15 Phylogenetic Analysis of DNA Sequence Data

SERGEI A. SUBBOTIN

Plant Pest Diagnostic Center, California Department of Food and Agriculture, California, USA; Center of Parasitology of A.N. Severtsov Institute of Ecology and Evolution, Moscow, Russia

15.1 Introduction

The goal of phylogenetics is to construct relationships that are true representations of the evolutionary history of a group of organisms or genes. The history inferred from phylogenetic analysis is usually depicted as branching in tree-like diagrams or networks. In nematology, phylogenetic studies have been applied to resolve a wide range of questions dealing with improving classifications and testing evolution processes, such as co-evolution, biogeography and many others. There are several main steps involved in a phylogenetic study:

- selection of ingroup and outgroup taxa for a study;
- selection of one or several gene fragments for a study;
- sample collection, obtaining PCR products and sequencing of gene fragments;
- visualization, editing raw sequence data and sequence assembling;
- search for sequence similarity in a public database;
- making and editing multiple alignment of sequences;
- selecting appropriate DNA model for a dataset;
- phylogenetic reconstruction using minimum evolution, maximum parsimony, maximum likelihood and Bayesian inference;
- visualization of tree files and preparation of tree for a publication; and
- sequence submission to a public database.

Molecular phylogenetic study requires particularly careful planning because it is usually relatively expensive in terms of the cost in reagents and time. The first and the most important step of any study is to define clearly the specific biological question to be answered. When the biological problem is formulated and the

© CAB International 2021. *Techniques for Work with Plant and Soil Nematodes* (eds R.N. Perry, D.J. Hunt and S.A. Subbotin)
10.1079/9781786391759.0015

literature survey pertaining to the group of interest is completed, selection of ingroup and outgroup taxa and appropriate genes should be done. These could be representatives of the same species from various locations or hosts, or different species of the same genus, or representatives of related genera or higher taxonomic categories. It is advisable to include as many samples as possible, as well as to choose several gene fragments to reduce artefactual associations between terminals.

The third stage, which includes sample collection, obtaining PCR products and sequencing, may require several weeks or months and is the most time-consuming stage in a study. Before launching a full-scale project, a pilot study with a limited sample number is recommended to determine if the selected genetic markers give sufficient resolution for phylogeny of the studied group.

The last seven stages require a computer with an Internet connection and suitable software. This chapter deals with these stages. The software and instructions for its use are discussed below. All of them are free and can be downloaded from different websites, except for PAUP* (Phylogenetic Analysis Using Parsimony*), which can be purchased from the official website. This chapter only considers data from nucleotide sequences and does not take into account amino acid sequences data.

There are many phylogenetics programs that perform similar functions (Table 15.1). The comprehensive list can be found in the websites:

- http://evolution.genetics.washington.edu/phylip/software.html;
- http://www.phylo.org;
- https://www.phylogeny.fr;
- https://isogg.org/wiki/Phylogeny_programs;
- https://en.wikipedia.org/wiki/List_of_phylogenetics_software; and
- https://molbiol-tools.ca/Phylogeny.htm.

Table 15.1. Phylogenetics programs.

Software	Website
BEAST (**B**ayesian **E**volutionary **A**nalysis **S**ampling **T**rees) is a cross-platform program for Bayesian phylogenetic analysis of molecular sequences using MCMC	https://beast.community https://www.beast2.org
BioEdit is a **Bio**logical Sequence **Edit**or	https://bioedit.software.informer.com
Clustal is a series of widely used computer programs used in Bioinformatics for multiple sequence alignment	http://www.clustal.org
GARLI (**G**enetic **A**lgorithm for **R**apid **L**ikelihood **I**nference) is a program for inferring phylogenetic trees	https://code.google.com/archive/p/garli/
MAFFT (**M**ultiple **A**lignment using **F**ast **F**ourier **T**ransform) is a program used to create multiple sequence alignments of amino acid or nucleotide sequences.	https://mafft.cbrc.jp/alignment/software/
MEGA (**M**olecular **E**volutionary **G**enetic **A**nalysis) is a software for conducting statistical analysis of molecular evolution and for constructing **phylogenetic** trees.	https://www.megasoftware.net
MrBayes is a free software tool that performs Bayesian inference of phylogeny	https://nbisweden.github.io/MrBayes/
MUSCLE (**Mu**ltiple **S**equence **C**omparison by **L**og-**E**xpectation) is a software for multiple sequence alignment of protein and nucleotide sequences.	http://www.drive5.com/muscle/
PAUP* - **P**hylogenetic **A**nalysis **U**sing **P**arsimony (*and other methods)	http://paup.phylosolutions.com
RAxML (**R**andomized **Ax**elerated **M**aximum **L**ikelihood) is a program for sequential and parallel maximum likelihood-based inference of large phylogenetic trees.	https://cme.h-its.org/exelixis/web/software/raxml/index.html http://www.trex.uqam.ca/index.php?action=raxml&project=trex
T-Coffee (**T**ree-based **C**onsistency **O**bjective **F**unction **f**or Alignm**e**nt **E**valuation) is a multiple sequence alignment software using a progressive approach	http://www.tcoffee.org

The reliability and practicality of the software depends on the structure and size of the data. The merits and pitfalls of various methods are the subject of often acrimonious debates in taxonomic and phylogenetic journals (Lemey *et al.*, 2009; Yang and Rannala, 2012). Generally, most software packages have been developed and maintained through the efforts of scientists in related fields and released under free software licences.

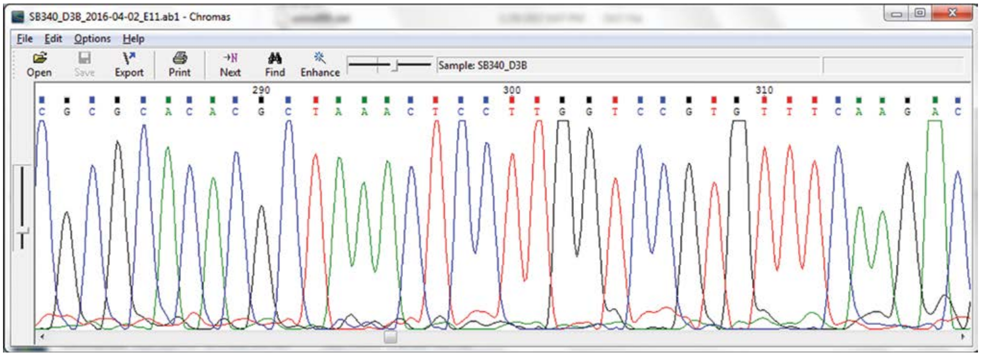
Phylogenetics is a rather complex and rapidly expanding field of research. In this chapter, some basic approaches for the analysis of nucleotide non-coding gene sequences are given and discussed with an assumption that the gene tree reflects the organism phylogeny. Many other important software with statistical testing for phylogenetic studies are not covered here, and specialized literature (e.g. Lemey *et al.*, 2009; Hall, 2017) may be recommended for researchers.

15.2 Visualization and Editing of Raw Sequence Data

The sequencer is a laser-based instrument that utilizes fluorescent labels to analyse the products of a sequencing reaction as they migrate through a gel. After the data are collected from a sequencing run, special software identifies and tracks the sample lanes of the gel and subsequently normalizes and integrates the data into files. Automated DNA sequencer generates two file types: (i) a four-colour chromatogram showing the results of the sequencing run; and (ii) a text file of sequence data. It is always highly recommended to check the quality of the chromatogram file of a studied sample before converting it to a sequence text file.

There are several programs for visualization of a raw sequence data, two of which can be freely downloaded: Chromas (http://technelysium.com.au/wp/chromas/) developed by Technelysium Pty Ltd and FinchTV (http://www.geospiza.com/ftvdlinfo.html) originally designed by Geospiza. Chromas is a simple, easy-to-use viewer and editor for chromatograms from automated Sanger sequencers. Chromas has several features including automatic removal of low-quality sequence or vector sequences, copying the sequence to the clipboard in plain text, FASTA format for pasting into other applications, performing reverse and complement the sequence and chromatogram, and displaying translations in three frames along with the sequence. A chromatogram shows a sequence of peaks in four colours, each representing the base: A, green or yellow; G, black; T, red; and C, blue.

Once the sequence is obtained, the quality of sequence reading should be proofread to ensure that all ambiguous sites are correctly resolved in a chromatogram file. Good-quality sequences are characterized by well-defined peak resolution, uniform peak spacing and high signal-to-noise ratios (Fig. 15.1). If the quality of a sequence is not good and contains double or asymmetrical peaks or strong background noise, sequencing reactions should be repeated with both forward and reverse primers (Fig. 15.2). There are many reasons why a sequence reaction can fail. A good sequence generally begins approximately around 20–30 bp and lasts up to 700–800 bp. Any bad-quality sequence should be eliminated from further analysis. Sometimes, the software may miscall or miss a nucleotide or add an additional nucleotide, making proofreading of chromatograph files an important step of data verification. A single peak position within a trace may have also two peaks of different colours instead of just one. This is a common problem when sequencing a PCR product derived


SB340_D3B_2016-04-02_E11.ab1 - Chromas

Sample: SB340_D3B
File
Edit
Options
Help
Open
Save
Export
Print
Next
Find
Enhance
290
300
310
C G C G C A C A C G C T A A A C T C C T T G G T C C G T G T T T C A A G A C

Fig. 15.1. Chromas with good-quality sequence containing well-defined peaks.

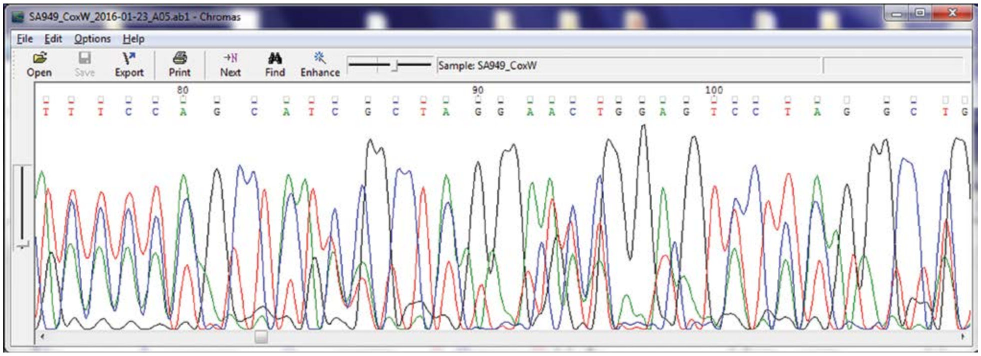
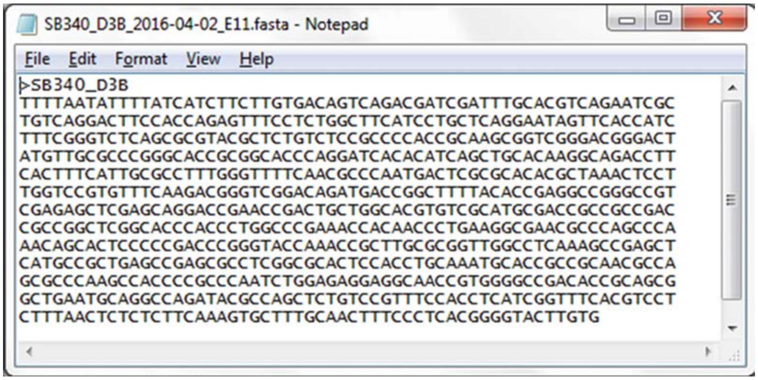

SA949_CoxW_2016-01-23_A05.ab1 - Chromas
File
Edit
Options
Help
Open
Save
Export
Print
Next
Find
Enhance
Sample: SA949_CoxW
80
90
100
T T T C C A G C A T C G C T A G G A A C T G G A G T C C T A G G C T G

Fig. 15.2. Chromas with bad-quality sequence containing double peaks and strong background.



```
>SB340_D3B
TTTTAATATTTTATCATCTTCTTGTGACAGTCAGACGATCGATTTGCACGTCAGAATCGC
TGTCAGGACTTCCACCAGAGTTTCCTCTGGCTTCATCCTGCTCAGGAATAGTTCACCATC
TTTCGGGTCTCAGCGCGTACGCTCTGTCTCCGCCCCACCGCAAGCGGTCGGGACGGGACT
ATGTTGCGCCCGGGCACCGCGGCACCCAGGATCACACATCAGCTGCACAAGGCAGACCTT
CACTTTCATTGCGCCTTTGGGTTTTCAACGCCCAATGACTCGCGCACACGCTAAACTCCT
TGGTCCGTGTTTCAAGACGGGTCGGACAGATGACCGGCTTTTACACCGAGGCCGGGCCGT
CGAGAGCTCGAGCAGGACCGAACCGACTGCTGGCACGTGTCGCATGCGACCGCCGCCGAC
CGCCGGCTCGGCACCCACCCTGGCCCGAAACCACAACCCTGAAGGCGAACGCCCAGCCCA
AACAGCACTCCCCCGACCCGGGTACCAAACCGCTTGCGCGGTTGGCCTCAAAGCCGAGCT
CATGCCGCTGAGCCGAGCGCCTCGGCGCACTCCACCTGCAAATGCACCGCCGCAACGCCA
GCGCCCAAGCCACCCCGCCCAATCTGGAGAGGAGGCAACCGTGGGGCCGACACCGCAGCG
GCTGAATGCAGGCCAGATACGCCAGCTCTGTCCGTTTCCACCTCATCGGTTTCACGTCCT
CTTTAACTCTCTCTTCAAAGTGCTTTGCAACTTTCCCTCACGGGGTACTTGTG
```

Fig. 15.3. Sequence in FASTA format.

from diploid genomic DNA having polymorphic positions. In this case, it is recommended to edit the base in a chromatogram file according to the single letter nucleic acid code recommendations: R, G or A (purine); Y, T or C (pyrimidine); K, G or T (keto); M, A or C (amino); S, G or C (strong); W, A or T (weak); B, G or T or C; D, G or A or T; H, A or C or T; V, G or C or A; N, A or G or C or T (any).

