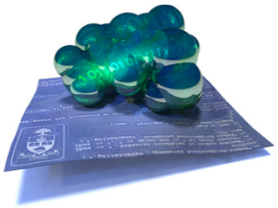
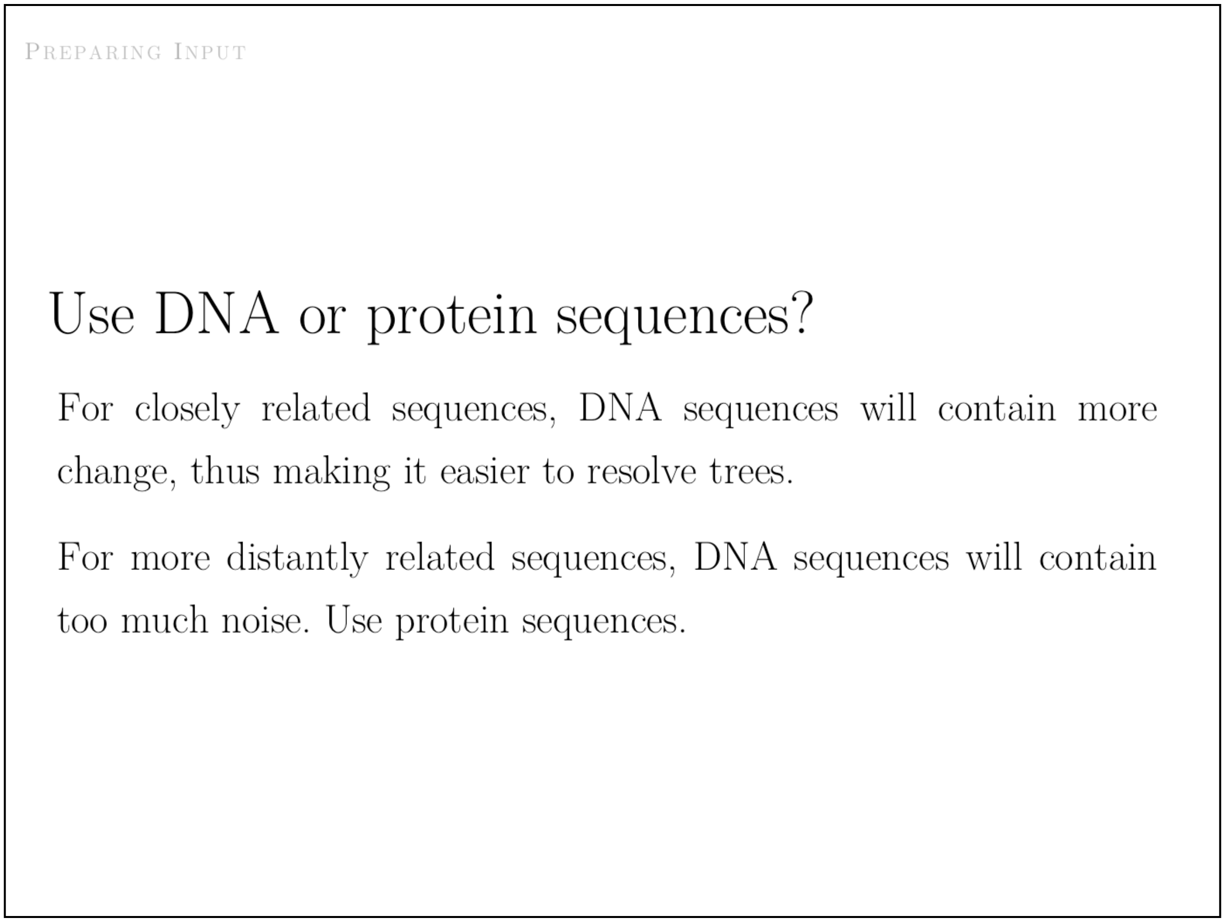
A
Bioinformatics
Course

