs to be an expert in very different areas of computer science such as databases, networking, scripting languages etc. Furthermore, algorithms/tools used to extract biologically relevant information tend to be developed at faster pace by researchers but in isolation. There is hardly any code sharing among the data analysis algorithms however there is an increase in code complexity.

Gene sequence analysis and study of evolutionary relationships among organisms are two major areas of interest to biologists. Gene sequence analysis involves identification of stretches of sequence in DNA that are biologically functional whereas evolutionary studies infer biological relationships among different organisms. *In-silico* identification of coding regions in genomes and phylogeny studies are important problems that have been brought into focus through advances in genomic sequencing. Availability of diversified tools available on different platforms, structures and heterogeneous databases makes this analysis, a difficult task for biologists. So, there is an urgent need for development of solutions for integration of various tools and database to assist biologist from burden on executing them independently on different platforms.

Workflow Management System (WMS) is the integration of several bioinformatics tools with multiple databases, to automate the analysis and storage of genomic sequences. Several WMSs were developed for researchers to perform computational analysis with ease using various computational tools. These workflow systems, differs in scope and approach of integration for their execution. Many of these WMS are available as web based servers to provide access to powerful computing resources through familiar graphical-based environment for inexperienced users. This saves time for installing software on their own computer and analyzing biological data. Standalone workflow systems integrate various bioinformatics tools within desktop applications using graphically specified workflows. This also provides access to distributed computational resources to biologists.

Although, many WMS are developed in the area of gene sequence analysis and evolutionary studies but no attempt has been made to compile these at one place along with bioinformatics tools used at each stage of analysis. The objective of this article is to provide comprehensive information on available WMS along with their limitations and practical considerations on their usage to biotechnologists, molecular biologists and other researchers. It also provides review of various tools to computational scientists, who are actively involved in the development of WMS.

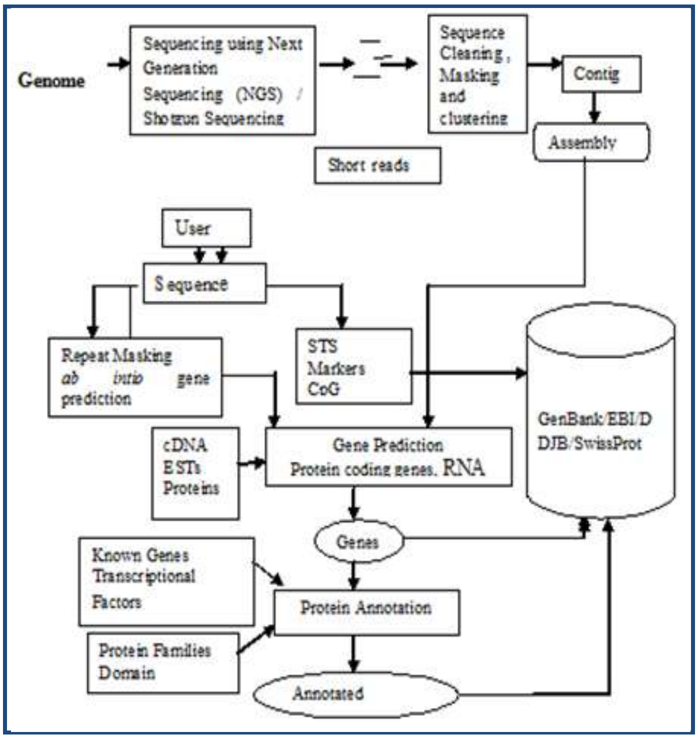
Genome
Sequencing using Next Generation Sequencing (NGS) / Shotgun Sequencing
Sequence Cleaning, Masking and clustering
Contig
Assembly
Short reads
User
Sequence
Repeat Masking ab initio gene prediction
STS Markers CoG
GenBank/EBI/DDJB/SwissProt
cDNA ESTs Proteins
Gene Prediction Protein coding genes. RNA
Genes
Known Genes Transcriptional Factors
Protein Annotation
Protein Families Domain
Annotated