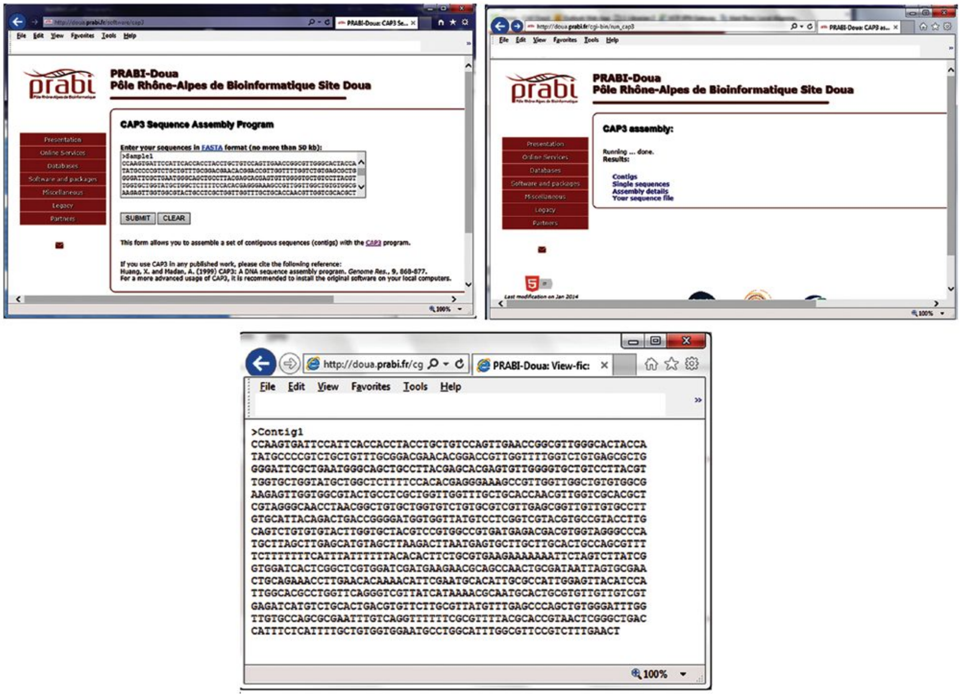
Work with Chromas version 2.62 includes a few steps:

- Open a chromatogram file. Choose **File > Open.**
- If an antisense chain (3′–5′) is open, use reverse and complement option. Choose **Edit > Reverse+Complement.**
- Check quality and trim low-quality data. Choose **Edit > Trim Low Quality.**
- Convert into a FASTA format, which begins with a single-line description, followed by lines of sequence data. Choose **File > Export.** After a chromatogram file has been examined and edited, it should be exported into a FASTA format file (Fig. 15.3).

15.3 Sequence Assembling

If a PCR product cannot be covered by a single sequencing reaction, several reactions should be performed. The chromatograms of these reactions should be verified and then displayed in a single FASTA format file. The goal of assembling is to create a single consensus sequence covering the whole length of a studied amplicon from several partly overlapped sequences.

A number of DNA sequence assembly programs have been developed including CAP3, which has also web-based version (http://doua.prabi.fr/software/cap3). For more advanced usage, it is recommended to install the software on your local computer. The program features include fast identification of pairs of reads



PRABI-Doua
Pôle Rhône-Alpes de Bioinformatique Site Doua

CAP3 Sequence Assembly Program

Enter your sequences in FASTA format (no more than 50 kb):

```
>Sample1
CCAAGTGATTCCATTCACCACCTACCTGCTGTCCAGTTGAACCGGCGTTGGGCACTACCA
TATGCCCCGTCTGCTGTTTGCGGACGAACACGGACCGTTGGTTTTGGTCTGTGAGCGCTG
GGGATTCGCTGAATGGGCAGCTGCCTTACGAGCACGAGTGTTGGGGTGCTGTCCTTACGT
TGGTGCTGGTATGCTGGCTCTTTTCCACACGAGGGAAAGCCGTTGGTTGGCTGTGTGGCG
AAGAGTTGGTGGCGTACTGCCTCGCTGGTTGGTTTGCTGCACCAACGTTGGTCGCACGCT
```

This form allows you to assemble a set of contiguous sequences (contigs) with the CAP3 program.

If you use CAP3 in any published work, please cite the following reference:
Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.*, 9, 868-877.
For a more advanced usage of CAP3, it is recommended to install the original software on your local computers.

PRABI-Doua
Pôle Rhône-Alpes de Bioinformatique Site Doua

CAP3 assembly:

Running .... done.
Results:

Contigs
Single sequences
Assembly details
Your sequence file

```
>Contig1
CCAAGTGATTCCATTCACCACCTACCTGCTGTCCAGTTGAACCGGCGTTGGGCACTACCA
TATGCCCCGTCTGCTGTTTGCGGACGAACACGGACCGTTGGTTTTGGTCTGTGAGCGCTG
GGGATTCGCTGAATGGGCAGCTGCCTTACGAGCACGAGTGTTGGGGTGCTGTCCTTACGT
TGGTGCTGGTATGCTGGCTCTTTTCCACACGAGGGAAAGCCGTTGGTTGGCTGTGTGGCG
AAGAGTTGGTGGCGTACTGCCTCGCTGGTTGGTTTGCTGCACCAACGTTGGTCGCACGCT
CGTAGGGCAACCTAACGGCTGTGCTGGTGTCTGTGCGTCGTTGAGCGGTTGTTGTGCCTT
GTGCATTACAGACTGACCGGGGATGGTGGTTATGTCCTCGGTCGTACGTGCCGTACCTTG
CAGTCTGTGTGTACTTGGTGCTACGTCCGTGGCCGTGATGAGACGACGTGGTAGGGCCCA
TGCTTAGCTTGAGCATGTAGCTTAAGACTTAATGAGTGCTTGCTTGCACTGCCAGCGTTT
TCTTTTTTTCATTTATTTTTTACACACTTCTGCGTGAAGAAAAAAATTCTAGTCTTATCG
GTGGATCACTCGGCTCGTGGATCGATGAAGAACGCAGCCAACTGCGATAATTAGTGCGAA
CTGCAGAAACCTTGAACACAAAACATTCGAATGCACATTGCGCCATTGGAGTTACATCCA
TTGGCACGCCTGGTTCAGGGTCGTTATCATAAAACGCAATGCACTGCGTGTTGTTGTCGT
GAGATCATGTCTGCACTGACGTGTTCTTGCGTTATGTTTGAGCCCAGCTGTGGGATTTGG
TTGTGCCAGCGCGAATTTGTCAGGTTTTTTCGCGTTTTACGCACCGTAACTCGGGCTGAC
CATTTCTCATTTTGCTGTGGTGGAATGCCTGGCATTTGGCGTTCCGTCTTTGAACT
```

Fig. 15.4. CAP3 home page and output.

with an overlap, clipping of 5′ and 3′ poor regions of reads, efficient computation and evaluation of overlaps, use of forward–reverse constraints to correct errors in construction of a contig, and generation of consensus sequence for a contig (Huang and Madan, 1999).

Work with web-based CAP3 is simple and the results are self-explanatory.

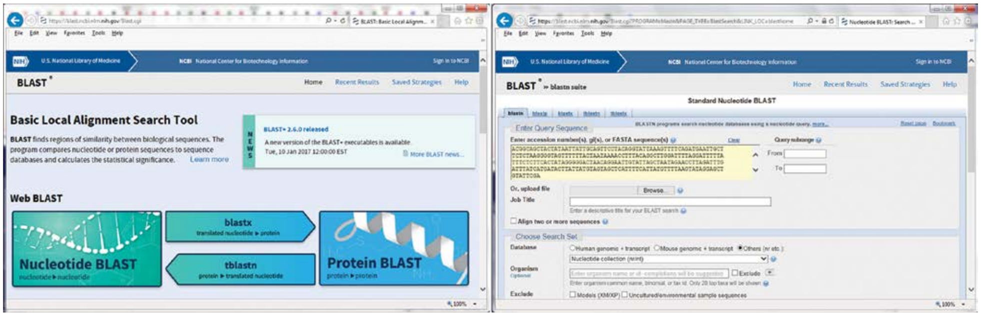
- Enter your sequences in FASTA format in a window and submit the data (Fig. 15.4).
- Results are displayed as contigs, single sequence, assembling details and your sequence file. Every base in an assembly must be covered by at least two sequences of high quality. Validating sequence coverage provides a high degree of confidence in the consensus base calls.

15.4 Similarity Search in a Sequence Database

One of the most important steps in the study is the identification of your sequence and comparison with all known sequences collected in different databases. This procedure is called similarity search. BLAST (Basic Local Alignment Search Tool) is a powerful program for rapid searching of nucleotide and protein databases. The BLAST program finds regions of local similarity between sequences. It was developed in 1989 at the National Center for Biotechnology Information (NCBI), Maryland, USA. A BLAST query uses statistical methods to compare a DNA or protein input sequence (‘query sequence’) to a database of sequences (‘subject sequences’) and returns those sequences that have a significant level of similarity to the query sequence. The BLAST algorithm calculates similarity scores for local alignments (i.e. the most similar regions between two sequences) between the query sequence and subject sequences using specific scoring matrices, and returns a table of the best matches (‘hits’) from the database (Altschul, 1990; Wheeler and Bhagwat, 2007).

A BLAST search includes a few steps:

- Point your browser to the NCBI BLAST server at: http://www.ncbi.nlm.nih.gov/BLAST.
- Select **Nucleotide BLAST** (Fig. 15.5).



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

BLAST+ 2.6.0 released

A new version of the BLAST+ executables is available

Tue, 10 Jan 2017 12:00:00 EST

Web BLAST

Standard Nucleotide BLAST

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database ○Human genomic + transcript ○Mouse genomic + transcript ●Others (nr etc.)

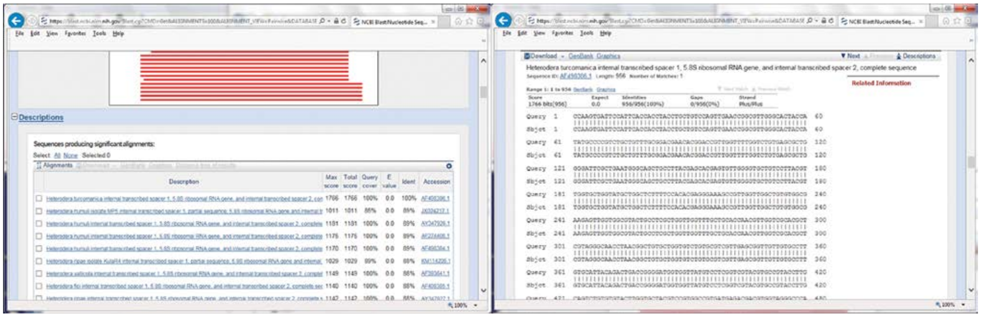
Nucleotide collection (nr/nt)

Organism Optional

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude □ Models (XM/XP) □ Uncultured/environmental sample sequences

Fig. 15.5. BLAST searching pages.



Descriptions

Sequences producing significant alignments:

Select: All None Selected 0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	Heterodera turcomanica internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, con	1766	1766	100%	0.0	100%	AF498306.1
□	Heterodera humuli isolate MP5 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal t	1011	1011	86%	0.0	89%	JX024217.1
□	Heterodera humuli internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete	1181	1181	100%	0.0	89%	AY347926.1
□	Heterodera humuli internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete	1175	1175	100%	0.0	89%	AF274408.1
□	Heterodera humuli internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete	1170	1170	100%	0.0	89%	AF498384.1
□	Heterodera ripae isolate KulaR4 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal	1029	1029	99%	0.0	88%	KM114226.1
□	Heterodera salixophila internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complet	1149	1149	100%	0.0	88%	AF393841.1
□	Heterodera fici internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sec	1140	1140	100%	0.0	88%	AF498385.1
□	Heterodera ripae internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete s	1142	1142	100%	0.0	88%	AY347927.1

Heterodera turcomanica internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence

Sequence ID: AF498306.1 Length: 956 Number of Matches: 1

Range 1: 1 to 956

Score	Expect	Identities	Gaps	Strand
1766 bits(956)	0.0	956/956(100%)	0/956(0%)	Plus/Plus

Fig. 15.6. BLAST output pages.

