Sequence Editing and Anaylsis (v 1 b 2)

Prepared by Dr Michael Monaghan

Introduction

This manual is intended to help you with each step involved in the process transforming raw chromatogram data to building phylogenetic trees. Here it is assumed that you have already downloaded data from the Zoology server, and that you are not using the STARS automated editing system. Below, the most useful file types and software programs are listed. There are many others e.g., http://wikiomics.org/wiki/Bioinfo_tutorial

1. File types

ABI (`*.ab1`) – raw data from ABI automated sequencer

FASTA (`*.fasta`) – text file with the following format, used for alignment:

```
>sequence_name1_gene[↵]
ATTAGCTGTACGATAAGCTAAGCTAGGCTAGGCATGCATTCGGATGCGTACGTGC
ATGCATGCATGCAGTACGTACGT[↵]
>sequence_name2_gene[↵]
ATTCGCTATTAGCTGTACGATAAGCTAAGCTAGGCTAGGCATGCATTCGGATGCG
TACGTGCATGCATGCATGCAGTACGTACGT[↵]
```

NB: there must be a hard return [↵] after the name of the sequence and at the end of the sequence *only*. Each name should end with the same identifier, i.e. _16S

NEXUS (`*.nex`) – text file for tree-building and other analyses (see appendix 1)

2. Essential Software (make sure these are all installed on your computer)

SEQUENCHER 4.5
: assembly of forward and reverse chromatograms, sequence editing

MACCLADE OSX
: sequence viewer, save files as `*.nex` for PAUP & TNT

SE-AL
: sequence viewer, export files in FASTA format

PAUP
: tree-search and general data handling

SAFARI or INTERNET EXPLORER
: web-based alignment (e.g., CLUSTALW)

MSWORD, EXCEL
: data handling

3. Additional Software (for other types of analyses)

TNT, PhyML, Mr Bayes
: other tree searching programs (parsimony, likelihood, Bayesian)

TreeRot

calculating tree (Bremer) support (see appendix 4)

I. Create a SEQUENCHER project

1. On a Mac, make a new folder in which you will keep all of your work. This should be a different folder from the one that has your sequences chromatograms. Give it your own name so that everyone will know whose it is, e.g., `Elvis_Data`.

2. Open SEQUENCHER v 4.5 software (NB: files that are saved in v 4.5 can not be read by v 4.1). A new, empty project will be opened.

3. Click and drag all of the chromatograms from the folder in which they are located into the open SEQUENCHER window [alternatively, choose `File-Import-Sequences`].

Save the project as `yourname_Sequences_Gene` in your folder

e.g., `Elvis_Sequences_CO1`

This project will contain all of your raw chromatogram data and may therefore be large. Save the project often, e.g., after editing each sequence.

II. Sequence Assembly

1. Choose two matching chromatograms, or paired forward-reverse reads, from the list of chromatograms, e.g.,

> **001_676767_CO1R**_032_A16.ab1
> **001_676767_CO1F**_032_A15.ab1
>
> where 001 is the plate number, 676767 is the BMNH number, CO1R is the reverse read and CO1F is the forward read (usually 5' – 3') of the *coxI* gene region. We can ignore the next two pieces of information (032_A16) – they are for the sequencing laboratory -- and .ab1 indicates it comes from the automated ABI sequencer.

2. Highlight both sequences (click and hold the shift key, or click and drag) and click the "Assemble Automatically" button. It is possible to assemble any number of sequences this way, but for now we will only look at the forward and reverse reads of one sequence.