Phylogenetic Data Preparation



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto



PREPARING INPUT

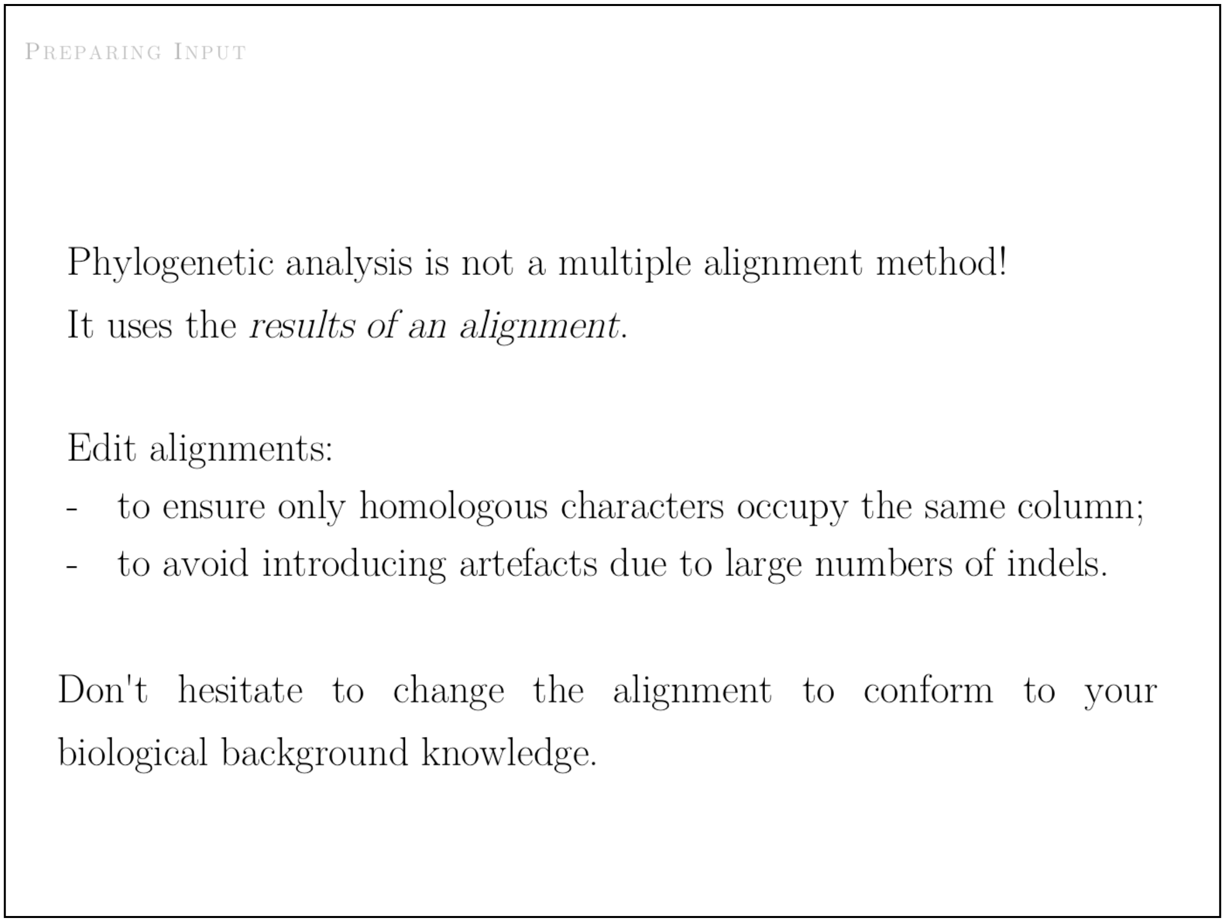
Use DNA or protein sequences?

For closely related sequences, DNA sequences will contain more change, thus making it easier to resolve trees.

For more distantly related sequences, DNA sequences will contain too much noise. Use protein sequences.

So you plan to compute a pphylogenetic tree. But how to begin?

Should you use DNA or protein sequences? How should you align them? Should you work with full-length sequences or domains? What about poorly aligned regions? How many sequences do you need?



Preparing Input

Phylogenetic analysis is not a multiple alignment method!
It uses the *results of an alignment*.