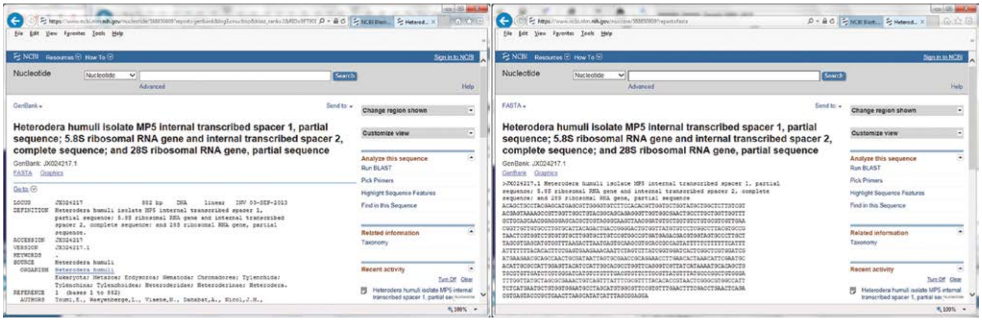
- Paste your sequence in the Query Sequence field, choose search set in the Database as **Others, Nucleotide collection. Organism group** could be also defined to limit your search to the DNA of a specific organism. Alignment parameters can be also modified (for example, increase: **Max target sequences**). The default setting uses a version of BLASTN called **megaBLAST**. Click the **BLAST** button and wait for the results.

The results from a BLAST search are divided into three sections: (i) the graphic pane; (ii) a results table; and (iii) the alignments between the query and the hits. Conclusions can be based on interpretations of the BLAST results table that provides basic information about the hits together with the statistics of each hit. Information includes: Max score (highest alignment score (bit-score) between the query sequence and the database sequence segment); Total score (sum of alignment scores of all segments from the same database sequence that match the query sequence (calculated over all segments)); Query cover (the percentage of the query that aligns with the hit); E value (number of alignments expected by chance with a particular score or better); Identity (percentage of identity between the query and the hit in a nucleotide to nucleotide alignment); and Accession (GenBank sequence number). The top hits are most significant and similar to the submitted sequence (Fig. 15.6).

15.5 Sequence Retrieval from the Database

There are several ways to retrieve sequences to build your dataset for phylogenetic analysis:

- Sequences can be obtained from the BLAST search result page. Click on **Accession** and the hyperlink takes you to the database entry that contains this sequence. Click on **FASTA** to convert the GenBank format into a FASTA format (Fig. 15.7).
- Point your browser to http://www.ncbi.nlm.nih.gov. Select **Nucleotide** in a search field and type organism name(s) and gene name or accession number(s).



Heterodera humuli isolate MP5 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence

GenBank: JX024217.1

```
LOCUS       JX024217                 882 bp    DNA     linear   INV 03-SEP-2013
DEFINITION  Heterodera humuli isolate MP5 internal transcribed spacer 1,
            partial sequence; 5.8S ribosomal RNA gene and internal transcribed
            spacer 2, complete sequence; and 28S ribosomal RNA gene, partial
            sequence.
ACCESSION   JX024217
VERSION     JX024217.1
KEYWORDS    .
SOURCE      Heterodera humuli
  ORGANISM  Heterodera humuli
            Eukaryota; Metazoa; Ecdysozoa; Nematoda; Chromadorea; Tylenchida;
            Tylenchina; Tylenchoidea; Heteroderidae; Heteroderinae; Heterodera.
REFERENCE   1  (bases 1 to 882)
  AUTHORS   Toumi,F., Waeyenberge,L., Viaene,N., Dababat,A., Nicol,J.M.,
```

```
>JX024217.1 Heterodera humuli isolate MP5 internal transcribed spacer 1, partial
sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2, complete
sequence; and 28S ribosomal RNA gene, partial sequence
ACAGCTGCCTACGAGCATGAGCGTTGGGTGTCTTCCACACGTTGGTGCTGGTATGCTGGCTCTTGTCGT
ACGAGTAAAAGCCGTTGGTTGGCTGTACGGCAGCAGAGGGTTGGTGGCGAACTGCCTTGCTGGTTGGTTT
GCTGCAGCAACGGAGGGAGCACGCTCGTAGGGCAAGCTAACGGATGTGCTGGTGTCTGTGCGTCGTTGAA
CGGTTGTTGTGCCTTGTGCATTACAGACTGACCGGGGACTGTGGTTATGTGTCCTCGGCCTTACGTGCCG
TAACTCGTGGTCTGTGTGTGCTTGGTGCTTGTCCGTGGCCGTGATGAGACGACGTGGTAGTGCCCTTGCT
TAGCGTGAGCATGTGGTTTAAGACTTAATCAGTGCAAGCGTGCACCGCCAGTATTTTTCTTTTTTCATTT
ATTTTTTTACACACTTCCGAGTGAAGAAACAATTCTAGTCTTATCGGTGGATCACTCGGCTCGTGGATCG
ATGAAGAACGCAGCCAACTGCGATAATTAGTGCGAACCGCAGAAACCTTGAACACTAAACATTCGAATGC
ACATTGCGCCATTGGAGTTACATCCATTGGCACGCCTGGTTCAGGGTCGTTATCATAAAATGCACAGCTG
TGCGTGTTGATCTCGTGGGATCATGTCTGTTTGACGTGTTCTTGCGTTATGTTTATGCCCGGCTGTGGGA
TTTGGTTATGCTAGCGCGAAACTGTCAGTTTATTTCGCGTTTTACACACCGTAACTCGGGCGTGGCCATT
TCTCATGAATGCTGTGGTGGAATGCCTAGCATGTGGCGTTCCGTGTTTGAACTTTCGACCTGAACTCAA
CGTGAGTACCCGCTGAACTTAAGCATATCATTTAGCGGAGGA
```

Fig. 15.7. GenBank entry JX024217.

15.5.1 Multiple alignment

The construction of alignment is the establishment of positional homology between nucleotides or amino acid bases that have descended from a common ancestral base. Errors incurred in this step can lead to an incorrect phylogeny. Multiple alignment construction is an exclusively mathematical process and is usually constructed using computer programs with particular algorithms. Most alignments are made based only on sequence information. However, aligning according to secondary structure is considered more reliable than sequence-based alignment because confidence in homology assessment is greater when comparisons are made based on structures rather than on simple characters. This approach is time consuming, requires information on secondary structure and can be done only with special alignment software.

Clustal (http://www.clustal.org) is the most popular multiple sequence alignment program and one of the most widely cited scientific publication (Larkin *et al.*, 2007). There are two main variations: ClustalW with command line interface (‘W’ stands for ‘weighting’ different parts of alignment differently) and ClustalX with a graphical user interface. Alignment can be made using web-based Clustal or software downloaded onto a computer. Clustal performs a global-multiple sequence alignment by the progressive method using a three-step process: (i) perform pairwise alignment of all the sequences by dynamic programming; it aligns each sequence against each other giving a similarity matrix; (ii) create a guide dendrogram using the similarity matrix; and (iii) start by aligning the two most similar sequences. Following the guide dendrogram, the next sequences are added in, aligning them to the existing alignment and insert gaps as necessary.

- Double click on ClustalX icon and run the program. Load sequence the sequence file. Choose **File > Load Sequences.** Under the alignment menu, choose the output format options.
- Under this menu, it is possible to change the alignment parameters (Gap Opening Penalty, Gap Extension Penalty), both for pairwise alignment and for the multiple alignment stages or run with default options.
- Choose **Alignment > Do Complete Alignment** and click **OK.** The sequence alignment is displayed in a window on the screen. The histogram below the ruler indicates the degree of similarity (Fig. 15.8).
- Clustal generates two output files, with extension ‘aln’ (alignment result in Clustal format) and extension ‘dnd’ (guide dendrogram).

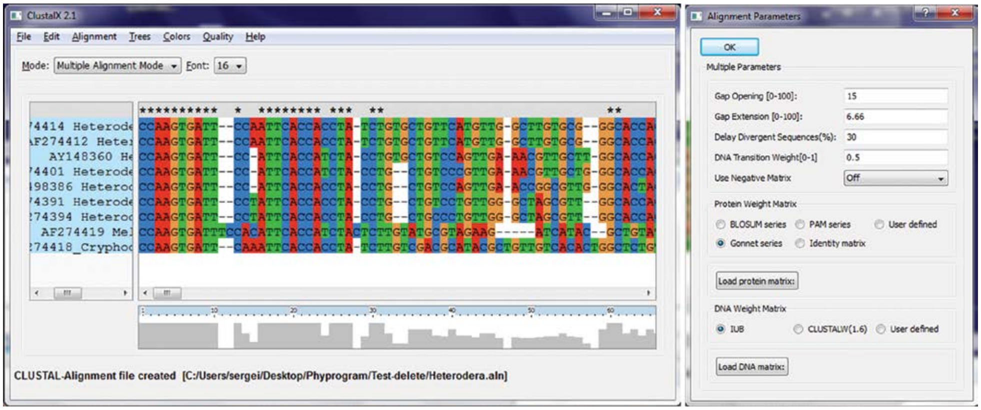
It is recommended to perform phylogenetic analyses based on a series of slightly modified alignments to determine how ambiguous regions in the alignment affect the results.

In addition to Clustal, other software can be used to make alignments (Table 15.1).

15.5.2 Sequence alignment editing

Any automatic alignment should be visually checked and then manually edited, if necessary. GeneDoc (https://genedoc.software.informer.com) is a multiple sequence editor with a full-featured alignment visualization including shading and structural definition features, editing and analysis tools. It has an easy-to-use point and click user interface with extensive keyboard mapping for advanced users (Nicholas *et al.*, 1997).

- **Run Genedoc.** To import the file generated by Clustal, choose **File > New > Import.** In Import Type window, select **Clustal** (**ALN**) and click >**Import.** In Open window, find your file and click> **Open** and then **Done.** The alignment will be displayed (Fig. 15.9).
- To see the full sequence names in the displayed alignment: use combination **Ctrl+G** or choose **Project** > **Configure** and then in the field 'Max NemLen' type '30' > **OK.**
- There are many features to edit alignment. To edit sequence list, use combination **Ctrl+Q** or select **Project > Edit Sequence list.** To move nucleotide(s) or insert gap(s), use **Grab+Drag bottom** (**Ctrl+A**), **Grab+Slide bottom** (F5) Insert/Delete gap in single sequence (F6), **Insert Gap Colum** (F7). To replace nucleotide, use **Ctrl+U** or **Edit** > **Residue Edit Mode.**
- To display the alignment with only nucleotide differences, use the combination **Ctrl+G** or choose **Project** > **Configure,** select **Shade** and in **Residue Display Mode > Differences** and in **Difference Mode Style > Diff/ Top Sequence,** click **OK.**



Mode: Multiple Alignment Mode ▾ Font: 16 ▾

```
               ********** *  ********* ***  **                  **
74414 Heterode CCAAGTGATT--CCAATTCACCACCTA-TCTGTGCTGTTCATGTTG-GCTTGTGCG--GGCACCA
F274412 Heter  CCAAGTGATT--CCAATTCACCACCTA-TCTGTGCTGTTCATGTTG-GCTTGTGCG--GGCACCA
  AY148360 He  CCAAGTGATT--CC-ATTCACCATCTA-CCTGTGCTGTCCAGTTGA-AACGTTGCTT-GGCACCA
74401 Heterode CCAAGTGATT--CC-ATTCACCATCTA-CCTG--CTGTCCCGTTGA-AACGTTGCTG-GGCACCA
98386 Heteroc  CCAAGTGATT--CC-ATTCACCACCTA-CCTG--CTGTCCAGTTGA-ACCGGCGTTG-GGCACTA
74391 Heterode CCAAGTGATT--CCTATTCACCACCTA-CCTG--CTGTCCTGTTGG-GCTAGCGTT--GGCACCA
274394 Heteroc CCAAGTGATT--CCTATTCACCACCTA-CCTG--CTGCCCTGTTGG-GCTAGCGTT--GGCACCA
 AF274419 Mel CCAAGTGATTTCCACATTCACCATCTACTCTTGTATGCGTAGAAG-----ATCATAC--GCTGTA
274418_Cryphoc CCAAGTGATT--CAAATTCACCACCTA-TCTTGTCGACGCATACGCTGTTGTCACACTGGCTCTG
```

1 10 20 30 40 50 60

CLUSTAL-Alignment file created [C:/Users/sergei/Desktop/Phyprogram/Test-delete/Heterodera.aln]

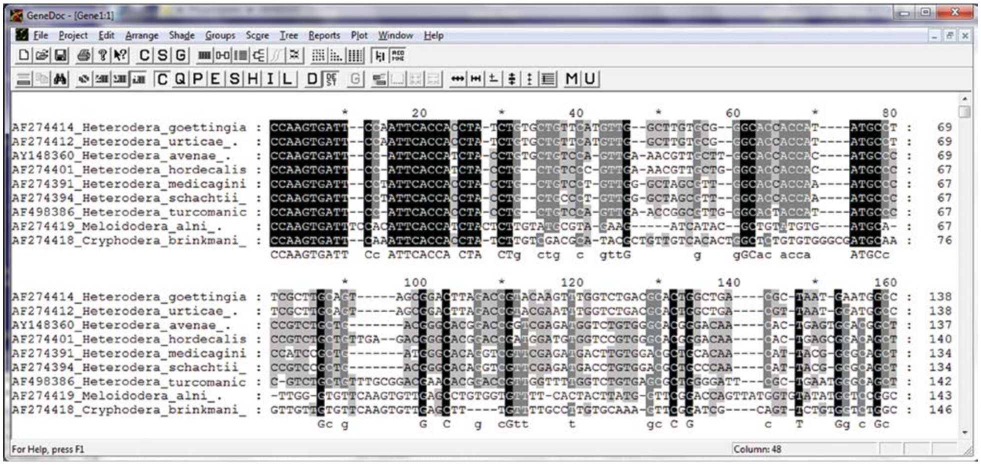
Multiple Parameters

Gap Opening [0-100]: 15
Gap Extension [0-100]: 6.66
Delay Divergent Sequences(%): 30
DNA Transition Weight[0-1] 0.5
Use Negative Matrix: Off ▾

Protein Weight Matrix
○ BLOSUM series ○ PAM series ○ User defined
● Gonnet series ○ Identity matrix

DNA Weight Matrix
● IUB ○ CLUSTALW(1.6) ○ User defined

Fig. 15.8. A multiple sequence alignment generated with ClustalX and alignment parameters.



```
                                      *        20         *        40         *        60         *        80
AF274414_Heterodera_goettingia : CCAAGTGATT--CCAATTCACCACCTA-TCTGTGCTGTTCATGTTG--GCTTGTGCG--GGCACCACCAT----ATGCCT :   69
AF274412_Heterodera_urticae_.  : CCAAGTGATT--CCAATTCACCACCTA-TCTGTGCTGTTCATGTTG--GCTTGTGCG--GGCACCACCAT----ATGCCT :   69
AY148360_Heterodera_avenae_.   : CCAAGTGATT--CC-ATTCACCATCTA-CCTGTGCTGTCCA-GTTGA-AACGTTGCTT-GGCACCACCAC----ATGCCC :   69
AF274401_Heterodera_hordecalis : CCAAGTGATT--CC-ATTCACCATCTA-CCTG--CTGTCCC-GTTGA-AACGTTGCTG-GGCACCACCAC----ATGCCC :   67
AF274391_Heterodera_medicagini : CCAAGTGATT--CCTATTCACCACCTA-CCTG--CTGTCCT-GTTGG-GCTAGCGTT--GGCACCACCAA----ATGCCC :   67
AF274394_Heterodera_schachtii_ : CCAAGTGATT--CCTATTCACCACCTA-CCTG--CTGCCCT-GTTGG-GCTAGCGTT--GGCACCACCAA----ATGCCC :   67
AF498386_Heterodera_turcomanic : CCAAGTGATT--CC-ATTCACCACCTA-CCTG--CTGTCCA-GTTGA-ACCGGCGTTG-GGCACTACCAT----ATGCCC :   67
AF274419_Meloidodera_alni_.    : CCAAGTGATTTCCACATTCACCATCTACTCTTGTATGCGTA-GAAG-----ATCATAC--GCTGTATGTG----ATGCA- :   67
AF274418_Cryphodera_brinkmani_ : CCAAGTGATT--CAAATTCACCACCTA-TCTTGTCGACGCA-TACGCTGTTGTCACACTGGCTCTGTGTGGGCGATGCAA :   76
                                 CCAAGTGATT  Cc ATTCACCA CTA  CTg  ctg  c  gttG        g    gGCac acca     ATGCc

                                      *       100         *       120         *       140         *       160
AF274414_Heterodera_goettingia : TCGCTTGCACT-----AGCGGACTTACACCGTACAAGTTTGGTCTGACGCACTGGCTGA----GGC-TAAT-GAATGGCC :  138
AF274412_Heterodera_urticae_.  : TCGCTTGCACT-----AGCGGACTTACACCGTACGAATTTGGTCTGACGCACTGGCTGA----GGT-TAAT-GCATGGCC :  138
AY148360_Heterodera_avenae_.   : CCGTCTGCTC-------ACGGGCACGCACCGGTCGAGATGGTCTGTGGGCACGGGACAA----GAC-TGAGTGCAGGCCT :  137
AF274401_Heterodera_hordecalis : CCGTCTGCTCTTGA--GACGGGCACGCACCGATGGATGTGGTCCGTGGGCACGGGACAA----GAC-TGAGCGCAGAGCT :  140
AF274391_Heterodera_medicagini : CCATCCGCTC-------ATGGGCACACGTCGTTCGAGATGACTTGTGGACGCTGCACAA----GAT-TACG-GCGCAGCT :  134
AF274394_Heterodera_schachtii_ : CCGTCCGCTC-------ACGGGCACACGTCGTTCGAGATGACCTGTGGACGCTGCCCAA----GAT-TACG-GCGCAGCT :  134
AF498386_Heterodera_turcomanic : C-GTCTGCTCTTTGCGGACGAACACGCACCGTTGGTTTTGGTCTGTGAGCGCTGGGGATT---GGC-TGAATGCGCAGCT :  142
AF274419_Meloidodera_alni_.    : -TTGG-GTGTTCAAGTGTTGAGCCTGTGGTGTTTT-CACTACTTATG-GTTCGGACCAGTTATGGTGTATATGCTCCGGC :  143
AF274418_Cryphodera_brinkmani_ : GTTGTTGTGTTCAAGTGTTGAGCTT----TGTTTTGCCTTGTGCAAA-GTTCGGATCG----CAGT-TCTGTGCTCTGGC :  146
                                         Gc g           G  C   g  cGtt      t          gc C G          c   T     Gg c Gc
```

For Help, press F1 Column: 48

Fig. 15.9. A multiple sequence alignment in GenDoc.

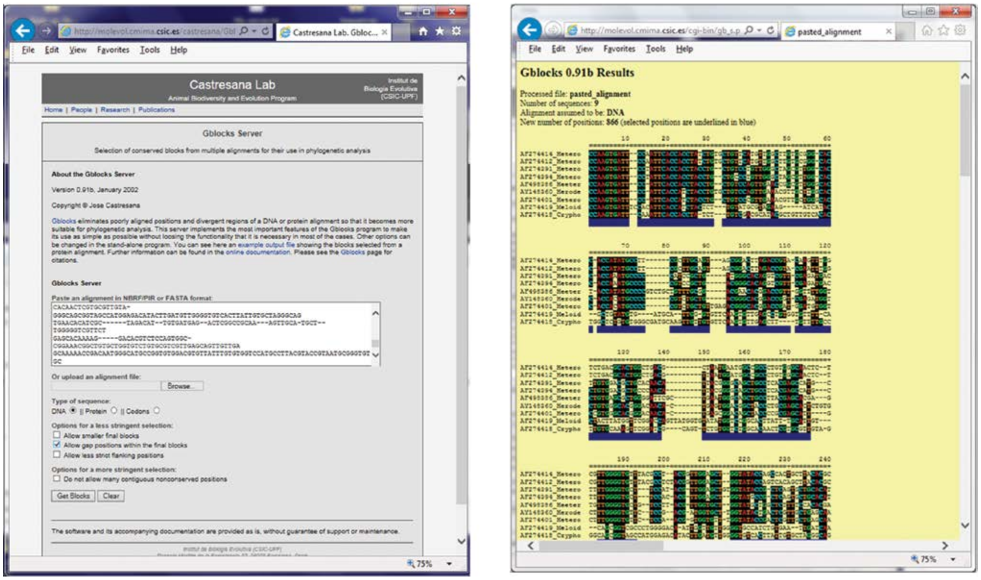
- To trim excess sequence in the beginning and end of alignment or remove any alignment region: use combination **Ctrl+L** or click **Edit,** select **> Select Columns** and then use the cursor to select alignment region. To delete this region: use combination **Ctrl+D,** or click **Edit,** select **> Delete All Data.**
- To prepare the alignment file for a publication, use combination **Ctrl+E** or **Edit > Select Blocks for Copy,** and use a cursor to mark block(s). When all blocks are marked, click **Edit > Copy Selected Blocks** to **> RTF file,** then **Save as.** The saved file could be opened in Word.
- To save the file use **Ctrl+S** or **File > Save.** Export the corrected GenDoc file into a FASTA format, choose **File > Export,** then type the file name, click on **Save > Done**. The result will be saved.

Multiple alignment can also be automatically edited using the computer program Gblocks (http://molevol.cmima.csic.es/castresana/Gblocks.html) that eliminates poorly aligned positions and divergent regions of an alignment of DNA or protein sequences (Castresana, 2000; Talavera and Castresana, 2007). These positions may not be homologous or may have been saturated by multiple substitutions and it is convenient to eliminate them prior to phylogenetic analysis. Gblocks selects blocks in a similar way as done by hand but following a reproducible set of conditions. The selected blocks must fulfil certain requirements with respect to the lack of large segments of contiguous non-conserved positions, lack of gap positions and high conservation of flanking positions, making the final alignment more suitable for phylogenetic analysis. Gblocks is very fast in processing alignments and it is therefore highly suitable for large-scale phylogenetic analyses. Gblocks can be run on a computer and run on web: http://molevol.cmima.csic.es/castresana/Gblocks_server.html.

- Go to Gblocks web-server. Insert the alignment in a FASTA format in a window and select options.
- Click on **Get Block** bottom. Gblocks outputs files to visualize the selected blocks (Fig. 15.10).
- Click on **Resulting alignment** to see the result.

15.6 File Format Converting

Each phylogenetic program requires an alignment prepared in certain file format. Several popular phylogenetic programs such as PAUP*, MrBayes, Mesquite, MacClade and others use the NEXUS format widely used in bioinformatics. ForCon is a user-friendly software tool (http://bioinformatics.psb.ugent.be/webtools/ForCon/)


http://molevol.cmima.csic.es/castresana/Gbl
Castresana Lab. Gbloc...
File
Edit
View
Favorites
Tools
Help
Home | People | Research | Publications
Browse...
Get Blocks
Clear
75%
http://molevol.cmima.csic.es/cgi-bin/gb_s.p
pasted_alignment
75%

Castresana Lab

Animal Biodiversity and Evolution Program

Institut de Biologia Evolutiva (CSIC-UPF)

Gblocks Server

Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis

About the Gblocks Server

Version 0.91b, January 2002

Copyright © Jose Castresana

Gblocks eliminates poorly aligned positions and divergent regions of a DNA or protein alignment so that it becomes more suitable for phylogenetic analysis. This server implements the most important features of the Gblocks program to make its use as simple as possible without loosing the functionality that it is necessary in most of the cases. Other options can be changed in the stand-alone program. You can see here an example output file showing the blocks selected from a protein alignment. Further information can be found in the online documentation. Please see the Gblocks page for citations.

Gblocks Server

Paste an alignment in NBRF/PIR or FASTA format:

```
CACAACTCGTGCGTTGTA-
GGGCAGCGGTAGCCATGGAGACATACTTGATGTTGGGGTGTCACTTATTGTGCTAGGGCAG
TGAACACATCGC------TAGACAT--TGTGATGAG--ACTCGGCCGCAA---AGTTGCA-TGCT--
TGGGGGTCGTTCT
GAGCACAAAAG-----GACACGTCTCCAGTGGC-
CGGAAACGGCTGTGCTGGTGTCTGTGCGTCGTTGAGCAGTTGTTGA
GCAAAAACCGACAATGGGCATGCCGGTGTGGACGTGTTATTTGTGTGGTCCATGCCTTACGTACCGTAATGCGGGTGT
GC
```

Or upload an alignment file:

Type of sequence:
DNA ● || Protein ○ || Codons ○

Options for a less stringent selection:
☐ Allow smaller final blocks
☑ Allow gap positions within the final blocks
☐ Allow less strict flanking positions

Options for a more stringent selection:
☐ Do not allow many contiguous nonconserved positions

The software and its accompanying documentation are provided as is, without guarantee of support or maintenance.

Institut de Biologia Evolutiva (CSIC-UPF)

Gblocks 0.91b Results

Processed file: **pasted_alignment**
Number of sequences: **9**
Alignment assumed to be: **DNA**
New number of positions: **866** (selected positions are underlined in blue)

```
AF274414_Hetero
AF274412_Hetero
AF274391_Hetero
AF274394_Hetero
AF498286_Heeter
AY145360_Herode
AF274401_Hetero
AF274419_Meloid
AF274418_Crypho
```

Fig. 15.10. Gblock Server home page and output.

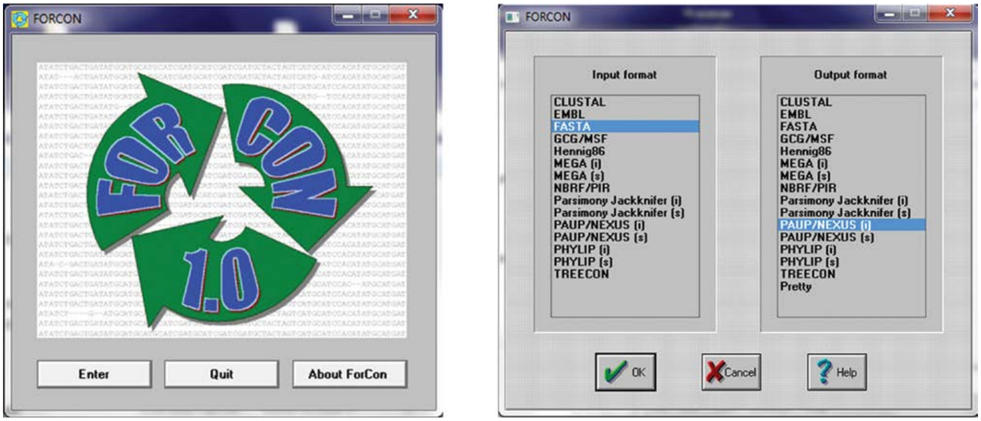


Fig. 15.11. ForCon windows.

for the conversion of nucleic acid and amino acid sequence alignments. ForCon is able to convert in both ways, i.e. reading and writing.

- Run ForCon program. Choose **Enter,** then select Input format > **FASTA** and for Output format >**PAUP/NEXUS(i)** > OK (Fig. 15.11).
- In Open window, find your file in FASTA format > **Open** > **OK.** In **Save as** window, **Type** file name with file extension 'nex' > **Save.** Then choose **Select All > OK.**

The file conversion could also be done via different on-line tools, for example, http://www.ebi.ac.uk/Tools/sfc/emboss_seqret/ and https://hcv.lanl.gov/content/sequence/FORMAT_CONVERSION/form.html.

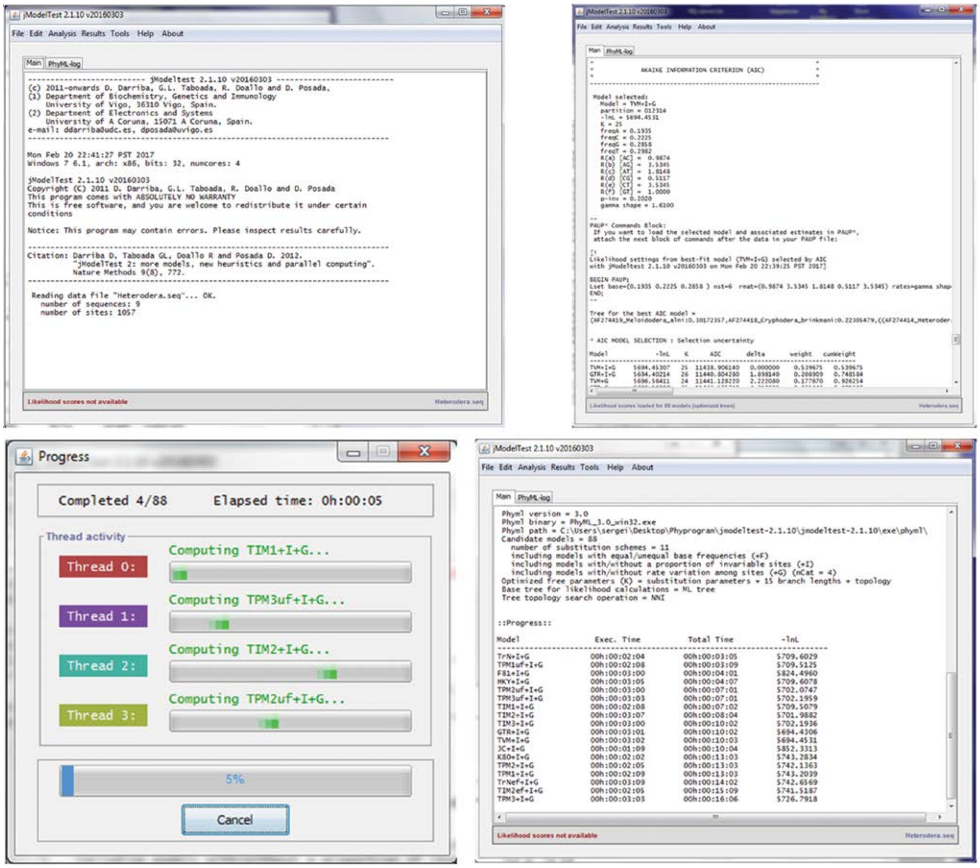
15.7 Selection of Model of Sequence Evolution

When using minimum evolution (distance), maximum likelihood methods or Bayesian inference to build trees, it needs to find a model of sequence evolution that fits the DNA changes in the aligned sequences that are being used. The substitution models differ in terms of the parameters used to describe the rates at which one nucleotide replaces another during evolution. The best-fit substitution model can be selected using the Modeltest program (Posada and Crandall, 1998), MrModeltest (Nylander, 2004) or jModelTest (Darriba *et al.*, 2012).

15.7.1 jModelTest: A tool to select the best-fit model of nucleotide substitution

jModelTest (http://code.google.com/p/jmodeltest2/) is a tool to carry out statistical selection of best-fit models of nucleotide substitution evolution that best fits the data and to use in constructing phylogenetic trees in PAUP* or MrBayes. It implements five different model selection strategies: hierarchical and dynamical likelihood ratio tests (hLRT and dLRT), Akaike and Bayesian information criteria (AIC and BIC), and a decision theory method. The jModelTest program is described by Posada (2008) and Darriba *et al.* (2012).

- Double click on jModelTest.jar to open it.
- Go **File** > **Load DNA alignment** and open the data set file (Fig. 15.12).
- Click on **Analysis** > **Compute likelihood.** A dialog box will appear that allows you to specify a number of likelihood settings, including the number of substitution schemes to be tested. Click on **Compute Likelihood** to start the analysis.
- When Likelihood score calculation is completed, click on **Analysis** and select either **Do AIC calculations** or **Do BIC calculations.** In the setting window, check a box for **Write PAUP* block** and click on **Do AIC calculation.**
- The best-fit model is displayed with PAUP* block, which can be added to the PAUP* input file. The results can be saved with **Ctrl+S.**


jModelTest 2.1.10 v20160303
File Edit Analysis Results Tools Help About
Main
PhyML-log
Likelihood scores not available
Heterodera.seq

```
-------------------------- jModeltest 2.1.10 v20160303 --------------------------
(c) 2011-onwards D. Darriba, G.L. Taboada, R. Doallo and D. Posada,
(1) Department of Biochemistry, Genetics and Immunology
    University of Vigo, 36310 Vigo, Spain.
(2) Department of Electronics and Systems
    University of A Coruna, 15071 A Coruna, Spain.
e-mail: ddarriba@udc.es, dposada@uvigo.es
--------------------------------------------------------------------------------

Mon Feb 20 22:41:27 PST 2017
Windows 7 6.1, arch: x86, bits: 32, numcores: 4

jModelTest 2.1.10 v20160303
Copyright (C) 2011 D. Darriba, G.L. Taboada, R. Doallo and D. Posada
This program comes with ABSOLUTELY NO WARRANTY
This is free software, and you are welcome to redistribute it under certain
conditions

Notice: This program may contain errors. Please inspect results carefully.

--------------------------------------------------------------------------------
Citation: Darriba D, Taboada GL, Doallo R and Posada D. 2012.
          "jModelTest 2: more models, new heuristics and parallel computing".
          Nature Methods 9(8), 772.
--------------------------------------------------------------------------------

 Reading data file "Heterodera.seq"... OK.
   number of sequences: 9
   number of sites: 1057
```

jModelTest 2.1.10 v20160303
File Edit Analysis Results Tools Help About
Main
PhyML-log
Likelihood scores loaded for 88 models (optimized trees)
Heterodera.seq

```
*                                                                  *
*              AKAIKE INFORMATION CRITERION (AIC)                  *
*                                                                  *
--------------------------------------------------------------------

 Model selected:
   Model = TVM+I+G
   partition = 012314
   -lnL = 5694.4531
   K = 25
   freqA = 0.1935
   freqC = 0.2225
   freqG = 0.2858
   freqT = 0.2982
   R(a) [AC] =  0.9874
   R(b) [AG] =  3.5345
   R(c) [AT] =  1.8148
   R(d) [CG] =  0.5117
   R(e) [CT] =  3.5345
   R(f) [GT] =  1.0000
   p-inv = 0.2020
   gamma shape = 1.6100

--
PAUP* Commands Block:
 If you want to load the selected model and associated estimates in PAUP*,
 attach the next block of commands after the data in your PAUP file:

[!
Likelihood settings from best-fit model (TVM+I+G) selected by AIC
with jModeltest 2.1.10 v20160303 on Mon Feb 20 22:39:25 PST 2017]

BEGIN PAUP;
Lset base=(0.1935 0.2225 0.2858 ) nst=6  rmat=(0.9874 3.5345 1.8148 0.5117 3.5345) rates=gamma shap
END;
--

Tree for the best AIC model =
(AF274419_Meloidodera_alni:0.30172357,AF274418_Cryphodera_brinkmani:0.22305479,((AF274414_Heteroder

* AIC MODEL SELECTION : Selection uncertainty

Model             -lnL      K       AIC         delta        weight    cumWeight
--------------------------------------------------------------------------------
TVM+I+G       5694.45307   25  11438.906140   0.000000     0.539675   0.539675
GTR+I+G       5694.40214   26  11440.804280   1.898140     0.208909   0.748584
TVM+G         5696.56411   24  11441.128220   2.222080     0.177670   0.926254
```

```
Completed 4/88        Elapsed time: 0h:00:05

Computing TIM1+I+G...
Computing TPM3uf+I+G...
Computing TIM2+I+G...
Computing TPM2uf+I+G...
```

jModelTest 2.1.10 v20160303
File Edit Analysis Results Tools Help About
Main
PhyML-log
Likelihood scores not available
Heterodera.seq

```
 Phyml version = 3.0
 Phyml binary = PhyML_3.0_win32.exe
 Phyml path = C:\Users\sergei\Desktop\Phyprogram\jmodeltest-2.1.10\jmodeltest-2.1.10\exe\phyml\
 Candidate models = 88
   number of substitution schemes = 11
   including models with equal/unequal base frequencies (+F)
   including models with/without a proportion of invariable sites (+I)
   including models with/without rate variation among sites (+G) (nCat = 4)
 Optimized free parameters (K) = substitution parameters + 15 branch lengths + topology
 Base tree for likelihood calculations = ML tree
 Tree topology search operation = NNI


::Progress::

Model                 Exec. Time          Total Time          -lnL
-------------------------------------------------------------------------
TrN+I+G             00h:00:02:04        00h:00:03:05        5709.6029
TPM1uf+I+G          00h:00:02:08        00h:00:03:09        5709.5125
F81+I+G             00h:00:03:00        00h:00:04:01        5824.4960
HKY+I+G             00h:00:03:05        00h:00:04:07        5709.6078
TPM2uf+I+G          00h:00:03:00        00h:00:07:01        5702.0747
TPM3uf+I+G          00h:00:03:03        00h:00:07:01        5702.1959
TIM1+I+G            00h:00:02:08        00h:00:07:02        5709.5079
TIM2+I+G            00h:00:03:07        00h:00:08:04        5701.9882
TIM3+I+G            00h:00:03:00        00h:00:10:02        5702.1936
GTR+I+G             00h:00:03:01        00h:00:10:02        5694.4306
TVM+I+G             00h:00:03:02        00h:00:10:03        5694.4531
JC+I+G              00h:00:01:09        00h:00:10:04        5852.3313
K80+I+G             00h:00:02:02        00h:00:13:03        5743.2834
TPM2+I+G            00h:00:02:05        00h:00:13:03        5742.1363
TPM1+I+G            00h:00:02:09        00h:00:13:03        5743.2039
TrNef+I+G           00h:00:03:09        00h:00:14:02        5742.6569
TIM2ef+I+G          00h:00:02:05        00h:00:15:09        5741.5187
TPM3+I+G            00h:00:03:03        00h:00:16:06        5726.7918
```

Fig. 15.12. jModelTest windows.

The widely used General Time Reversible (GTR) model has six substitution types (**lset nst=6**), one for each pair of nucleotides and it is the most complex model. In addition to models describing the rates of change from one nucleotide to another, there are models to describe rate variation among sites in a sequence. The following are the two most commonly used models: (i) gamma distribution (G) or gamma distributed rate variation among sites proportion of invariable sites; and (ii) (I) or extent of static, unchanging sites in a dataset.

15.8 Phylogenetic Analysis with PAUP*

Once the sequences are aligned, there are several methods of phylogenetic analysis that can be implemented. The main methods include minimum evolution, parsimony and likelihood methods, and Bayesian inference.

Minimum evolution (distance) calculates a measure of the distance between each pair of species and then finds a tree that predicts the observed set of distances as closely as possible. A topology showing the smallest value of the sum of all branches is chosen as an estimate of the correct tree.

Maximum parsimony (MP) is a character-based method that builds a phylogenetic tree by minimizing the total tree length. It searches for the minimum number of evolutionary steps required for a given set of data.

Maximum likelihood (ML) is a statistical method for reconstructing trees. It requires three elements: (i) sequence alignment; (ii) tree topologies; and (iii) model of character evolution. ML operates by trying to maximize the likelihood value, and the tree with the highest likelihood value is considered the best tree.

These three approaches can be performed using PAUP* software (Swofford, 2003) that is available from Sinauer Associates Inc. Publishers, Sunderland, Massachusetts, USA. PAUP* (Phylogenetic Analysis Using

Parsimony* and other methods) (http://paup.phylosolutions.com) performs phylogenetic analyses using parsimony, maximum likelihood and distance methods. The program features an extensive selection of analysis options and model choices, and accommodates DNA, RNA, protein and general data types.

The Windows versions of PAUP* require data and commands, which should be typed in NEXUS format. PAUP* is run by entering the command blocks. The PAUP* manual with command explanation is given in the PDF documentation in the PAUP* folder: http://paup.phylosolutions.com/tutorials/quick-start/ The PDF command summary can be found in the **Quick Start tutorial and PAUP* FAQ: Answers** pages.