Figure 1: General gene sequence analysis workflow system

Workflow Management Systems for Gene Sequence Analysis:

Gene sequence analysis involves identification of features such as genes, transcription initiation and poly(A) cleavage sites, 5′ as well as 3′-untranslated regions (UTRs) and promoter regions etc. in genome, derived through, transformation of raw genomic sequences into information by integrating computational tools, auxiliary biological data, and biological knowledge. Identification and *in-silico* annotation of coding and non-coding sequences from a variety of genomes is necessary due to exponential increase in raw sequence data. Due to availability of advanced sequencing technologies, large volume of multi-species genomic data is generated. Manual curation and annotation of this data is a difficult and time consuming task. The development of automatic *in-silico* computational solution to aid the manual curation process is highly desirable. As, genome annotation involves performing various tasks like gene finding, repeat finding, Expressed Sequenced Tags (EST)/cDNA alignment, homology searching and protein family searching etc. Attempts have been made to develop various biological workflows through integration of various computational tools through development of automatic pipelines to perform this genomic annotation. The generic solution of this workflow is given in **(Figure 1). Table 1 (see supplementary material)** lists some of the important computational tools used for performing different tasks in the process. Major workflows for gene sequence analysis along with important tools which are integrated are compared in **Table 2 (see supplementary material).**

ESTpass **[1]** is a workflow, used for processing and annotating sequence data from ESTs. The major advantages of ESTpass are, the integration of cleansing and annotating processes, rigorous chimeric EST detection, exhaustive annotation, email reporting to inform users about progress and to send results. PSEUDOPIPE **[2]** is a homology-based computational pipeline, which helps to search a mammalian genome and identify pseudogene sequences in a comprehensive and consistent manner. The output of PSEUDOPIPE is the complete annotation of pseudogenes in genome, their chromosomal location, nucleotide sequences, name and sequence of the parent gene, and alignment of the pseudogene with the functional gene. Tiger Gene Indices Clustering tools (TGICL) **[3]** is a pipeline for analysis of large EST and mRNA databases in which sequences are first clustered based on pair wise sequence similarity, and then assembled by individual clusters to produce longer, more complete consensus sequences. TGICL is used to generate TIGR Gene Indices representing independent analyses for nearly 60 species with EST collections of fewer than 10000 to more than 4000000 sequences. EGene **[4]** is a generic, flexible and modular pipeline generation system that makes pipeline construction a modular job. EGene allows for third-party programs to be used and integrated according to the needs of distinct projects and without any previous experience of programming or formal language. A series of components to build pipelines for sequence processing is provided along with this.

MAKER **[5]** is a portable and easy to configure genome annotation pipeline. MAKER identifies repeats, aligns ESTs and proteins to a genome, produces *Ab-initio* gene predictions and automatically synthesizes these data into gene annotations having evidence-based quality values. MAKER’s modular construction allows it to break annotation process down into a series of five discrete activities that are easily interoperable: compute, filter/cluster, polish, synthesis, and annotate. Protein Annotation Toolkit (PAT) **[6]** is an integrated bio-computing server that provides a standardized web interface to a wide range of protein analysis tools. It is designed as a streamlined analysis environment that implements many features, which strongly simplify studies dealing with protein sequences/structures and improve productivity. Pipeline for Protein Annotation (PIPA) **[7]** annotates protein functions by combining the results of multiple programs and databases, such as InterPro and the Conserved Domains Database, into common Gene Ontology (GO) terms. The major algorithms implemented in PIPA are: (1) a profile database generation algorithm, which generates customized profile databases to predict particular protein functions, (2) an automated ontology mapping generation algorithm, which maps various classification schemes into GO, and (3) a consensus algorithm to reconcile annotations from the integrated programs and databases. Automatic and manual Functional Annotation in a Web services Environment (AFAWE) **[8]** simplifies the task of manual functional annotation by running different tools and workflows for automatic function prediction and displaying results in a way that facilitates comparison. AFAWE includes analyses for homolog detection, protein domain search and phylogenomics.