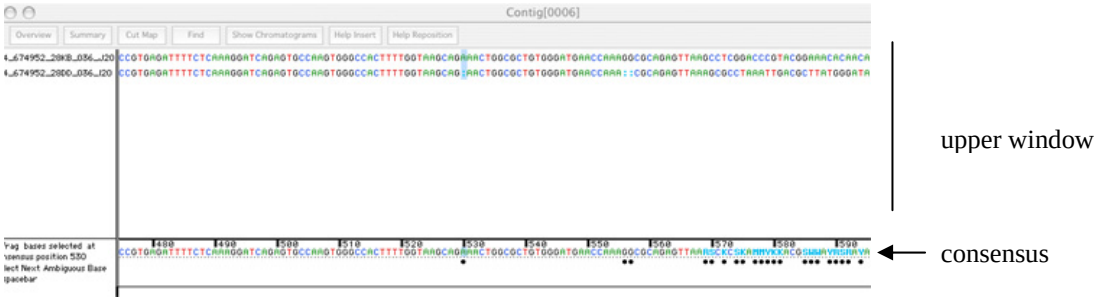
3. You should now have a single highlighted contig displayed on your screen. If so, proceed to the next section. If not go to step 4.

4. Lower the assembly matching paramaters and try to contig the sequences with a lower threshold. If necessary, reduce the minimum match to 60 (default is 85). If this works, proceed to the next section. If not go to step 5.

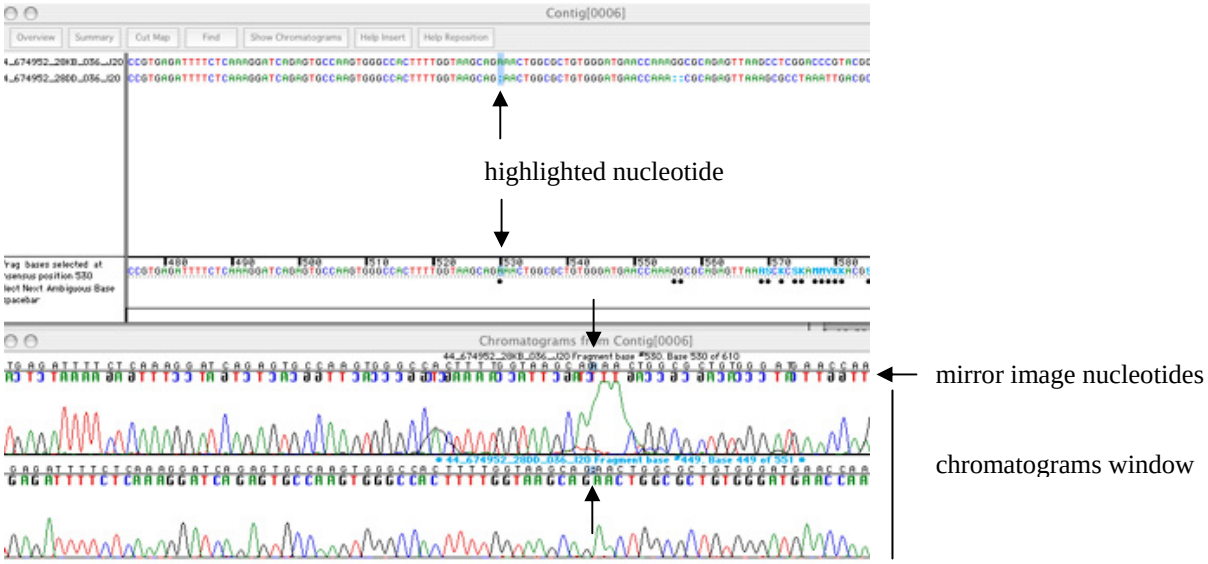
5. Open each chromatogram individually to try and determine whether wither one has good sequence data. If only one has good data, then it must be edited alone, later in the editing process. For now, choose another pair of contigs and repeat this section (**II**) and after proceeding to the next section (**III**) we can edit this single contig during section **IV** step 4 .

III. Viewing Sequences and Chromatograms

1. Double click the “Contig” you have created. In the top window are the two sequences as text; one is displayed the reverse complement to the other, thus they *appear* identical. In the bottom window is the “consensus” sequence, with black dots indicating bases that differ between the two reads (e.g., an A in one chromatogram but a T in the other). You can view any number of sequences that you have assembled, but for now we will look at only two.



2. Click on any base in the **consensus** (lower window) and then click “Show chromatograms” . A second window opens and the two sequences are displayed as chromatograms – the reverse complement looks identical except for the coloured letters which are a mirror image. The nucleotide that you clicked on will be highlighted in all windows.