NEXUS data files always begin with the characters #nexus but are otherwise organized into major units known as blocks. The necessary commands can be put in a PAUP* block in the original nexus file after **Assumption block** in any text editor or in PAUP* using **Edit File Open Mode**. Each command begins with a command-name and ends with a semicolon. NEXUS files can contain text comments surrounded by square brackets. The following sections present several examples of the command blocks for distance, parsimony, and likelihood analyses.

15.8.1 Distance method

A. Command block for reconstruction of neighbour-joining tree:

```
begin paup; [start PAUP running]
log file = NJ.log; [command starts a log file]
set criterion = distance; [command defines the optimality criterion]
outgroup 8; [command specifies that the resulting trees should be rooted to given taxon]
bootstrap nreps = 1000 search = nj; [commands specify bootstrap number and method of search]
nj; [command calculates a tree using the neighbor-joining method]
showtrees; [command to request to display one or more trees]
savetrees file = NJ.tre brlens = yes root = yes; [commands save the best rooted tree
  found during the search with branch length information in a file]
log stop; [command stops the logging of output]
end; [stop PAUP running]
```

B. Command block for calculation of nucleotide differences

```
begin paup;
log file = distance.log;
dset distance = mean; [command gives mean number of pairwise character differences,
  adjusted for missing data]
BaseFreq; [command shows base frequencies for each taxon]
showdist; [command shows output a matrix of ‘distances’ between taxa in a PAUP window]
log stop;
end;
```

15.8.2 Maximum parsimony

A. Command block for reconstruction of maximum parsimony trees

```
begin paup;
log file = MP.log;
set increase = auto; [setting automatically be increased by a number of trees equal
  to the default number 100, could be changed into ‘no’, if search takes long time]
set autoclose; [closes the status window automatically]
set criterion = parsimony;
outgroup 8;
bootstrap nreps = 1000 search = heuristic;
hsearch nreps = 10 addseq=random; [search for optimal trees using heuristic algorithm
  with 10 replicates using random-addition-sequence replications to be performed]
showtrees;
describetrees /apolist = yes; [command produces a depiction of the tree and set a
  list of the apomorphic characters is displayed]
```

```
savetrees file = MP.tre brlens = yes root = yes;
log stop;
end;
```

B. Command block for obtaining a strict consensus tree

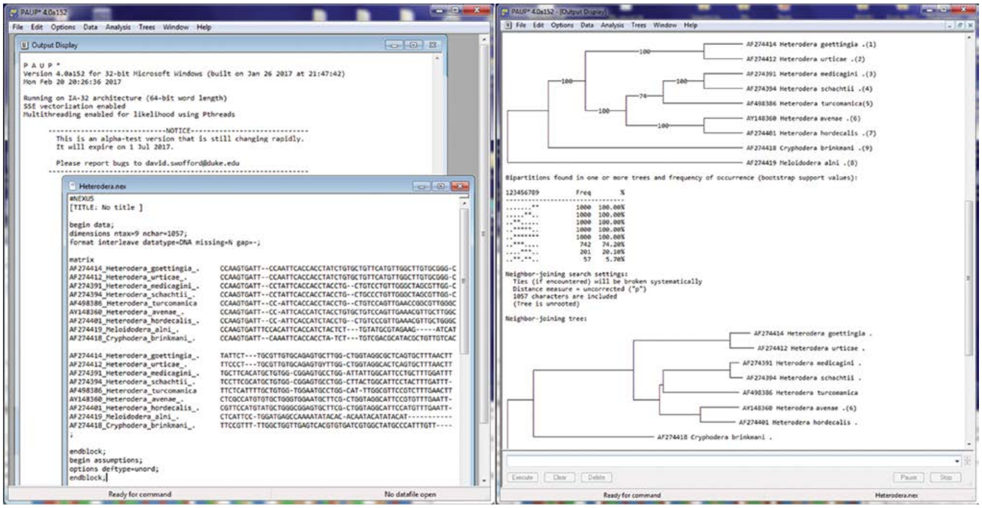
```
begin paup;
log file = conMP.log;
gettrees file = MP.tre; [command to load trees into memory from a file]
outgroup 8;
contree / root = outgroup; [root a strict consensus tree]
showtrees;
contree / treefile = conMP.tre; [save strict consensus file]
log stop;
end;
```

15.8.3 Maximum likelihood

A. Command block for reconstruction of maximum likelihood trees

```
begin paup;
log file = ML.log;
set autoclose;
set criterion = likelihood;
Lset base=(0.1943 0.2232 0.2846) nst=6 rmat=(0.9577 3.5283 1.7964 0.5168 3.5283) rates=gamma
  shape=0.7150 ncat=4 pinvar=0; [model parameters obtained from jModelTest results]
bootstrap nreps = 100 search = faststep; [bootstrap with tree searches in each replication
  are performed using one random-sequence-addition replication and no branch swapping]
hsearch;
savetrees file = ML.tre brlens = yes root = yes;
log stop;
end;
```

- Double click on PAUP icon and run the program. Change **File Open Mode** to **Edit**. Select the executable file and click on **Open** (Fig. 15.13).
- Insert command block or blocks after assumption block in the file.


PAUP* 4.0a152
File Edit Options Data Analysis Trees Window Help
Output Display
Heterodera.nex
Ready for command
No datafile open
PAUP* 4.0a152 - [Output Display]
100
100
100
74
100
100
AF274414 Heterodera goettingia .(1)
AF274412 Heterodera urticae .(2)
AF274391 Heterodera medicagini .(3)
AF274394 Heterodera schachtii .(4)
AF498386 Heterodera turcomanica(5)
AY148360 Heterodera avenae .(6)
AF274401 Heterodera hordecalis .(7)
AF274418 Cryphodera brinkmani .(9)
AF274419 Meloidodera alni .(8)
AF274414 Heterodera goettingia .
AF274412 Heterodera urticae .
AF274391 Heterodera medicagini .
AF274394 Heterodera schachtii .
AF498386 Heterodera turcomanica
AY148360 Heterodera avenae .(6)
AF274401 Heterodera hordecalis .
AF274418 Cryphodera brinkmani .
Execute
Clear
Delete
Pause
Stop
Ready for command
Heterodera.nex

```
P A U P *
Version 4.0a152 for 32-bit Microsoft Windows (built on Jan 26 2017 at 21:47:42)
Mon Feb 20 20:26:36 2017

Running on IA-32 architecture (64-bit word length)
SSE vectorization enabled
Multithreading enabled for likelihood using Pthreads

     -----------------------------NOTICE-----------------------------
       This is an alpha-test version that is still changing rapidly.
       It will expire on 1 Jul 2017.

       Please report bugs to david.swofford@duke.edu
     ----------------------------------------------------------------
```

```
#NEXUS
[TITLE: No title ]

begin data;
dimensions ntax=9 nchar=1057;
format interleave datatype=DNA missing=N gap=-;

matrix
AF274414_Heterodera_goettingia_.    CCAAGTGATT--CCAATTCACCACCTATCTGTGCTGTTCATGTTGGCTTGTGCGGG-C
AF274412_Heterodera_urticae_.       CCAAGTGATT--CCAATTCACCACCTATCTGTGCTGTTCATGTTGGCTTGTGCGGG-C
AF274391_Heterodera_medicagini_.    CCAAGTGATT--CCTATTCACCACCTACCTG--CTGTCCTGTTGGGCTAGCGTTGG-C
AF274394_Heterodera_schachtii_.     CCAAGTGATT--CCTATTCACCACCTACCTG--CTGCCCTGTTGGGCTAGCGTTGG-C
AF498386_Heterodera_turcomanica     CCAAGTGATT--CC-ATTCACCACCTACCTG--CTGTCCAGTTGAACCGGCGTTGGGC
AY148360_Heterodera_avenae_.        CCAAGTGATT--CC-ATTCACCATCTACCTGTGCTGTCCAGTTGAAACGTTGCTTGGC
AF274401_Heterodera_hordecalis_.    CCAAGTGATT--CC-ATTCACCATCTACCTG--CTGTCCCGTTGAAACGTTGCTGGGC
AF274419_Meloidodera_alni_.         CCAAGTGATTTCCACATTCACCATCTACTCT---TGTATGCGTAGAAG-----ATCAT
AF274418_Cryphodera_brinkmani_.     CCAAGTGATT--CAAATTCACCACCTA-TCT---TGTCGACGCATACGCTGTTGTCAC

AF274414_Heterodera_goettingia_.    TATTCT---TGCGTTGTGCAGAGTGCTTGG-CTGGTAGGCGCTCAGTGCTTTAACTT
AF274412_Heterodera_urticae_.       TTCCCT---TGCGTTGTGCAGAGTGYTTGG-CTGGTAGGCACTCAGTGCTTTAACTT
AF274391_Heterodera_medicagini_.    TGCTTCACATGCTGTGG-CGGAGTGCCTGG-ATTATTGGCATTCCTGCTTTGGATTT
AF274394_Heterodera_schachtii_.     TCCTTCGCATGCTGTGG-CGGAGTGCCTGG-CTTACTGGCATTCCTACTTTGATTT-
AF498386_Heterodera_turcomanica     TTCTCATTTTGCTGTGG-TGGAATGCCTGG-CAT-TTGGCGTTCCGTCTTTGAACTT
AY148360_Heterodera_avenae_.        CTCGCCATGTGTGCTGGGTGGAATGCTTCG-CTGGTAGGCATTCCGTGTTTGAATT-
AF274401_Heterodera_hordecalis_.    CGTTCCATGTATGCTGGGCGGAGTGCTTCG-CTGGTAGGCATTCCATGTTTGAATT-
AF274419_Meloidodera_alni_.         CTCATTCC-TGGATGAGCCAAAATATACAC-ACAATACATATACAT----------
AF274418_Cryphodera_brinkmani_.     TTCCGTTT-TTGGCTGGTTGAGTCACGTGTGATCGTGGCTATGCCCATTTGTT----
;

endblock;
begin assumptions;
options deftype=unord;
endblock;
```

```
Bipartitions found in one or more trees and frequency of occurrence (bootstrap support values):

123456789        Freq       %
--------------------------------
.......**        1000  100.00%
.....**..        1000  100.00%
..**.....        1000  100.00%
..*****..        1000  100.00%
..*******        1000  100.00%
..***....         742   74.20%
....***..         201   20.10%
..**.**..          57    5.70%

Neighbor-joining search settings:
  Ties (if encountered) will be broken systematically
  Distance measure = uncorrected ("p")
  1057 characters are included
  (Tree is unrooted)

Neighbor-joining tree:
```

Fig. 15.13. PAUP* with data matrix and outputs.

- Open **File** and then select **Execute** 'file.extension'.
- PAUP performs analysis.

PAUP generates two files: (i) a file with extension**.log** containing information on search, resulted bootstrap tree and table and optimal tree, which can be opened in any text Editor program; and (ii) a file with extension.tre with resulting tree, which can be visualized with the TreeView.

15.9 Phylogenetic Analysis with MrBayes

MrBayes is a program for the Bayesian estimation of phylogeny (http://mrbayes.scs.fsu.edu/). Bayesian inference (BI) of phylogeny is based upon a quantity called the posterior probability distribution of trees, which is the probability of a tree conditioned on the observations. The conditioning is accomplished using Bayes's theorem. MrBayes uses a simulation technique called Markov chain Monte Carlo (or MCMC) to approximate the posterior probabilities of trees (Ronquist and Huelsenbeck, 2003; Ronquist *et al.*, 2012).

- Open your nexus format file in any text editor program.
- Type the following command block with appropriate model (model GTR+G+I is used as an example) after the data block.
  Command block for reconstruction of BI trees:

```
begin mrbayes; [command to start MrBayes]
log start filename = BIfile.log; [command to create the log file under certain name]
outgroup 8; [command specifies that the resulting trees should be rooted to given taxon]
lset nst=6 rates=invgamma; [commands to set parameters for the model]
showmodel; [command to show model settings]mcmc ngen=1000000 printfreq=100
  samplefreq=100 nchains=4 savebrlens=yes filename=BItree; [commands to run the Markov
  chain with 1000000 generations, print the results to the screen every 100 generations,
  record the current tree and parameters values to files every 100 generations, run 4
  independent chains, save the trees with branch lengths in the tree file]
sumt filename=BItree burnin=1000 contype=halfcompat; [commands to summarize the saved
  trees, discard the first 1000 trees or 10% of total saved trees and write a majority
  rule consensus tree in BItree.con]
sump filename=BItree burnin=1000; [commands to summarize statistics for trees sampled
  during analysis]
log stop; [command to quit recording in the log file]
end;
```

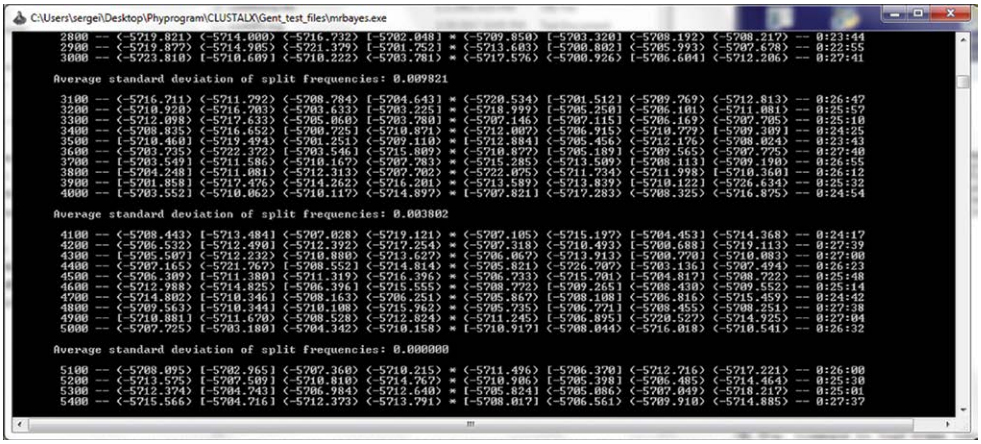
  Command to set models: JC - nst = 1; K2, HKY, T3P, TN93 - nst = 2; GTR - nst = 6 and rates: Uniform - rates = equal; Gamma Distributed (G) - rates = gamma; Invariant Sites (I) - rates = propinv; G + I - rates = invgamma

- **Save** the file. Copy MrBayes in the same directory.
- Double click on MrBayes icon and run program. Type the following command: **execute** and after a space type **file name with extension.** Click > **Enter**

The program runs (Fig. 15.14). Each result line shows the generation number at the left, then a series of four numbers, each enclosed in brackets. Those numbers are the log likelihood of the trees in each of the four chain in run 1. The cold chain is enclosed in square brackets. After symbol * – chain is in run 2. The last number in each line of output to the screen is the estimated time until the run ends (Figs 15.15 and 15.16). If the average standard deviation of split frequencies is stabilized for many generations below 0.01, the run can be stopped. The results of BI analysis are saved in several files: BIfile.log, BItree.mcmc, BItree.run1.p, BItree.run2.p, BItree.run1.t, BItree.run2.t, BItree.con.

The BItree.con file can be opened in TreeView to visualize the tree.

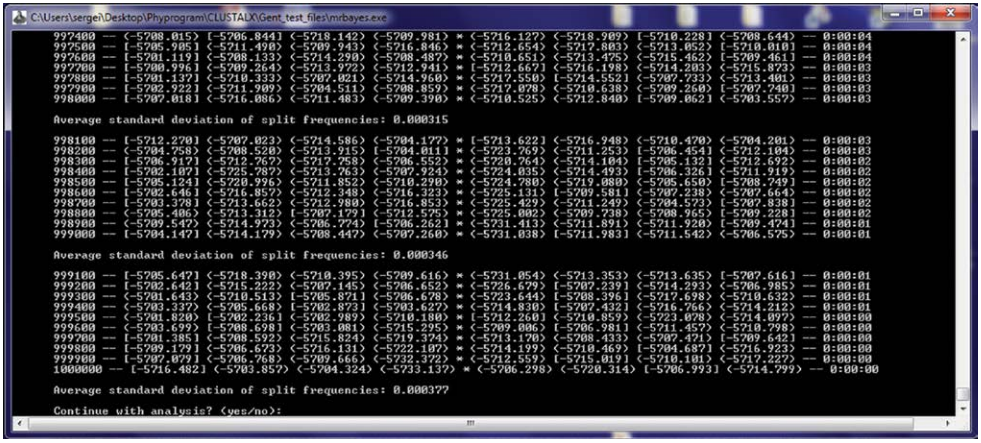
Several other software applications can be used to reconstruct phylogenetic relationships: RAxML (https://sco.h-its.org/exelixis/web/software/raxml/) and BEAST (http://tree.bio.ed.ac.uk/software/beast) and MEGA (https://www.megasoftware.net).



C:\Users\sergei\Desktop\Phyprogram\CLUSTALX\Gent_test_files\mrbayes.exe

```
   2800 -- (-5719.821) (-5714.000) (-5716.732) [-5702.048] * (-5709.850) [-5703.320] (-5708.192) (-5708.217) -- 0:23:44
   2900 -- (-5719.877) (-5714.905) (-5721.379) [-5701.752] * (-5713.603) [-5700.802] (-5705.993) (-5707.678) -- 0:22:55
   3000 -- (-5723.810) [-5710.609] (-5710.222) (-5703.781) * (-5717.576) (-5700.926) [-5706.604] (-5712.206) -- 0:27:41

  Average standard deviation of split frequencies: 0.009821

   3100 -- (-5716.711) (-5711.792) (-5708.784) [-5704.643] * (-5720.534) [-5701.512] (-5709.769) (-5712.813) -- 0:26:47
   3200 -- (-5710.920) (-5716.703) (-5703.633) [-5703.225] * (-5718.999) [-5705.250] (-5706.101) (-5711.081) -- 0:25:57
   3300 -- (-5712.098) (-5717.633) (-5705.060) [-5703.780] * (-5707.146) [-5707.115] (-5706.169) (-5707.705) -- 0:25:10
   3400 -- (-5708.835) (-5716.652) [-5700.725] (-5710.871) * (-5712.007) (-5706.915) (-5710.779) [-5709.309] -- 0:24:25
   3500 -- [-5710.460] (-5719.494) (-5701.251) (-5709.110) * [-5712.884] (-5705.456) (-5712.176) (-5708.024) -- 0:23:43
   3600 -- (-5703.735) (-5722.372) [-5703.546] (-5715.809) * (-5710.877) [-5705.189] (-5709.565) (-5707.775) -- 0:27:40
   3700 -- [-5703.549] (-5711.586) (-5710.167) (-5707.783) * (-5715.285) (-5713.509) [-5708.113] (-5709.190) -- 0:26:55
   3800 -- [-5704.248] (-5711.081) (-5712.313) (-5707.702) * (-5722.075) (-5711.734) (-5711.998) [-5710.360] -- 0:26:12
   3900 -- [-5701.858] (-5717.476) (-5714.262) (-5716.201) * (-5713.589) (-5713.839) [-5710.122] (-5726.634) -- 0:25:32
   4000 -- [-5703.552] (-5710.062) (-5710.117) (-5714.897) * [-5707.821] (-5717.283) (-5708.325) (-5716.875) -- 0:24:54

  Average standard deviation of split frequencies: 0.003802

   4100 -- (-5708.443) [-5713.484] (-5707.028) (-5719.121) * (-5707.105) (-5715.197) [-5704.453] (-5714.368) -- 0:24:17
   4200 -- (-5706.532) [-5712.490] (-5712.392) (-5717.254) * (-5707.318) (-5710.493) [-5700.688] (-5719.113) -- 0:27:39
   4300 -- [-5705.507] (-5712.232) (-5710.880) (-5713.627) * (-5706.067) (-5713.913) [-5700.770] (-5710.083) -- 0:27:00
   4400 -- (-5707.165) (-5721.767) [-5708.552] (-5714.814) * (-5705.821) (-5726.707) [-5703.136] (-5707.494) -- 0:26:23
   4500 -- (-5706.309) [-5711.380] (-5711.319) (-5716.396) * (-5706.733) (-5715.701) [-5704.817] (-5708.722) -- 0:25:48
   4600 -- (-5712.988) (-5714.825) [-5706.396] (-5715.555) * (-5708.772) [-5709.265] (-5708.430) (-5709.552) -- 0:25:14
   4700 -- (-5714.802) [-5710.346] (-5708.163) (-5706.251) * (-5705.867) [-5708.108] (-5706.816) (-5715.459) -- 0:24:42
   4800 -- (-5709.563) [-5710.344] (-5710.108) (-5715.962) * (-5705.735) [-5706.771] (-5708.455) (-5708.251) -- 0:27:38
   4900 -- [-5710.881] (-5711.670) (-5708.528) (-5712.824) * (-5711.245) [-5706.895] (-5720.527) (-5714.925) -- 0:27:04
   5000 -- (-5707.725) [-5703.180] (-5704.342) (-5710.158) * [-5710.917] (-5708.044) (-5716.018) (-5710.541) -- 0:26:32

  Average standard deviation of split frequencies: 0.000000

   5100 -- (-5708.095) [-5702.965] (-5707.360) (-5710.215) * (-5711.496) [-5706.370] (-5712.716) (-5717.221) -- 0:26:00
   5200 -- (-5713.575) [-5707.509] (-5710.810) (-5714.767) * (-5710.906) [-5705.398] (-5706.485) (-5714.464) -- 0:25:30
   5300 -- (-5712.374) [-5704.743] (-5706.984) (-5712.640) * [-5705.824] (-5705.086) (-5707.049) (-5718.217) -- 0:25:01
   5400 -- (-5715.566) [-5704.716] (-5712.373) (-5713.791) * [-5708.017] (-5706.561) (-5709.910) (-5714.885) -- 0:27:37
```

Fig. 15.14. Output while MrBayes is running.



C:\Users\sergei\Desktop\Phyprogram\CLUSTALX\Gent_test_files\mrbayes.exe

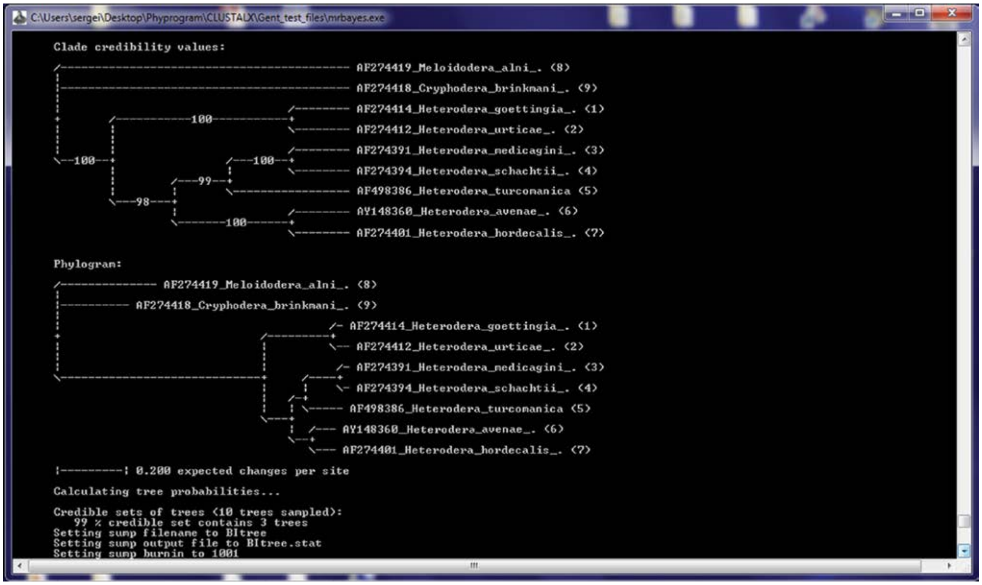
```
  997400 -- (-5708.015) [-5706.844] (-5718.142) (-5709.981) * (-5716.127) (-5718.909) [-5710.228] (-5708.644) -- 0:00:04
  997500 -- [-5705.905] (-5711.490) (-5709.943) (-5716.846) * (-5712.654) (-5717.803) (-5713.052) [-5710.010] -- 0:00:04
  997600 -- [-5701.119] (-5708.133) (-5714.290) (-5708.487) * (-5710.651) (-5713.475) (-5715.462) [-5709.461] -- 0:00:04
  997700 -- [-5700.996] (-5709.264) (-5713.972) (-5712.941) * [-5712.667] (-5716.198) (-5714.203) (-5715.873) -- 0:00:03
  997800 -- [-5701.137] (-5710.333) (-5707.021) (-5714.960) * (-5717.550) [-5714.552] (-5707.733) (-5713.401) -- 0:00:03
  997900 -- [-5702.922] (-5711.909) (-5704.511) (-5708.859) * (-5717.078) (-5710.638) (-5709.260) [-5707.740] -- 0:00:03
  998000 -- [-5707.018] (-5716.086) (-5711.483) (-5709.390) * (-5710.525) (-5712.840) [-5709.062] (-5703.557) -- 0:00:03

  Average standard deviation of split frequencies: 0.000315

  998100 -- [-5712.270] (-5707.023) (-5714.586) (-5704.177) * [-5713.622] (-5716.948) (-5710.470) (-5704.201) -- 0:00:03
  998200 -- (-5704.758) (-5708.520) (-5713.915) [-5704.011] * (-5723.769) (-5711.253) [-5706.454] (-5712.104) -- 0:00:03
  998300 -- [-5706.917] (-5712.767) (-5717.758) (-5706.552) * (-5720.764) (-5714.104) [-5705.132] (-5712.692) -- 0:00:02
  998400 -- [-5702.107] (-5725.787) (-5713.763) (-5707.924) * (-5724.035) (-5714.493) [-5706.326] (-5711.919) -- 0:00:02
  998500 -- [-5705.124] (-5720.996) (-5711.852) (-5710.290) * (-5724.780) (-5719.080) (-5705.650) [-5708.749] -- 0:00:02
  998600 -- [-5702.646] (-5716.857) (-5712.348) (-5716.323) * (-5725.131) [-5709.581] (-5707.238) (-5707.664) -- 0:00:02
  998700 -- [-5703.378] (-5713.662) (-5712.980) (-5716.853) * (-5725.429) (-5711.249) (-5704.573) [-5707.838] -- 0:00:02
  998800 -- (-5705.406) (-5713.312) [-5707.179] (-5712.575) * (-5725.002) (-5709.738) (-5708.965) [-5709.228] -- 0:00:02
  998900 -- (-5709.547) (-5714.973) (-5706.774) [-5706.262] * (-5731.413) (-5711.891) (-5711.920) [-5709.474] -- 0:00:01
  999000 -- [-5704.147] (-5714.179) (-5708.447) (-5707.260) * (-5731.038) [-5711.983] (-5711.542) (-5706.575) -- 0:00:01

  Average standard deviation of split frequencies: 0.000346

  999100 -- [-5705.647] (-5718.390) (-5710.395) (-5709.616) * (-5731.054) (-5713.353) (-5713.635) [-5707.616] -- 0:00:01
  999200 -- [-5702.642] (-5715.222) (-5707.145) (-5706.652) * (-5726.679) [-5707.239] (-5714.293) (-5706.985) -- 0:00:01
  999300 -- (-5701.643) (-5710.513) [-5705.871] (-5706.678) * (-5723.644) [-5708.396] (-5717.698) (-5710.632) -- 0:00:01
  999400 -- (-5703.337) (-5705.668) [-5702.873] (-5703.627) * (-5714.830) [-5707.432] (-5716.766) (-5714.212) -- 0:00:01
  999500 -- (-5701.820) [-5702.236] (-5702.989) (-5710.180) * [-5712.260] (-5710.859) (-5723.078) (-5714.097) -- 0:00:00
  999600 -- (-5703.699) [-5708.698] (-5703.081) (-5715.295) * (-5709.006) [-5706.981] (-5711.457) (-5710.798) -- 0:00:00
  999700 -- [-5701.385] (-5708.592) (-5715.824) (-5719.374) * (-5713.170) (-5708.433) (-5707.471) [-5709.642] -- 0:00:00
  999800 -- [-5709.179] (-5706.673) (-5716.131) (-5722.107) * (-5714.199) (-5710.469) [-5704.687] (-5716.923) -- 0:00:00
  999900 -- [-5707.079] (-5706.768) (-5709.666) (-5732.372) * (-5712.559) [-5715.019] (-5710.101) (-5717.227) -- 0:00:00
  1000000 -- [-5716.482] (-5703.857) (-5704.324) (-5733.137) * (-5706.298) (-5720.314) [-5706.993] (-5714.799) -- 0:00:00

  Average standard deviation of split frequencies: 0.000377

  Continue with analysis? (yes/no):
```

Fig. 15.15. Output of MrBayes run.

15.10 Visualization of Phylogenetic Trees

TreeView is a program for displaying and printing phylogenies (https://code.google.com/archive/p/treeviewx/). TreeView provides a simple way to view the contents of a NEXUS, PHYLIP, Hennig86, Clustal, or other format tree file (Page, 1996). The PAUP* for Windows does not have a graphical interface, hence TreeView allows you to create publication quality trees from PAUP files, either directly, or by generating graphics files for editing by other programs.

- Run TreeView program. Choose **File** > **Open**. Then select the file in your folder. Click **Open**. The tree appears in TreeView (Fig. 15.17).
- Click on **Phylogram** icon to see the tree with length branches. Choose **Tree > Order > Select Ladderize left** > **OK.** Outgroup taxa appear in the bottom of the tree.
- If the tree needs to be re-rooted, select **Tree > Define outgroup** and then select Outgroup taxa. Click OK.
- The displayed tree can be saved in a graph format (.emf). Choose **File** > **Save as graph,** then type file name, select format**. emf** and then click **Save.**


C:\Users\sergei\Desktop\Phyprogram\CLUSTALX\Gent_test_files\mrbayes.exe

```
Clade credibility values:

/------------------------------------------ AF274419_Meloidodera_alni_. (8)
|
|------------------------------------------ AF274418_Cryphodera_brinkmani_. (9)
|
|                                 /-------- AF274414_Heterodera_goettingia_. (1)
+       /-----------100-----------+
|       |                         \-------- AF274412_Heterodera_urticae_. (2)
|       |
|       |                         /-------- AF274391_Heterodera_medicagini_. (3)
\--100--+               /---100---+
        |               |         \-------- AF274394_Heterodera_schachtii_. (4)
        |       /---99--+
        |       |       \------------------ AF498386_Heterodera_turcomanica (5)
        \---98--+
                |                 /-------- AY148360_Heterodera_avenae_. (6)
                \-------100-------+
                                  \-------- AF274401_Heterodera_hordecalis_. (7)

Phylogram:

/-------------- AF274419_Meloidodera_alni_. (8)
|
|---------- AF274418_Cryphodera_brinkmani_. (9)
|
|                                        /- AF274414_Heterodera_goettingia_. (1)
+                              /---------+
|                              |         \-- AF274412_Heterodera_urticae_. (2)
|                              |
|                              |          /- AF274391_Heterodera_medicagini_. (3)
\------------------------------+     /----+
                               |     |    \- AF274394_Heterodera_schachtii_. (4)
                               |   /-+
                               |   | \------ AF498386_Heterodera_turcomanica (5)
                               \---+
                                   |  /--- AY148360_Heterodera_avenae_. (6)
                                   \--+
                                      \--- AF274401_Heterodera_hordecalis_. (7)

|---------| 0.200 expected changes per site

Calculating tree probabilities...

Credible sets of trees (10 trees sampled):
   99 % credible set contains 3 trees
Setting sump filename to BItree
Setting sump output file to BItree.stat
Setting sump burnin to 1001
```

Fig. 15.16. Tree output of MrBayes run.

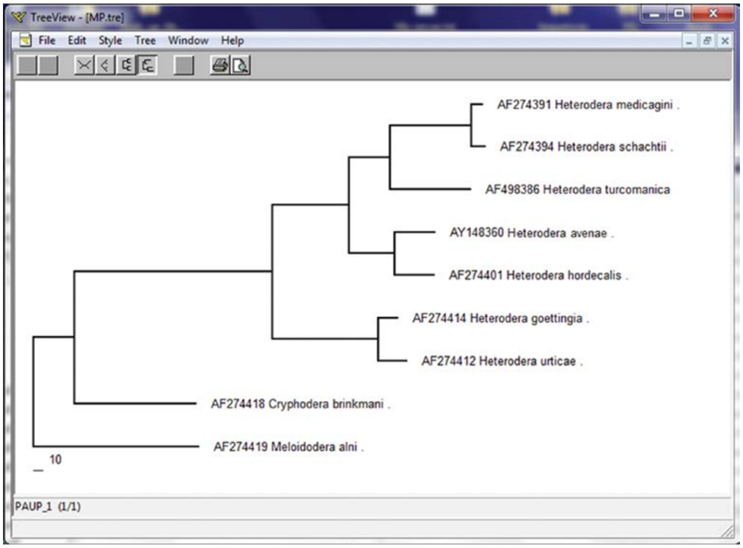
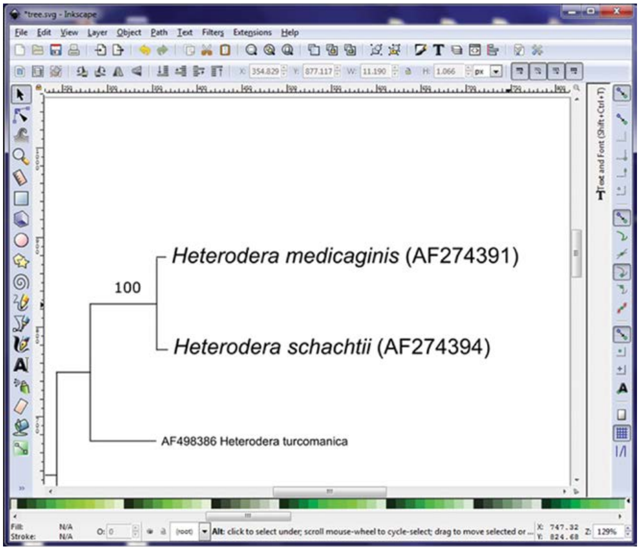

TreeView - [MP.tre]
File
Edit
Style
Tree
Window
Help
AF274391 Heterodera medicagini .
AF274394 Heterodera schachtii .
AF498386 Heterodera turcomanica
AY148360 Heterodera avenae .
AF274401 Heterodera hordecalis .
AF274414 Heterodera goettingia .
AF274412 Heterodera urticae .
AF274418 Cryphodera brinkmani .
AF274419 Meloidodera alni .
10
PAUP_1 (1/1)

Fig. 15.17. Tree visualized by TreeView.

This file can be opened and edited by Adobe Illustrator, CorelDraw, Inkscape or other Graph Editors.

FigTree is another graphical viewer of phylogenetic trees and a program for producing publication-ready figures (http://tree.bio.ed.ac.uk/software/figtree/). In particular, it is designed to display summarized and annotated trees produced by BEAST.

Inkscape (https://inkscape.org) is professional quality vector graphics software that runs on Windows, Mac and Linux. It is used for creating a wide variety of graphics and can be used for final preparation of a phylogenetic tree for a publication. The program can be freely downloaded from the website.



*tree.svg - Inkscape
File Edit View Layer Object Path Text Filters Extensions Help
X: 354.829 Y: 877.117 W: 11.190 H: 1.066 px
Heterodera medicaginis (AF274391)
100
Heterodera schachtii (AF274394)
AF498386 Heterodera turcomanica
Text and Font (Shift+Ctrl+T)
Fill: N/A
Stroke: N/A
O: 0
(root)
X: 747.32
Y: 824.68
Z: 129%

Alt: click to select under; scroll mouse-wheel to cycle-select; drag to move selected or ...

Fig. 15.18. Inkscape with a tree.

The phylogenetic tree should contain clear terminal labels with species names and GenBank accession numbers and numbers with bootstrap or posterior probability values placed on appropriate branches (Fig. 15.18).

15.11 Sequence Submission

The final and important step of a phylogenetic study is the submission of new sequences in public databases. Once sequences are submitted and accession numbers are assigned, these numbers must be included in a publication and published tree. New sequences should be submitted in one of the public database: GenBank, a comprehensive public database of nucleotide sequences and supporting bibliographical and biological annotation built and distributed by NCBI (Benson *et al.*, 2010); EMBL (the European Molecular Biology Laboratory Nucleotide Sequence Database in Europe); or DDBJ (DNA Data Bank of Japan). Daily data exchange within these databases ensures worldwide coverage.

There are some options for submitting data to GenBank:

- **BankIt** (https://www.ncbi.nlm.nih.gov/WebSub/), a WWW-based submission tool with wizards to guide the submission process; or
- **Submission Portal** (https://submit.ncbi.nlm.nih.gov), a unified system for multiple submission types. Currently only ribosomal RNA (rRNA), rRNA-ITS, Influenza or Norovirus sequences can be submitted with the GenBank component of this tool.

15.12 References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215, 403–410. DOI: 10.1016/S0022-2836(05)80360-2

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2010) GenBank. *Nucleic Acids Research* 38, D46–D51. DOI: 10.1093/nar/gkp1024

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17, 540–552. DOI: 10.1093/oxfordjournals.molbev.a026334

Darriba, D., Taboada, G.L., Doallo, R. and Posada, D. (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* 9, 772. DOI: 10.1038/nmeth.2109

Hall, B.G. (2017) *Phylogenetic Trees Made Easy.* 5th edn. Oxford University Press, Oxford, UK.

Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Research* 9, 868–877. DOI: 10.1101/gr.9.9.868

Larkin, M.A., Blackshields, G., Brown, N.P. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. DOI: 10.1093/bioinformatics/btm404

Lemey, P. Salemi, M. and Vandmme, A.-M. (eds) (2009) *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing.* 2nd edn. Cambridge University Press, Cambridge, UK. DOI: 10.1111/j.1541-0420.2010.01388.x

Nicholas, K.B., Nicholas Jr, H.B. and Deerfield II, D.W. (1997) GeneDoc: analysis and visualization of genetic variation. *EMBNEW News* 4, 14.

Nylander, J.A.A. (2004) *MrModeltest Version 2. Program distributed by the author*. Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden.

Page, R.D.M. (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* 12, 357–358.

Posada, D. (2008) jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution* 25, 1253–1256. DOI: 10.1093/molbev/msn083

Posada, D. and Crandall, K.A. (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics* 14, 817–818. DOI: 10.1093/bioinformatics/14.9.817

Ronquist, F. and Huelsenbeck. J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574. DOI: 10.1093/bioinformatics/btg180

Ronquist, F., Teslenko, M., van der Mark, P. *et al.* (2012) MrBayes3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* 61, 532–542. DOI: 10.1093/sysbio/sys029

Swofford, D.L. (2003) PAUP*: phylogenetic analysis using parsimony (*and other methods), version 4.0b 10. Sinauer Associates, Sunderland, Massachusetts, USA.

Talavera, G. and Castresana, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* 56, 564–577. DOI: 10.1080/10635150701472164

Wheeler, D. and Bhagwat, M (2007) Chapter 9. BLAST QuickStart In: Bergman N.H. (ed.). *Methods in Molecular Biology, Vol. 395: Comparative Genomics*. Humana Press Inc., Totowa, New Jersey, USA, pp. 149–175*.*

Yang, Z. and Rannala, B. (2012) Molecular phylogenetics: principles and practice. *Nature Review: Genetics* 13, 303–314. DOI: 10.1038/nrg3186