Workflow Management Systems for Phylogenetic Analysis:
Phylogeny and evolutionary analyses of sequences are among the most often used methodologies in laboratories working on functional, comparative and structural genomics. Phylogenetics analysis involves performing various tasks like multiple sequence alignment of uploaded sequences, curation of alignment obtained, construction of phylogenetics tress and their visualization as shown in **(Figure 2)**.

Further, execution of each of these tasks requires, use of specialized bioinformatics tools. As, there were many tools or web servers were developed for phylogenetic and evolutionary analysis, many workflows have been developed to automate this process. Several web sites offer phylogenetic tree reconstruction. Some offer a single tool, while others bring together many of the most popular programs for phylogenetic reconstruction. The workflow pipeline integrates these commonly used computational tools in a flexible way and allows the user to plug in custom sequence databases as well as alternative analysis tools. This section describes the important workflow management systems developed for phylogenetic analysis. **Table 3 (see supplementary material)** lists some of the important computational tools used for performing different tasks in this process. Major workflows for phylogenetic analysis along with important tools which are integrated are compared in **Table 4 (see supplementary material)**.

Phylogena **[9]** is a user-friendly, interactive graphical user interface running on desktop computers that automatically performs a Basic Local Alignment Search Tool (BLAST) with respect toquery sequences, selects a representative subset of them, then creates a multiple alignment from the selected sequences, and finally computes a phylogenetic tree. Phylemon [10] is an online platform for phylogenetic and evolutionary analyses of molecular sequence data. Phylemon also provides facilities for file format conversion, gene concatenation, tree visualization and the computation of distances between trees. Automated Simultaneous Analysis Phylogenetics (ASAP) **[11]** is an automated technique developed to assemble multigene/multi species matrices and to evaluate the significance of individual genes within the context of a given phylogenetic hypothesis. Matrix assembly at the genome scale involves the acquisition of hundreds to thousands of gene regions for the taxa of interest, the formatting of these sequences for use in an alignment program, aligning them, and finally eexport of the data partitions into formats used by phylogenetic analysis packages. Hal **[12]** is command line programs that brings together a number of bioinformatic applications into an efficient pipeline that inputs unaligned proteins sequences in fasta format and generate species trees from super alignments containing several orthologous protein sequences in a fully automated manner. The BioExtract **[13]** Server was used to create a workflow for comparing and aligning a number of nucleotide sequences to build a phylogenetic evolutionary tree. The web server Phylogeny.fr **[14]** is designed for non-specialists and has up-to-date programs that are often designed for experts. Armadillo v1.1 **[15]** is a novel workflow platform dedicated to designing and conducting phylogenetic studies, including comprehensive simulations. As Armadillo is an open-source project, it allows scientists to develop their own modules as well as to integrate existing computer applications. TreeDomViewer **[16]** is a visualization tool available as a web-based interface that combines phylogenetic tree description, multiple sequence alignment and InterProScan data of sequences and generates a phylogenetic tree projecting the corresponding protein domain information onto the multiple sequence alignment.

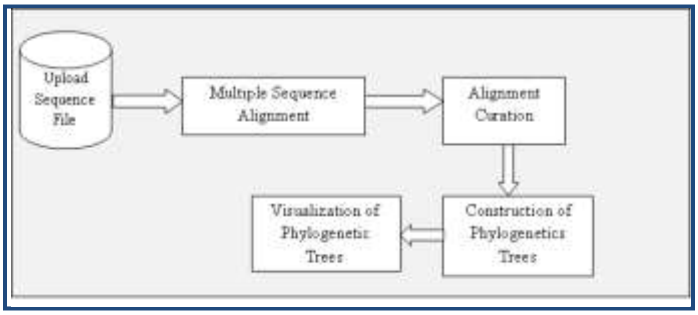
Upload Sequence File
Multiple Sequence Alignment
Alignment Curation
Construction of Phylogenetics Trees
Visualization of Phylogenetic Trees

Figure 2: General phylogenetic workflow system

Conclusion:
Analysis of 'omics' data using integrated bioinformatics tools through workflow management systems will help in increasing the productivity of researchers by reducing the time and effort spent on searching and executing each tool independently on different platforms. This article attempt to compare the features and performance of workflows developed for gene sequence analysis and evolutionary studies. Some of the important issues that must be addressed by these workflows are security, scheduling, load balancing and resource pooling. There is a need to design workflows through object oriented approach for its better re-usability, transportability, code sharing and ultimately reducing the efforts.

References:

[1] Lee B *et al. Nucleic Acids Res*. 2007 **35:** W159 [PMID: 17526512]
[2] Zhaolei Z *et al. Oxford Journals*. 2006 **22**: 1437 [PMID: 16574694]
[3] Pertea G *et al. Oxford Journals*. 2003 **19:** 651 [PMID: 12651724].
[4] Durham A *et al. Oxford Journals*. 2005 **21:** 2812 [PMID: 15814554]
[5] Cantarel B *et al. Genome Research*. 2007 **18**: 186 [PMID: 18025269]
[6] Je´ro^me Gracy & Laurent Chiche, *Nucleic Acids Res*. 2005 **33:** W65 [PMID: 15980554]
[7] Yu C *et al. Biomed Central*. 2008 **9:** 2109 [PMID:18221520]
[8] Jocker A *et al. Oxford Journals*. 2008 **24:** 2393 [PMID: 18697771]
[9] Hanekamp K *et al. Oxford Journals* 2007 **23:** 793 [PMID: 17332025]
[10] Gabaldo´ T *et al. Nucleic Acids Res*. 2007 **35:** W39 [PMID: 17452346]

[11] Sarkar I *et al. BioMed Central.* 2008 **9:** 2105 [PMID: 18282301]
[12] Robbertse B *et al. PLoS Currents Tree of Life.* 2011 [PMID: 21327165]
[13] Dereeper A *et al. Nucleic Acids Res.* 2008 **36**: W465 [PMID: 18424797]
[14] Carol M Lushbough *et al. Nucleic Acids Res.* 2011 1-5 [PMID: 21546552].
[15] Etienne L *et al. PLoS ONE.* 2012 **7**: e29903 [PMID: 22253821]
[16] Alako B *et al. Nucleic Acids Res.* 2006 **34**: W104 [PMID: 16844970]

Edited by P Kangueane
Citation: Sharma *et al.* Bioinformation 9(13): 663-672 (2013)

Supplementary material:

Table1: Tools used in the gene sequence analysis

Tool	Common Link	Description	Operating systems	HPC Compatibility*
Sequence Cleaning, Masking and Clustering				
EGAssembler	http://egassembler.hgc.jp/cgi-bin/eassembler4.cgi	Aligns and merges sequence fragments	Online	-
DNAStar	http://www.dnastar.com/	DNA and protein sequence analysis; next and third generation sequence assembly and analysis	Windows based	-
SeqTrim	http://www.scbi.uma.es/cgi-bin/seqtrim/seqtrim_login.cgi	High throughput reprocessing of sequence reads	Online	-
SeqClean	http://seqclean.sourceforge.net/	Automated trimming and validation of ESTs or other DNA sequences by screening for various contaminants, low quality and low-complexity sequences	Linux	Yes
VecScreen	http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html	Identifying segments of a nucleic acid sequence that may be of vector origin	Online	-
RepeatMasker	http://www.repeatmasker.org/	Screens DNA sequences for interspersed repeats and low complexity DNA sequences	Linux	-
Cap3	http://seq.cs.iastate.edu/cap3.html	DNA sequence assembly program	Windows, MacOS, Linux, Solaris	-
Assembly				
Phred & Phrap	http://www.phrap.com/	Sequence assembly	Windows, MacOS, Linux	Yes
Consed	http://www.phrap.org/consed/consed.html	Assembly finishing package	Windows, MacOS, Linux	-
CpG Island prediction				
CpG Island Searcher	http://www.uscnorris.com/cpgislands/cpg.cgi	screens for CpG islands	Online	-
CpG Plot	http://www.ebi.ac.uk/emboss/cpgplot/	Detection of regions of genomic sequences that are rich in the CpG pattern	Online	-
CpGPAP	http://bio.kuas.edu.tw/CpGPAP/	predicting CpG islands in genome sequences	Online	-
tRNA prediction				
tRNAscan	lowelab.ucsc.edu/tRNAscan-SE/	Search for tRNA genes in genomic sequence	Online	-
Gene Prediction				
GeneMark	http://topaz.gatech.edu/GeneMark/gmchoice.html	family of gene prediction programs	Linux, Solaris, Mac OS	
GeneMark.hmm	http://topaz.gatech.edu/GeneMark/hmmchoice.html	gene prediction program for prokaryotes and eukaryotes	Windows, Mac OS X, and Linux	-
GLIMMER	http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi	finding genes in microbial DNA	Linux	-
GENESCAN	http://genes.mit.edu/GENSCAN.html	finding genes using Fourier transform	Windows, Linux and Mac OS X	-
GENOMESCAN	http://probcons.stanford.edu/	Predicts locations and exon-intron structures of genes in genomic sequences from a variety of organisms.	Linux	-
ATGpr	http://atgpr.dbcls.jp/	identifying translational initiation sites in cDNA sequences	Unix, Linux, cygwin and MacOSX	Yes
AUGUSTUS	http://bioinf.uni-greifswald.de/augustus/	predicts genes in eukaryotic genomic sequences	Unix, Linux, cygwin and MacOSX	-

GlimmerHMM	http://www.cbcb.umd.edu/software/GlimmerHMM/	GlimmerHMM's predicts introns of each phase, intergenic regions, and four types of exons (initial, internal, final, and single).	Linux RedHat 6.x+, Sun Solaris, and Alpha OSF1	-
tRNAscan - SE	http://lowelab.ucsc.edu/tRNAscan-SE/	identifies transfer RNA genes in DNA sequence	Online Unix	-

*HPC – High Performance Computing

Table 2: Details of workflow management systems available for gene sequence analysis

Name	Year	Implementation	Tools	Performance & Limitations	URL
ESTpass	2007	-Accepts a FASTA-formatted EST file. -A web server developed using MySQL, HTML and JSP, Perl, Python, Java, and an Apache Ant -Based on Linux machine with four dual-core AMD Opteron 875 CPUs (8 cores) and 16 GB of RAM	d2_cluster and CAP3 programs for clustering and assembling. -Annotating the putative transcript sequences using RefSeq, InterPro, GO and KEGG gene databases	-EST analysis is generally time-consuming due to the large number of EST sequences—it may take more than 1 day depending on the number of EST sequences. -ESTpass cannot accept chromatogram files due to file-size limitations of web-based uploading -The maximum number of input EST sequences in a single submission is 10 000 EST sequences	http://estpass.kobic.re.kr/.
Pseudopipe	2006	-Accepts genomic sequence after repeat-masking, the comprehensive and non-redundant set of protein sequences in the genome, and the chromosomal coordinates of the functional gene. -Phython Based	BLAST	-Except for the step of whole-genome BLAST search, the annotation pipeline can be run on an entire genome in a few hours, on a reasonably robust Linux workstation (3.0 GHz, 1 GB RAM). -Multiple concurrent independent pipeline runs could be started on multiple computers, e.g. several chromosomes can be grouped together and processed on a single computer.	http://pseudogene.org/
TGICL	2003	-Developed using C and PERL Linux Based multi-CPU architectures including SMP and PVM. -Pairwise searches are performed by mgblast, written in C	Lucy, SeqClean, UniVec, RepeatMasker, megablast	-Clustering is very fast due to the modified megablast engine used for pairwise searches and distributed processing makes TGICL even faster: on a PVM cluster with 20 Pentium III nodes, an input file of 1 700 000 entries was clustered in approximately one hour and assembly was completed the following day. Sets of 150 000 sequencescan be fully clustered and assembled overnight on a single CPU. -TGICL has difficulty with highly expressed genes that have several thousand ESTs in a single cluster. For these, CAP3 or other assemblers generally run out of memory.	http://www.tigr.org/tdb/tgi/software/
EGene	2005	EGene was written in Perl and is designed to run on Unix/Linux operating systems.	CAP3, Phrap, BLAST, Cross_match, Phred,	- pipelines can be executed in concurrent mode	http://www.lbm.fmvz.usp.br/egene/
MAKER	2012	It is written in Perl, Bioperl and	RepeatMasker, BLAST,,	-MPI based and capable of parallelization across computer	http://www.yandell-

		outputs are in GFF3 or FASTA format	Exonerate, SNAP,	cluster -For C. elegans genes, performance in genomic and overlap was 90.75% compared with 91.12% and 93.26% for Gramene and Augustus. -In case of exon overlap, it is under performing with 3.67% and 5.02% for sensitivity relative to Gramene and Augustus.	lab.org/maker.
PAT	2005	It is written in CGI and PERL	-Automatic retrieval of protein entries SWISSPROT, TREMBL, PDB and PFAM)using specific identifier or accession number indexes. -It also launches the 1D, 2D and 3D protein analysis tools through a uniform web interface.	-The processing of query is automatically aborted for CPU time is larger than 10 minutes; process requiring more than 500Mb of RAM; output files are bigger than 10Mb. -Does not run on multi-processors linux cluster	http://pat.cbs.cnrs.fr
PIPA	2008	input protein sequences in FASTA format	CatFam, CDD,COG, Pfam, TIGRfam, SMART Gene3D, FprintScan PANTHER,,SUPERFAMILY, ProDom,PIR, PROSITE, COIL, Phobius ,PSORTb	-LINUX computer cluster integrated programs are executed in parallel using 64 computing processors -PIPA can annotate a typical bacterial genome consisting of 4,000 proteins in about six hours.	
AFAWE	2008	-	WU-Blast, UniProt database, SwissProt database, InterProScan, RPSBlast, Conserved Domain Database	Running a complete automatic annotation with AC144389_35.2 as query, takes about 5 min on our machines, if all found domains and homologous proteins are in the AFAWE database.	http://bioinfo.mpiz-koeln.mpg.de/afawe/

Table 3: Tools used in the phylogenic workflow management system

Tool	Common Link	Description	Operating systems	HPC Compatiblity*
Local Alignment and Sequence Search				
BLASTp	http://blast.ncbi.nlm.nih.gov/Blast.cgi	Search protein database using a protein query	Windows, Linux and Macintosh	Yes
Multiple Sequence Alignment				
ClustalW	http://clustal.org	Computes a multiple sequence alignment for protein or DNA sequences	Windows, Linux and Macintosh	
ClustalW2	http://www.ebi.ac.uk/Tools/msa/clustalw2/	Multiple sequence alignment program for DNA or proteins	Windows, Linux and Macintosh	Yes
BAli-phy	http://www.biomath.ucla.edu/msuchard/bali-phy/	MCMC software for simultaneous Bayesian estimation of alignment and phylogeny (and other parameters)	Windows, Mac OS X, and Linux	-
Kalign	http://www.ebi.ac.uk/Tools/msa/kalign/	Multiple sequence alignment program	Linux	-

MUSCLE	http://www.ebi.ac.uk/Tools/msa/muscle/	**MU**ltiple **S**equence **C**omparison by **L**og- **E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee	Windows, Linux and Mac OS X	-
ProbCons	http://probcons.stanford.edu/	Protein multiple sequence alignment program	Linux	-
T-Coffee	http://www.tcoffee.org/Projects_home_page/t_coffee_home_page.html /	Protein multiple sequence alignment program	Unix, Linux, cygwin and MacOSX	Yes
3D-Coffee	http://www.tcoffee.org/Projects_home_page/expresso_home_page.html	Combining sequences and structure for multiple sequence alignment	Unix, Linux, cygwin and MacOS X	-
Phylogenetic Tree Inference				
MrBayes	http://mrbayes.csit.fsu.edu/	Estimate phylogeny upon Bayesian inference which is based on the probability of a tree conditioned on the observations	Macintosh, Windows, UNIX	Yes
PAUP	http://paup.csit.fsu.edu/index.html	Tool for inferring and interpreting phylogenetic trees. It analyzes the molecular sequences and morphological data using maximum likelihood, parsimony and distance methods	Macintosh, Windows, UNIX/DOS	No
1. MacClade	http://macclade.org/intro.html	Tool for phylogenetic analysis	MacOS X	-
2. PHYLIP	http://evolution.genetics.washington.edu/phylip.html	Phylogenetic inference package using maximum parsimony, distance matrix, maximum likelihood	Windows, Mac OS and Linux	No
3. PHYML	http://www.atgc-montpellier.fr/organization/	Estimates maximum likelihood phylogenies	Windows, Mac OS and Linux	Yes
4. Tree-Puzzle	http://www.tree-puzzle.de/	Maximum likelihood and statistical analysis	Windows, Mac OS and Linux	Yes
5. TNT	http://www.zmuc.dk/public/phylogeny/TNT/	Phylogenetic inference using parsimony, weighting, ratchet, tree drift, tree fusing, sectorial searches	Windows, Mac OS and Linux	Yes
6. PAML	http://abacus.gene.ucl.ac.uk/software/paml.html	Phylogenetic analysis by maximum likelihood and Bayesian inference	UNIX,Linux,Mac OS X, Windows	Yes
7. IQPNNI	http://www.cibiv.at/software/iqpnni/	Iterative maximum likelihood tree search with stopping rule	Linux, MacOS and Windows	Yes
8. RAxML-HPC	http://phylobench.vital-it.ch/raxml-bb/	Randomized Axelerated Maximum Likelihood for High Performance Computing (nucleotides and aminoacids)	Linux, MacOS and Windows	Yes
9. GARLI	https://www.nescent.org/wg_garli/Main_Page	performs phylogenetic inference using the maximum-likelihood criterion	Mac OS and Linux	Yes
10. Mafft	http://mafft.cbrc.jp/alignment/software/	MAFFT offers various multiple alignment strategies. They are classified into three types, (**a**) the progressive method, (**b**) the iterative refinement method with the WSP score, and (**c**) the iterative refinment method using both the WSP and consistency score.	Linux, MacOS and Windows	-
11. POY	http://research.amnh.org/scicomp/scripts/download.php	A phylogenetic analysis program that supports multiple kinds of data and can perform alignment and phylogeny inference. A variety of heuristic algorithms have been developed for this purpose	Mac OS and WinXP	Yes
Phylogenetic Tree Visualization				

PhyloWidget	http://www.phylowidget.org/	View, edit, and publish phylogenetic trees online; interfaces with databases	Linux	-
Archaeopteryx	http://www.phylosoft.org/archaeopteryx/	Java tree viewer and editor (used to be ATV)	Linux, MacOS and Windows	-
ScripTree	www.scriptree.org	Tool for the automation of tree rendering	Windows and Unix-like systems including OS X	-
TreeDyn	http://www.treedyn.org/	TreeDyn links unique leaf labels to lists of variables/values pairs of annotations (meta-information), independently of the tree topologies, remaining fully compatible with the basic newick format.	Windows and Unix-like systems including OS X	-
Drawgram	http://cmgm.stanford.edu/phylip/drawgram.html	DRAWGRAM plots rooted phylogenies, cladograms, circular trees and phenograms in a wide variety of user-controllable formats. The program is interactive and allows previewing of the tree .	Linux, MacOS and Windows	-
Drawtree	http://cmgm.stanford.edu/phylip/drawtree.html	DRAWTREE interactively plots an unrooted tree diagram, with many options including orientation of tree and branches, label sizes and angles, margin sizes.	Linux, MacOS and Windows	-

*HPC – High performance computing

Table 4: Comparison of workflow management sytems available for phylogenetic analyses

Name	Year	Implementation	Tools Used	Performance & Limitations	URL
Phylogena	2007	Java, Biojava and knowledge base is written in Prolog	ATV , JalView, BLAST, ClustalW,T-Coffee,DIALIGN POA, Mafft,, MUSCLE, Kalign	-Output files are very large and are stored in the memory -Approximately 200 and 400 ORFs can be analysed with 1GB memory -Not available for Macintosh platform	http://www.awi.de/en/phylogena
Phylemon	2007	-Web based, accepts input in Fasta and PHYLIP data format. -Developed using Java applet environment	ClustalW, MUSCLE, Lagan, M-Lagan, TrimAl, CDS-ProtAl, ConcatenAl, ReadAl, Seqboo, Consense, Dnadist, Protdist, DnaML, ProML, DnaPars, ProtPars, Neighbor, Fitch, ETE, PhyML-Best-AIC-Tree, PhyML, Tree-Puzzle, MrBayes, ProtTest, jModelTest, RRTree, SLR, YN00, CodeML	-asynchronous use of tools (a program can be left running to later come back to see the results)	http://phylemon.bioinfo.cipf.es.
ASAP	2008	-Accepts input in Fasta or a list of NCBI accession numbers. -Developed using PERL	PAUP	Requires PAUP(command-line) and MUSCLE	http://sarkarlab.mbl.edu/ASAP

HAL	2011	-Accepts inputs as unaligned proteins sequences in fasta format -Developed using Perl -Based on 64-bit Linux architecture	BLASTp, MCL, Muscle, Mafft PROBCONS ClustalW GBlocks RAxML PhyML PAU, PHYLIP	-It is most efficient when using a Sun Grid Engine (SGE), which significantly decreases processing time since serial jobs are run on several processors -Running on a 32-bit machine may present a problem of insufficient memory for larger analyses. -Also run on a single machine, but this will take considerably more time depending on the number of taxa and size of the input genomes.	http://sourceforge.net/projects/bio-hal/
BioExtract		Protein sequences shell script utilizing the Vmatch tool	BLASTp, xmknr ClustalW, MrBayes	-The execution of any created workflow generates the running of all the tools at once, and provides access to all the results via the general workflow report. Consequently, the results are obtained in an extremely reduced time compared to conventional methods.	http://www.myexperiment.org/workflows/1941.html
Phylogeny.fr	2008	Developed using Perl	Many tools for phylogenetic analysis	The platform currently runs on a dedicated server (PowerEdge 2850-Xeon 2.8 GHz/2_2 MB Dual Core), except for the BLAST module which is parallelized on a 25-CPU cluster. MUSCLE and ClustalW are limited to 200 sequences, while T-Coffee and 3D-Coffee limitations are <50 sequences and <2000 sites. Distance-based phylogeny programs (i.e. NJ and BioNJ) have no limitation, while all other phylogeny programs are limited <10 000 000	http://www.phylogeny.fr/
Armadillo 1.1	2012	Developed in Java	Many tools for phylogenetic analysis	User has to cope with particular memory and parameters limitations imposed by the applications included in the Armadillo platform as well as with the RAM overflow that can be caused by executing those applications on large datasets	http://www.bioinfo.uqam.ca/armadillo/
TreeDomViewer	2006	Perl CGI Apache 2.0 web server on a Linux platform (SuSE linux Enterprise Server 9). Input is a set of aligned or unaligned sequences	ClustalW, PHYLIP, InterProScan	Running parallel on 10 nodes of a small Linux cluster, the analysis of 60 protein sequences of 1000 amino residues each is performed in <3 minutes	http://www.bioinformatics.nl/tools/treedom/