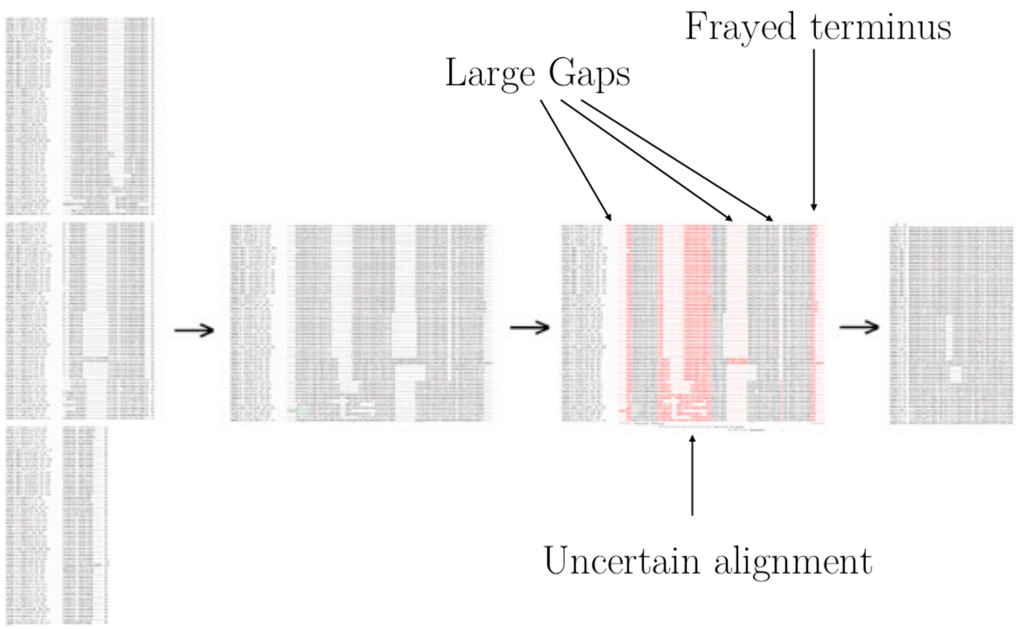
Edit alignments:

- to ensure only homologous characters occupy the same column;
- to avoid introducing artefacts due to large numbers of indels.

Don't hesitate to change the alignment to conform to your biological background knowledge.

Of course, the quality of the tree depends on the quality of the alignment. Use additional sequences to resolve ambiguities, use sevaral MSA algorithms, and don’t hesitate to edit the alignment in case you have additional information eg. regarding functional sites in your sequences.

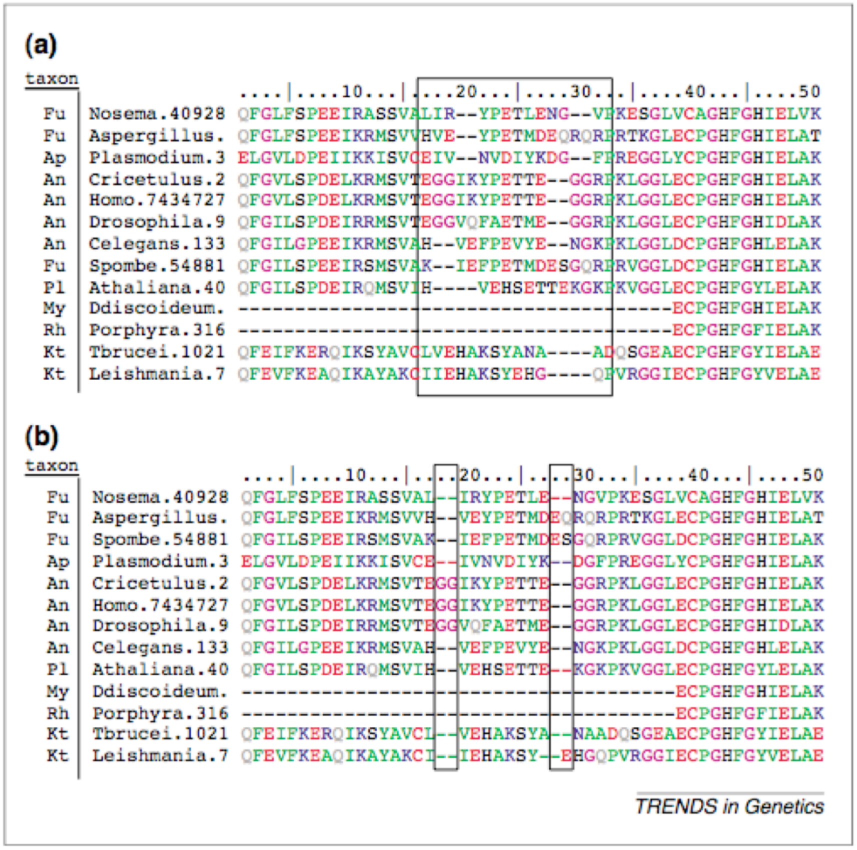
Preparing Input



Regions of the alignment that contain large numbers of gap-characters appear unreasonably similar to tree inference algorithms that count gaps as identical states. Delete all but a few of these columns. Regions where the alignments itself appear uncertain and have no apparent similarity should likewise be deleted: such regions only add noise. The same goes for (frayed) termini and other regions that contribute little if anything to selection on the sequence.