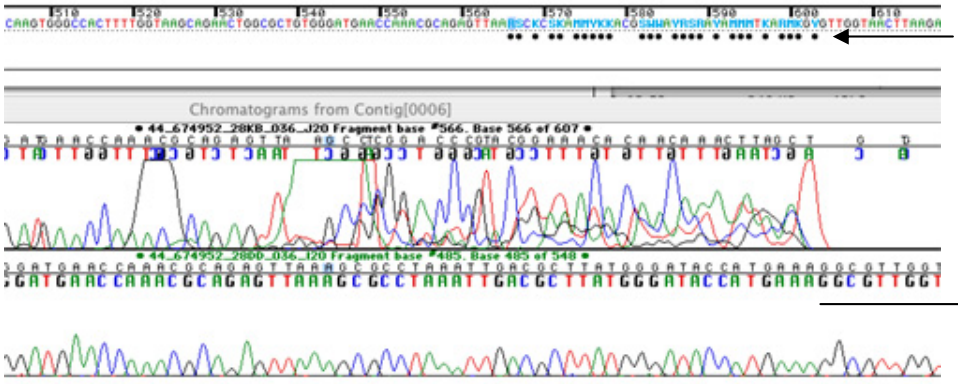


IV. Editing Sequences

Don't worry, it is common to have poor sequence data at the ends of chromatograms: close to the primer binding site, toward the end of the sequencing read, or both.

1. When there is poor sequence on the **right** of your screen: highlight the poor sequence in the **upper window** and delete using the **backspace** button. Below is an example of poor sequence quality close to the priming site. The upper row of sequence in the **upper window** is represented by the upper row of chromatogram data (in the chromatograms window).

highlight and delete with backspace key



mismatch indicators

poor sequence quality, in this case near the primer binding site

primer binding site

2. When there is poor sequence on the **left** in one of the strands, highlight the poor sequence in the **upper** section and delete using the **delete** button

*NB if you accidently use the delete key from the right or the backspace from the left, you can "undo" this using apple-z or using the edit menu. If this does not work (you can only "undo" a single action) then dissolve the contig and then re-assmeble it.

3. Poor sequence in the interior: highlight the miscalled base in the **consensus** (lower) window: and type the correct base (A,C,G, or T), or remove gaps using the delete button. Use the better of the two chromatograms to decide which nucleotide is present. Use N when it is not possible to assign a base unambiguously.

4. If you have only one good strand of a sequence, i.e. if you were not able to contig your paired reads but one (or both) of them appear to be good quality sequence data, then choose one of the good contigs you have already edited and contig the good sequence to this contig. The result will be a window like the one above, but with three sequences and three chromatograms. Edit the third sequence strand by viewing the chromatogram in the window, but remember that you are *not* comparing it to the other two strands (it is likely that some nucleotides will be different). Edit the third strand in the **upper** window, not the consensus.

V. Name contig

1. When you have finished editing the sequence, close the two windows and name the contig using the following template, e.g.,

> **001_676767**_Baerho_EN_**CO1**
>
> Where the plate number (001) and BMNH numbers (676767) are first, followed by any information you want to add (e.g., here a code for the Linnaen binomial *Baetis rhodani* and country England) and ending with _CO1 to specifiy the gene region.

Rules for naming Contigs

1. Use only alphanumeric characters and avoid + / * ; - .
2. Do not leave white space (use_underscore_marks_instead)
3. Do not exceed 30 characters.
4. Make sure *every* contig has the same ending e.g., _CO1 above. For the other genes, use logical contig names such as _16S, _28S.

2. Save the SEQUENCHER project.

> ** SEQUENCHER runs from a file server machine, not your local hard disc. Sometimes the server has to be rebooted or the connection is lost, meaning any unsaved data could be lost. *Please* save your work frequently!

VI. Export Consensus Sequences

Sequences are now ready to exported from SEQUENCHER and entered into an alignment program.

1. In your working folder, i.e. in `Elvis_Data`, create a new folder named `TEMP`.

2. Highlight all of your contigs in the SEQUENCHER window (click and drag) and choose `File-Export-Consensus`. In the window choose **`ASCII`** as the output type. Export all sequences to `TEMP` when prompted.

3. In the SEQUENCHER window, choose `File-Import-Folder of sequences`. Import "all" when prompted.

4. Highlight all of these newly imported sequences (do not include any of your previous contigs) and click the "Assemble automatically" button.

** If necessary, change the assembly settings by reducing the minimum match to 60 (default is 85) as in section **II** step 4.

5. Choose `File-Export-Consensus`. In the window choose `NEXUS/PAUP` as the output type and add the file extension `.nex`, e.g., `Elvis_Sequences_CO1.nex`

VI. Create the FASTA file

FASTA files are used for the alignment program CLUSTALW

1. Open MACCLADE and open the nexus file you just exported from SEQUENCHER (`Elvis_Sequences_CO1.nex`).

2. Choose `File-Options` for saving and **unclick** the "interleaved" box.

** Sequencher exports nexus files in an "interleaved" format, meaning that line breaks are inserted every 50 bp of each sequence. This format can not be read by many programs, so we remove the interleaving with MACCLADE.

3. Open SE-AL and open the file `Elvis_Sequences_CO1.nex` that you just un-interleaved. Choose `Alignment-Reorder sequences`. This will place your sequences in alphaneumerical order. This is an important step when you have more than one data set e.g., *coxI* and 18S data.

4. Choose `File-Export` and select FASTA as the file type. Export to your folder. The file should have the fasta extension, i.e., `Elvis_Sequences_CO1.fasta`

5. Open MSWORD and open your .fasta file. Choose `Edit-Replace`.

6. In the window, in `Replace` type `^p` and leave `With` blank (this will delete all paragraph marks). Click `Replace All`.

7. `Replace >` with `^p>` (adds a paragraph mark before each sequence name)

8. Replace all `_CO1` (or _28S etc) with `_CO1^p`

9. Save the file (in **MS-Dos** format) as `Elvis_Sequences_CO1.dos`

** The web-based ClustalW we use runs on a server that accepts only `.dos` format

VII. Combine FASTA files from multiple sources

1. If you want to add sequences that the others have edited (e.g. from GenBank or other members of the group), add them in `.fasta` format to the bottom of your FASTA file.

2. All sequences must be in the same orientation. Make sure this is the case. If necessary, import your fasta file, with your own as well as others' sequences, into a new, empty SEQUENCHER project and go back to section **VI** steps 4 and 5.

3. If, after alignment, you find that some of the sequences align very poorly, the first thing you should consider is that not all sequences were in the same orientation when you created your FASTA file. See step 2 above.

VIII. Sequence alignment

1. Open INTERNET EXPLORER or, if you do not have it, SAFARI. Open http://align.genome.jp This is an extremely fast web server for alignment (and other) applications. There are other www servers dedicated to sequence alignment and if this one is down you can search for others using Google (e.g., Institut Pasteur has a server as well).

2. Load your fasta file by clicking "Browse". Remember that it must be **dos** formatted.

3. Choose DNA as the data type. For multiple alignments, notice the default weight matrix and gap penalties are `IUB` and open = 15, extension = 6.66. These can be changed in order to compare different alignments. Other web servers will have a slightly different interface but you can always change gap penalties. To begin we will use the defaults.

4. In the Options box type `–OUTORDER=INPUT.` This will keep all of your sequences in the same alphanumeric order you created with `Se-Al` in section **VI** step 3.. This is very helpful when you want to combine the three genes later, as the sequences *must* be in the same order.

5. Click the "Execute multiple alignment" button.

6. Click the hyperlink `clustalw.aln`

7. Hold the **control** button while you click and hold (left mouse click on PC) the `clustalw.aln` hyperlink. If you are using INTERNET EXPLORER, select "Download file to disc" and save the `pushfile` to your TEMP directory as `Elvis_CO1.aln`. If you are using SAFARI then you must download it to whatever directory has been specified in the settings.

IX. View sequences with MACCLADE and trim primers

1. Open MACCLADE and then open your aligned sequences file `Elvis_CO1.aln`. At the prompt click "ok" for CLUSTALW format.

2. If you see only black-and-white nucleotides, choose `Display-Color cells` to view the sequences more easily.

3. Check the general quality of the alignment and check if there are any major problems. A common problem is that some sequences were 5'-3' and others were 3'-5' in the fasta file. This will be immediately obvious! See section **VII** step 3.

4. Next check sequences for obvious errors. For example, protein-coding *coxI* or H3 should not have any gaps in the sequence except perhaps at the ends.

5. Locate the primers in your sequences. Parts of the primer sequences will be visible and the ends of some, but probably not all, of your sequences. These nucleotides are primer sequence rather than the gene sequence of the organism you are studying.

6. Hold the shift key and click the left-most column of primer sequence, then click the right-most column to highlight the primer binding region. Delete using the backspace key. Click "yes" at the prompt.

7. Choose `Utilities-Change All-Terminal gaps to missing`. Most analyses will treat indels (the loss or gain of nucleotides) as evolutionarily important characters, so we want to code the gaps (`-`), but missing data at the beginning or end of a sequence must be coded as missing (`?` or `X` or `N`).

8. Choose `File-Options` for saving and **unclick** the "interleaved" box. Save the file as `Elvis_CO1.nex`

X. Amino acid translation with MacClade (optional)

Protein-coding genes (e.g., *coxI*) can be analysed using a combination of 1st, 2nd, and 3rd position nucleotides, based on the assumption that different character positions evolve at different rates. To prepare your CO1 data to be able to use these options in PAUP, do the following steps in order:

1. Choose `Edit-Select all`

2. Choose `Characters-Genetic code-Drosophila mtDNA`

3. Choose `Characters–Codon positions–Calculate–1 (or 2 or 3)`

4. Choose `Display–Color cells–Don't color`

5. Choose `Display-Amino acid translation–show translated amino acids`

6. Choose `Display-Amino acid translation-color translated amino acids`

7. Choose `Display-Amino acid translation-dim nucleotides`

8. Check the sequence for stop codons (black). If visible, recalculate codon positions. If all 3 positions produce stop codons then you may be looking at the reverse complement the sequences.

9. To view the reverse complement, in order Choose `Edit-Select all`, Choose `Utilities-Complement`, Choose `Utilities-Reverse`

10. Save your file as `Elvis_CO1_AA.nex`

XI. Create the NEXUS file for PAUP

1. Open PAUP.

2. Open your nexus file. If you have only one gene, rename it `Elvis_CO1.paup` and the file is ready for **XII**.

3. If you have several genes for your data set, open each one in a separate window (`*_CO1.nex`, `*_16S.nex` and `*_28S.nex`).

4. Note the number of characters in each dataset (`nchar=` ??? in the `Dimensions` settings line`)` because you will need this information later.

5. Choose one of the files (e.g., the `CO1.nex` file) and delete all text below the matrix (i.e., delete everything from "`BEGIN MacClade`" onward). The last few lines of your file should now look like

```
>last_sequence_of_your_dataset_CO1   ATGCTGATGCTAGGC etc…
;
end;
```

6. After your last sequence and before the semicolon (;), copy and paste the sequences from your other data matrices (e.g., 18S) into this file. Each group of sequences should have the same species, in the same order (section **VI** step 3).

7. If a species is missing from one of the files, then the name must be added in the list and the data be filled with "?????". In the following example, there were no 16S data available for the specimen "Cadeltocps"

```
CAcantholu_CO1    ATGCTAGTC
CAdeltocps_CO1    ATCCTAGTG

CAcantholu_16S    AATGATGAT
CAdeltocps_16S    ?????????

CAcantholu_28S    ATGCTAGTC
CAdeltocps_28S    ATCCTAGTG
;
End;
```

8. In the header, change the `nchar=` in the `Dimensions` line to reflect the new, *total* number of nucleotides (e.g, CO1+16S+28S). Type `INTERLEAVE` in the `Format` line. The header should be something like this (`ntax` and `nchar` will vary):

```
#NEXUS

Begin DATA;
     Dimensions ntax=2 nchar=27;
     Format datatype=NUCLEOTIDE gap=- INTERLEAVE;
     Matrix
```

9. Save the NEXUS file as `Elvis_combined.paup` and save it in your folder.

XII. Create character sets (optional)

1. In the footer of your nexus file `Elvis_combined.paup`, delineate the gene regions by typing the following information starting two lines below `End`. See Appendix I for an example of a complete file(The order of gene regions and the number of characters will be different for everyone, this is only an example)

```
Begin sets;
 charset CO1 = 1-746; (the first group of sequences are CO1, 746 bp)
 charset 28S = 747-1548; (second group is 28S, 802 bp)
;
End;
```

Where the range of numbers corresponds to the number of characters in the aligned matrix in section **XI** step 4.

NB People often make mistakes in the mathematics! 746 + 802 = 1548.

2. If you used MACCLADE for the amino acid translation (section **X**) the program will have generated the information for codon positions. It should look like this:

```
Begin codons
 Codonposset * codon positions =
       1: 2-356\3
       2: 3-357\3 357,
       3: 1-355\3;
 Codeset * untitled = mtDNA.dros: all ;
end;
```

XII. PAUP parsimony tree-searches

1. In PAUP, execute your matrix (`File-execute` or apple+R).

2. Under `Analysis` choose `Parsimony` (this is normally the default)

3. Under `Analysis` choose `Parsimony settings`

4….

Coming soon:

TreeRot
GapCoder
r8s
BEAST
PhyML
MrBayes
Arlequin

and many exciting others…

Appendix I. Example PAUP file

```
#NEXUS
Begin DATA;
     Dimensions ntax=2 nchar=27; [this is the total of 3 genes]
     Format datatype=NUCLEOTIDE gap=- INTERLEAVE;
     Matrix

CAcantholu_CO1 ATGCTAGTC
CAdeltocps_CO1 ATCCTAGTG

CAcantholu_16S AATGATGAT
CAdeltocps_16S ?????????

CAcantholu_28S ATGCTAGTC
CAdeltocps_28S ATC--AGTG
;
End;

Begin sets;
     charset CO1 = 1-9; [the first group is CO1, 9 bp]
     charset 16S = 10-18; [second group is 16S, 9 bp]
     charset 28S = 19-27, [third group is 28S, 9bp]
;
End;

Begin codons
     Codonposset * codon positions =
          1: 2-356\3
          2: 3-357\3 357,
          3: 1-355\3;
     Codeset * untitled = mtDNA.dros: all ;
end;
```

Appendix II. Other software programs (freeware unless stated) and web-based tools:

This list is not exhaustive. When no specific link is given, programs can be found by typing the name (in bold) into a search engine. In Google, perhaps add "sequence", "alignment" or "phylogeny" and click "I'm feeling lucky."

General information:

**The most comprehensive source of programs (freeware and otherwise) is here:
http://evolution.genetics.washington.edu/phylip/software.html

EvolDir (the evolution directory) contains discussion strings about software and analyses, as well as studentships, jobs, and conferences related to evolutionary biology.
http://life.biology.mcmaster.ca/~brian/evoldir.html

Chromatogram assembly and editing:

BioEdit allows you to assemble, view and edit contigs, and contains a clustal alignment algorithm. In practice it is not as user-friendly as Sequencher.

Sequence viewing:

seaview is a sequence viewer (like MacClade)

Sequence alignment:

clustalW is available for local (i.e. on your computer) use

Muscle is an alignment program

Sequence databases:

GenBank

Tree search programs:

TNT is a parsimony-based program that incorporates many newly developed search methods (US$ 80)

PhyML is a maximum likelihood tree search program
http://atgc.lirmm.fr/phyml/

MrBayes is a Bayesian tree search program

POY uses an optimisation procedure to align sequences and build trees simultaneously

Branch support:

Treerot calculates Bremer support by creating a file that executes in PAUP

PRAP calculates Bremer support similarly to Treerot but incorporates ratchet searches