However, also place these decisions into the context of your entire set of sequences. If sequences are overall too similar, they carry little information of use for tree-construction.

Preparing Input



(a)

```
taxon
                     ....|....10...|....20...|....30...|....40...|....50
 Fu | Nosema.40928  QFGLFSPEEIRASSVALIR--YPETLENG--VPKESGLVCAGHFGHIELVK
 Fu | Aspergillus.  QFGLFSPEEIKRMSVVHVE--YPETMDEQRQRPRTKGLECPGHFGHIELAT
 Ap | Plasmodium.3  ELGVLDPEIIKKISVCEIV--NVDIYKDG--FPREGGLYCPGHFGHIELAK
 An | Cricetulus.2  QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
 An | Homo.7434727  QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
 An | Drosophila.9  QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
 An | Celegans.133  QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
 Fu | Spombe.54881  QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
 Pl | Athaliana.40  QFGILSPDEIRQMSVIH----VEHSETTEKGKPKVGGLECPGHFGYLELAK
 My | Ddiscoideum.  -------------------------------------ECPGHFGHIELAK
 Rh | Porphyra.316  -------------------------------------ECPGHFGFIELAK
 Kt | Tbrucei.1021  QFEIFKERQIKSYAVCLVEHAKSYANA----ADQSGEAECPGHFGYIELAE
 Kt | Leishmania.7  QFEVFKEAQIKAYAKCIIEHAKSYEHG----QPVRGGIECPGHFGYVELAE
```

(b)

```
taxon
                     ....|....10...|....20...|....30...|....40...|....50
 Fu | Nosema.40928  QFGLFSPEEIRASSVAL--IRYPETLE--NGVPKESGLVCAGHFGHIELVK
 Fu | Aspergillus.  QFGLFSPEEIKRMSVVH--VEYPETMDEQRQRPRTKGLECPGHFGHIELAT
 Fu | Spombe.54881  QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
 Ap | Plasmodium.3  ELGVLDPEIIKKISVCE--IVNVDIYK--DGFPREGGLYCPGHFGHIELAK
 An | Cricetulus.2  QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
 An | Homo.7434727  QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
 An | Drosophila.9  QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
 An | Celegans.133  QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
 Pl | Athaliana.40  QFGILSPDEIRQMSVIH--VEHSETTE--KGKPKVGGLECPGHFGYLELAK
 My | Ddiscoideum.  -------------------------------------ECPGHFGHIELAK
 Rh | Porphyra.316  -------------------------------------ECPGHFGFIELAK
 Kt | Tbrucei.1021  QFEIFKERQIKSYAVCL--VEHAKSYA--NAADQSGEAECPGHFGYIELAE
 Kt | Leishmania.7  QFEVFKEAQIKAYAKCI--IEHAKSY--EHGQPVRGGIECPGHFGYVELAE
```

TRENDS in Genetics

(a) The raw output from a ClustalX alignment of rpb1 sequences, which predicts six insertion/deletion events (boxed), some of which are blatantly inconsistent with known taxonomy.

(b) The refined alignment makes much better evolutionary sense, because it shows only two insertion events in well-defined taxonomic groups (animals and higher fungi).

Taxon labels are
Fu (fungi),
An (animals),
Pl (green plant),
Ap (apicomplexan),
Rh (rhodophyte),
My (mycetozoan),
Kt (kinetoplastids).

In (b), the sequence from *S. pombe* has been placed adjacent to the other fungi to make these relationships more obvious.

From Baldauf, S. (2003) Phylogeny for the faint of heart. *TiG*B **19**:345-351 http://steipe.biochemistry.utoronto.ca/abc

BORIS.STEIPE@UTORONTO.CA

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA