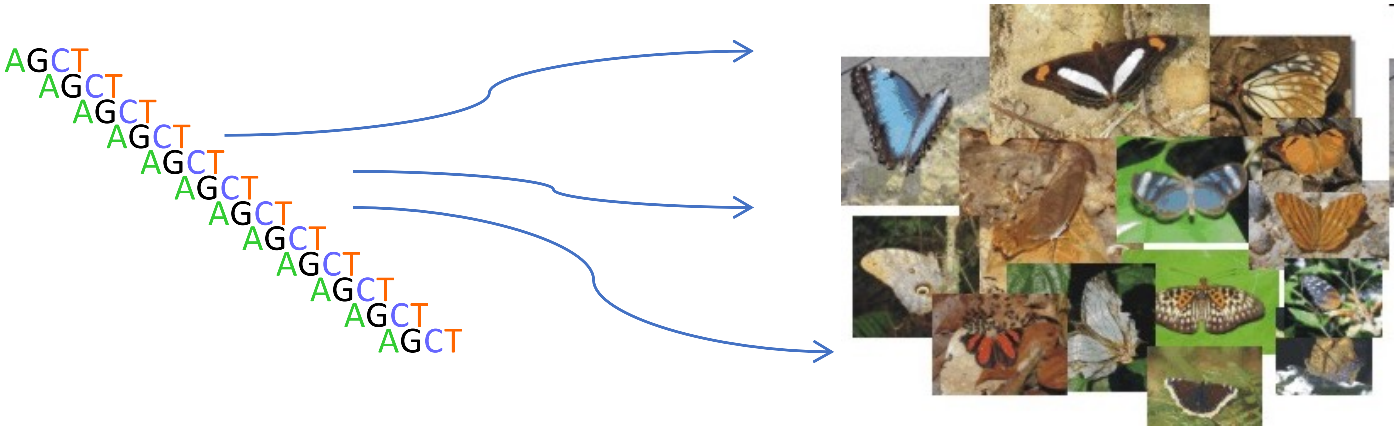
BIOR90 Evolutionary Biology - Methods and Applications 2023

Molecular phylogenetics

Teachers Jadranka Rota, Etka Yapar, Niklas Wahlberg, Sridhar Halali

Recap: Why *molecular* systematics?

- Ease of data generation for large numbers of taxa
- Ease of generating a large number of independent data sets for given taxa
- Molecular characters behind the morphological characters we see



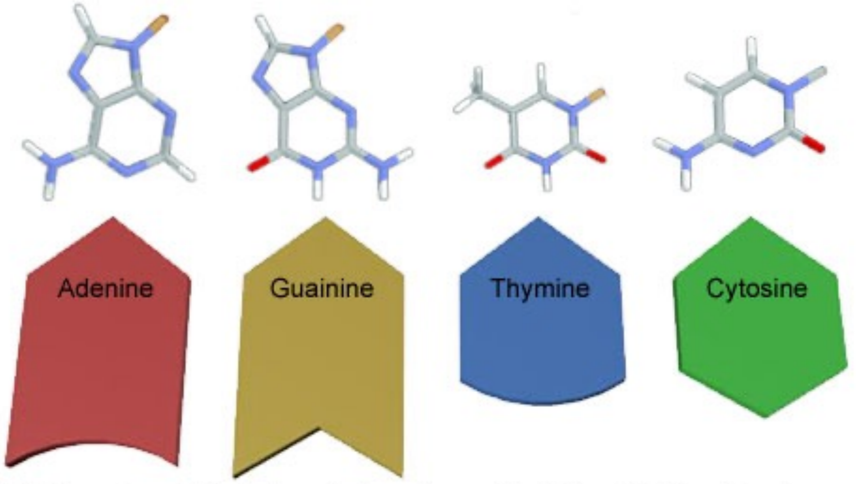
DNA as a source of information

- **DNA has four characters**

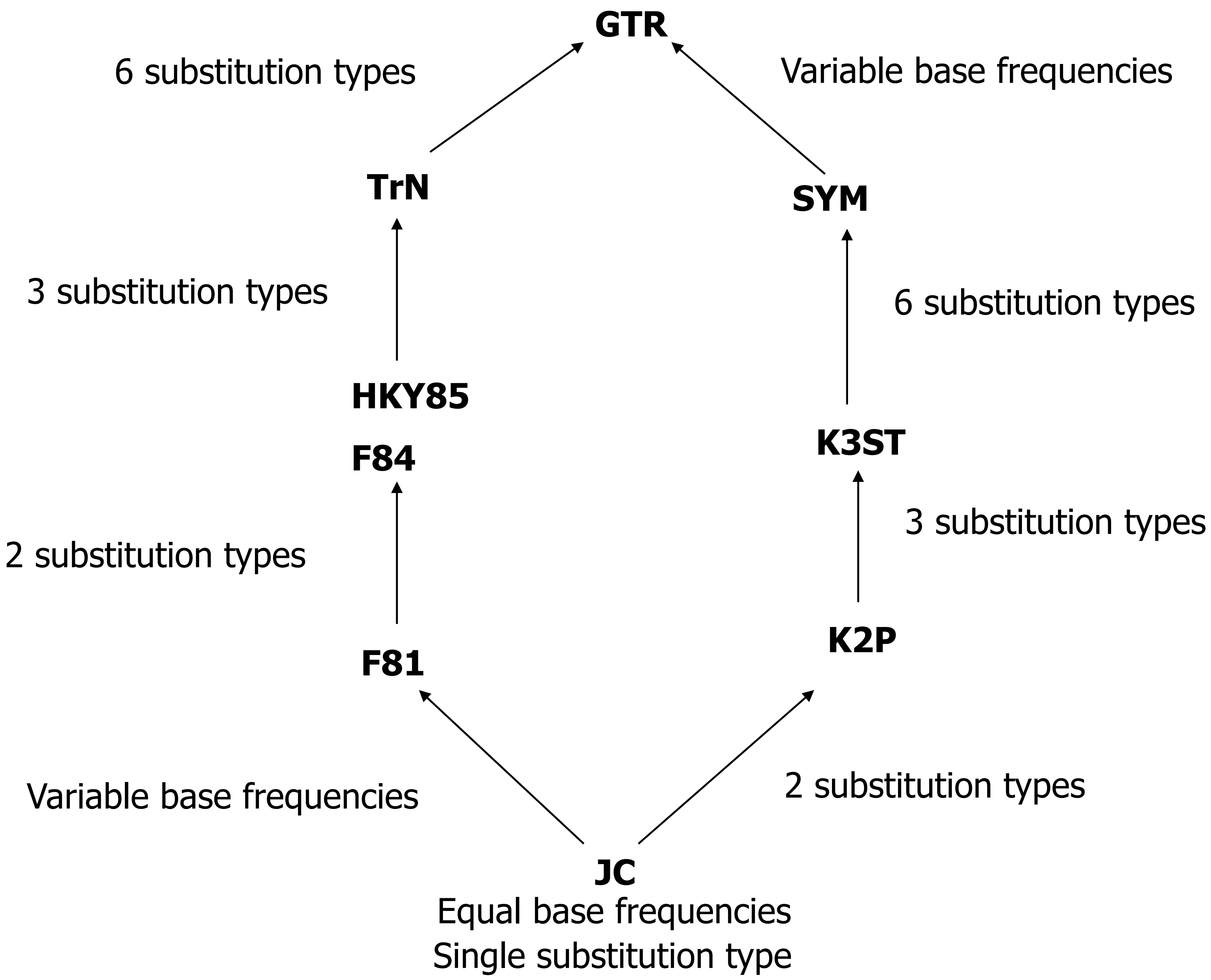
Purines

Pyrimidines

Figure B-3: The Four Nitrogenous Bases



Each base has a distinct shape that can be used to distinguish it form the others. 3D representations of the four bases are shown, with the corresponding chemical structures drawn above.



GTR

6 substitution types

Variable base frequencies

TrN

SYM

3 substitution types

6 substitution types

HKY85

F84

K3ST

3 substitution types

2 substitution types

F81

K2P

Variable base frequencies

2 substitution types

JC

Equal base frequencies
Single substitution type

Maximum Likelihood

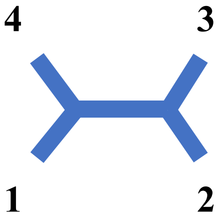
- For reconstructing phylogenies

Model

- which tree topology (τ), branch lengths, and parameters of DNA evolution model (θ) (e.g. transition/transversion ratio, base frequencies, …) are maximizing the probability of observing the sequences at hand?

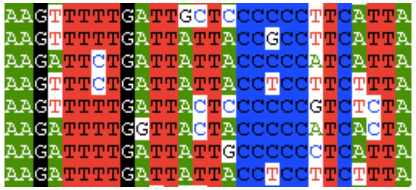
Data

$$L(\tau,\theta) = \Pr(\text{Data} \mid \tau,\theta)$$

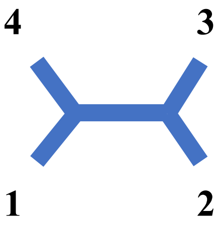


	A	C	G	T
A		α	α	α
C	α		α	α
G	α	α		α
T	α	α	α	

=



```
AAGTTTTTGATTGCTCCCCCCTTCATTA
AAGTTTTTGATTATTACCGCCTTCATTA
AAGATTCTGATTATTACCCCCATCATTA
AAGTTTCTGATTATTACCTCCTTCTTTA
AAGTTTTTGATTACTCCCCCCGTCTCTA
AAGATTTTGGTTACTACCCCCATCACTA
AAGATTTTGATTATTGCCCCCCTCATTA
AAGATTTTGATTATTACCTCCTTCTTTA
```



	A	C	G	T
A		α	α	α
C	α		α	α
G	α	α		α
T	α	α	α	

ML analysis in short

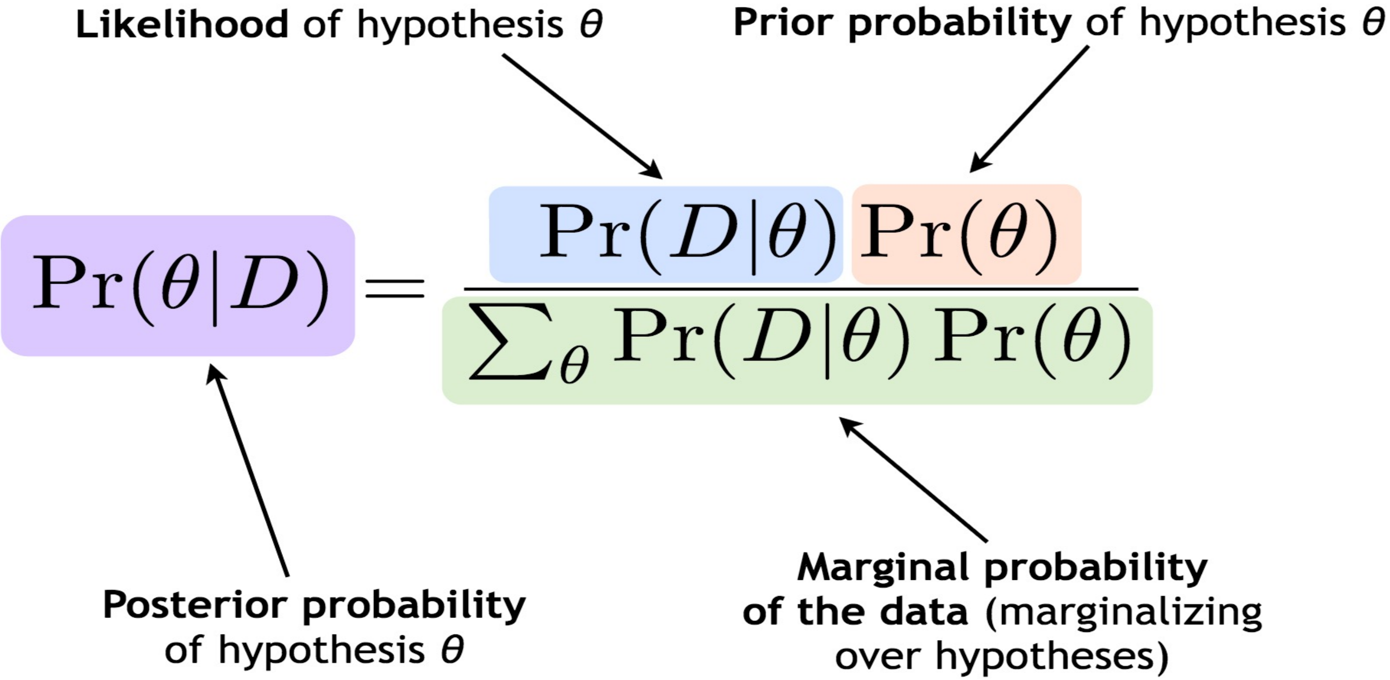
- Tree topology is obtained
- Branch lengths and parameters of the DNA substitution model are optimized
- Different topologies (with branch lengths and DNA substitution model parameters optimized) are compared based on their likelihood as the optimality criterion
- The topology with the highest likelihood needs to be found

A Bayesian approach compared to ML

- The likelihood is the probability of observing the data given a hypothesis
    - L = Pr (D | **θ**).
- **In ML** we search for the parameter values of the model that maximize the likelihood function

- **In a Bayesian analysis, we get the probability of a hypothesis given the data (probability of the tree given the sequences)**
    - We combine the likelihood of a given hypothesis with a prior expectation for this hypothesis to obtain a posterior probability of the hypothesis

Bayes' rule in statistics



Likelihood of hypothesis θ

Prior probability of hypothesis θ

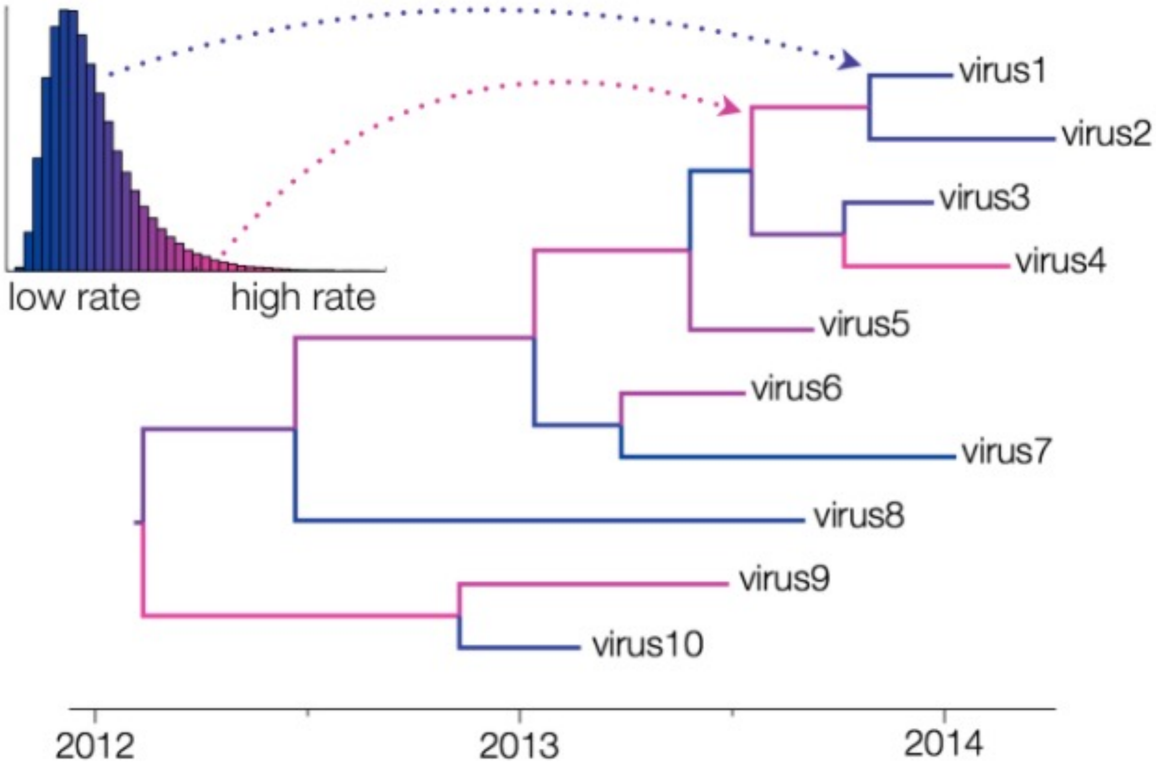
$$\Pr(\theta|D) = \frac{\Pr(D|\theta)\,\Pr(\theta)}{\sum_\theta \Pr(D|\theta)\,\Pr(\theta)}$$

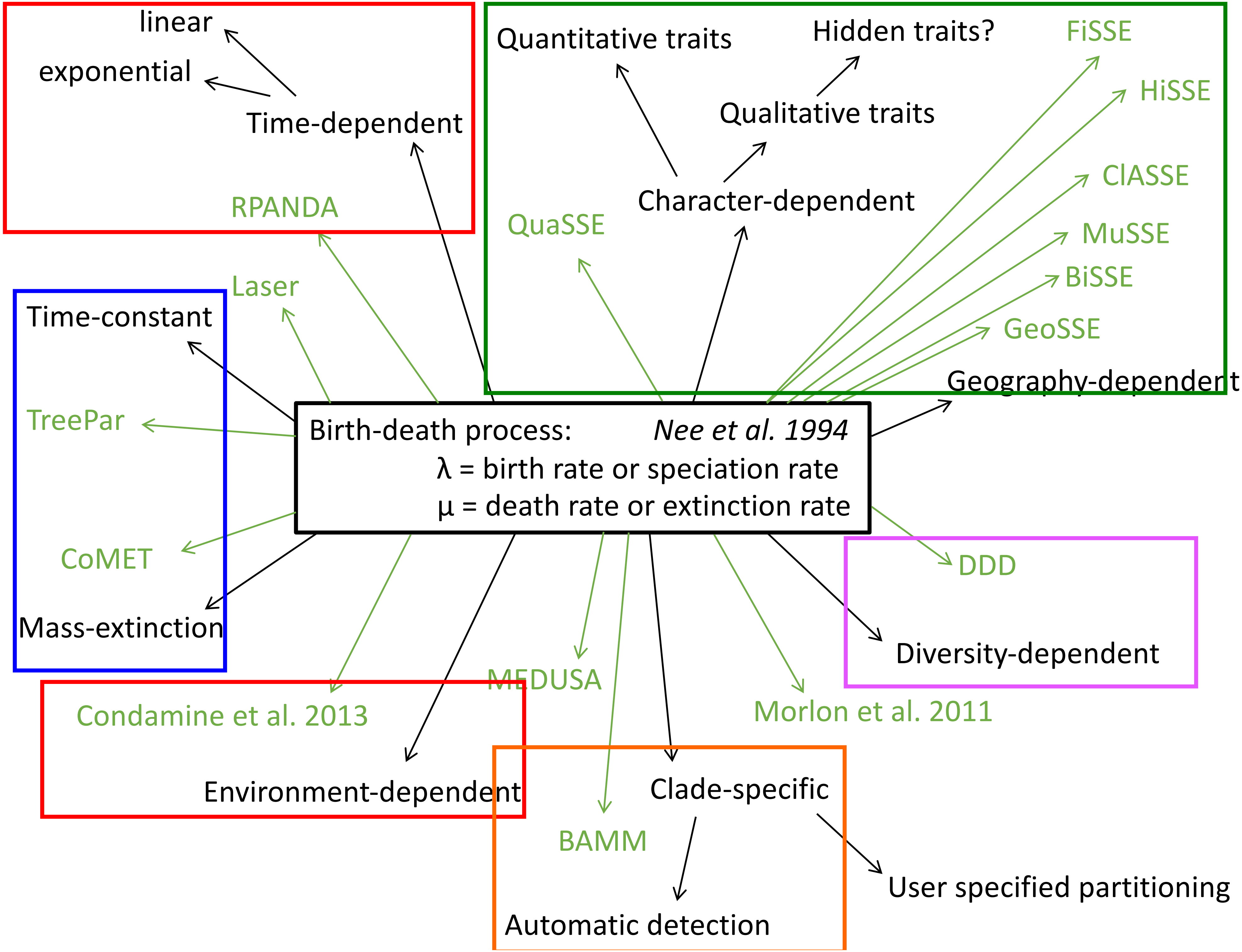
Posterior probability
of hypothesis θ

Marginal probability
of the data (marginalizing
over hypotheses)

Uncorrelated relaxed clocks

- Models available in *BEAST*
  - **Lognormal distribution**
    Most rates cluster around the mean
  - **Exponential distribution**
    Most rates are quite low





Birth-death process: *Nee et al. 1994*

λ = birth rate or speciation rate

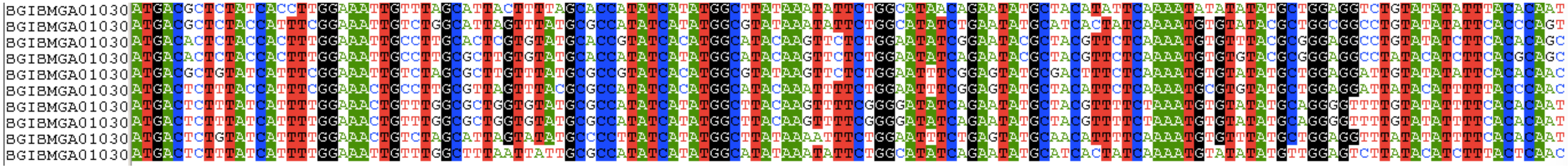
μ = death rate or extinction rate

This week in BIOR90 – how to analyse data

Hours\Days	May 2	May 3	May 4	May 5
9:00-10:00	Module outline by Charlie Cornwallis	Tutorials 3-5 – models, ML, Bayesian (JR, EY)	Tutorial 8 – diversification (JR, EY)	Tutorial 8 – diversification (cont.) (JR, EY)
10:00-12:00	Introduction to alignments, different file formats (NW)	Tutorials 3-5 – models, ML, Bayesian (cont.) (JR, EY)	Tutorial 8 – diversification (cont.) (JR, EY)	Tutorial 9 – mapping characters (SH, EY)
12:00-13:00	Lunch	Lunch	Lunch	Lunch
13:00-14:30	Tutorial 1 – creating datasets (NW, EY)	Tutorial 6 – timing of divergence (EY, NW)	free	Tutorial 9 – mapping characters (cont.) (SH, EY)
14:30-16:00	Tutorial 2 – alignment (NW, EY)	Tutorial 7 – tree manipulation (EY, NW)	free	Tutorial 9 – mapping characters (cont.) (SH, EY)

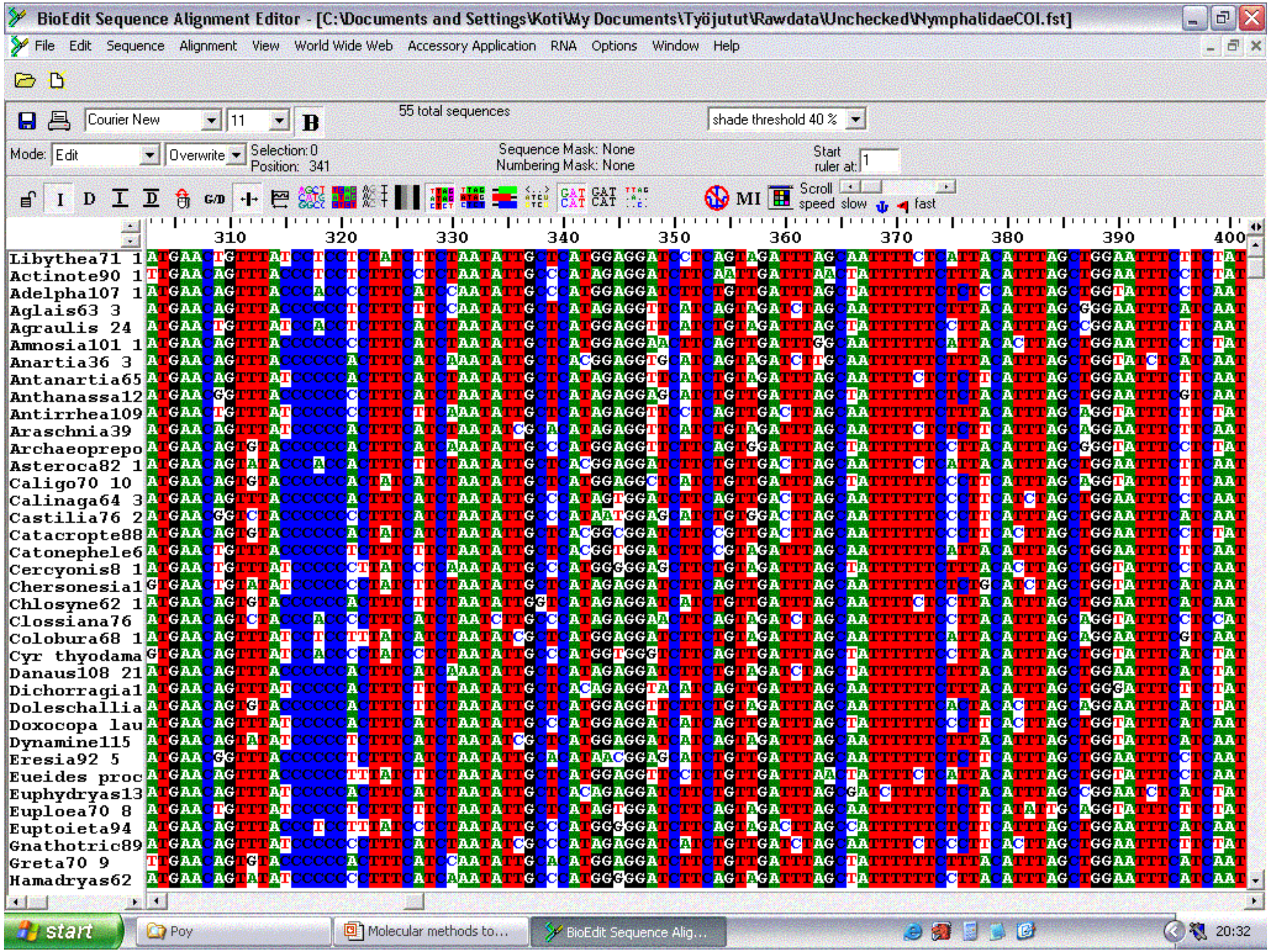
Tutorials on: **https://github.com/NymphalidNiklas/EB2_2023**

Multiple Sequence Alignment



```
BGIBMGA01030 ATGACGCTCTATCACCTTGGAAATTGTTTAGCATTACTTTTAGCACCATATCATATGGCTTATAAATATTCTGGCATAACAGAATATGCTACATATTCAAAATATATATATGCTGGAGGTCTGTATATATTTACACAAT
BGIBMGA01030 ATGACGCTCTACCATTTCGGAAATTGTCTGGCATTAGTTTATGCGCCATATCATATGGCGTATAAATATTCTGGCATATCTGAATATGCATCACTATCAAAATGTGTATACGCTGGCGGCCTGTATATATTCACCCAGT
BGIBMGA01030 ATGACACTCTACCACTTTGGAAATTGCCTTGCACTCGTGTATGCACCGTATCACATGGCATACAAGTTCTCTGGAATATCGGAATACGCTACGTTCTCAAAATGTGTTTACGCGGGAGGCCTGTATATCTTCACACAGC
BGIBMGA01030 ATGACACTCTACCACTTTGGAAATTGCCTTGCGCTTGTGTATGCACCATATCATATGGCATACAAGTTCTCTGGAATATCAGAATACGCTACGTTCTCAAAATGTGTGTACGCGGGAGGCCTATACATCTTCACGCAGC
BGIBMGA01030 ATGACGCTGTATCATTTCGGAAATTGTCTAGCGCTTGTTTATGCGCCGTATCACATGGCGTATAAGTTCTCTGGAATTTCGGAGTATGCGACTTTCTCAAAATGTGTATATGCTGGAGGATTGTATATATTCACACAAC
BGIBMGA01030 ATGACTCTTTACCATTTCGGAAACTGCCTTGCGTTAGTTTACGCGCCATATCACATGGCATACAAATTTTCTGGAATTCGGAGTATGCTACATTCTCAAAATGCGTGTATGCTGGAGGATTATACATTTTTACCCAAC
BGIBMGA01030 ATGACTCTTTATCATTTTGGAAACTGTTTGGCGCTGGTGTATGCGCCATATCATATGGCTTACAAGTTTTCGGGGATATCAGAATATGCTACGTTTTCTAAATGTGTATATGCAGGGGTTTTGTATATTTTCACACAAT
BGIBMGA01030 ATGACTCTTTATCATTTTGGAAACTGTTTGGCGCTGGTGTATGCGCCATATCATATGGCTTACAAGTTTTCGGGGATATCAGAATATGCTACGTTTTCTAAATGTGTATATGCAGGGGTTTTGTATATTTTCACACAAT
BGIBMGA01030 ATGACTCTGTATCATTTTGGAAACTGTCTAGCATTAGTATATGCCCCTTATCATATGGCTTATAAAATTTCTGGAATTTCTGAGTATGCAACATTTTCAAAATGTGTTTATGCTGGAGGTTTATATATTTTCACACAAT
BGIBMGA01030 ATGACTCTTTATCATTTTGGAAATTGTTTGGCTTTAATTATTGCGCCATATCATATGGCATATAAATATTCTGGCATATCAGAATATGCATCACTATCAAAATGTATATATGTTGGAGTCTTATACATCTTTACTCAAC
```

Alignment can be easy…



Courier New 11

55 total sequences

shade threshold 40 %

Mode: Edit Overwrite

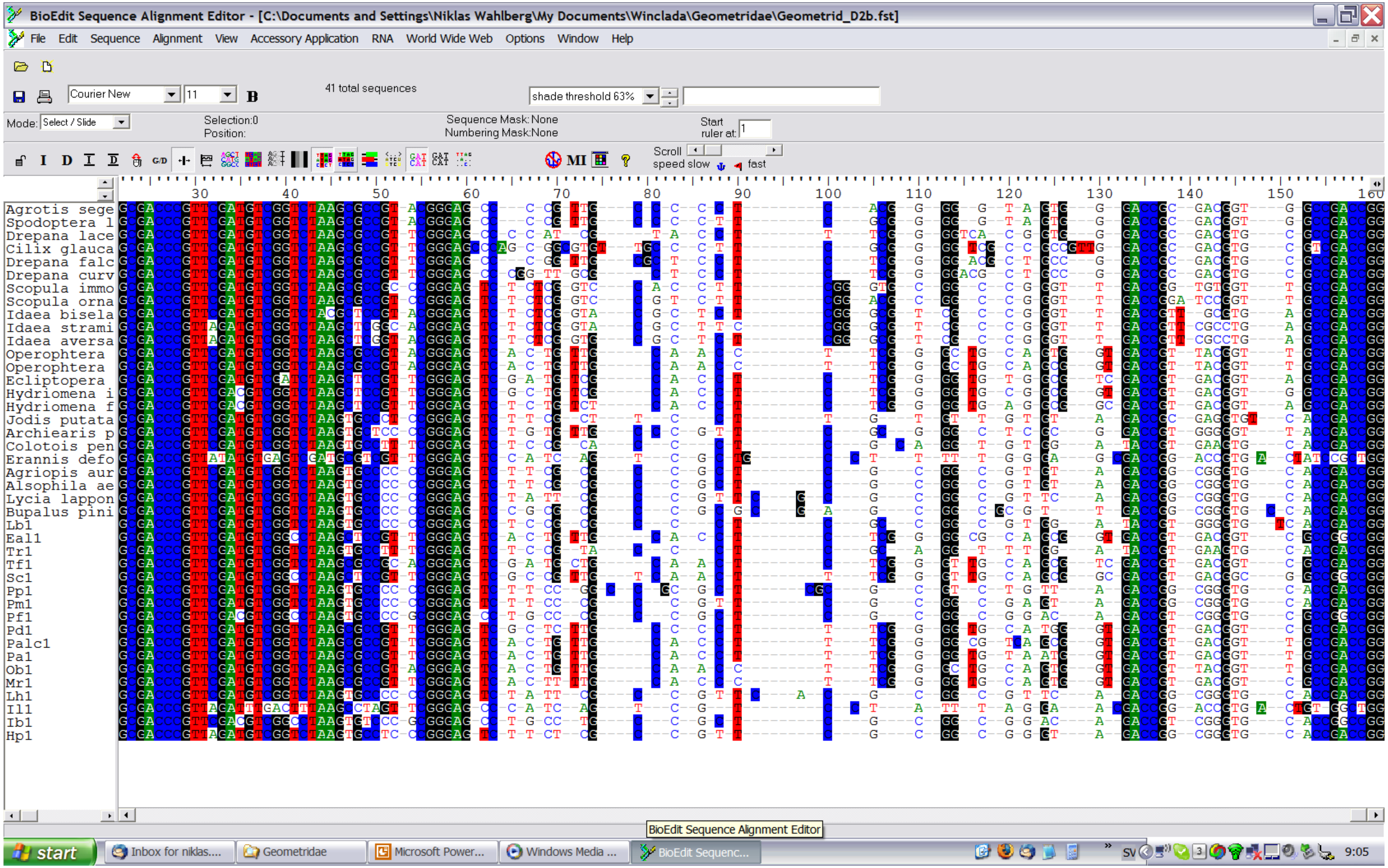
Selection: 0
Position: 341

Sequence Mask: None
Numbering Mask: None

Start ruler at: 1

```
                 310       320       330       340       350       360       370       380       390       400
Libythea71 1 ATGAACTGTTTATCCTCCTCTATCTTCTAATATTGCTCATGGAGGATCCTCAGTAGATTTAGCAATTTTCTCATTACATTTAGCTGGAATTTCTTCTAT
Actinote90 1 TTGAACAGTTTACCCTCCTCTTTCCTCTAATATTGCCCATAGAGGATCTTCAATTGATTTAACTATTTTTTCTTTACATTTAGCTGGAATTTCCTCTAT
Adelpha107 1 ATGAACAGTTTACCCACCCCTTTCATCCAATATTGCCCATGGAGGATCTTCTGTTGATTTAGCTATTTTTTCTCTCCATTTAGCTGGTATTTCCTCAAT
Aglais63 3   ATGAACAGTTTACCCCCCTCTTTCTTCCAATATTGCTCATAGAGGTTCATCAGTAGATCTAGCAATTTTTTCTTTACATTTAGCGGGAATTTCATCAAT
Agraulis 24  ATGAACTGTTTATCCACCTCTTTCATCTAATATTGCTCATGGAGGTTCATCTGTAGATTTAGCTATTTTTTCCTTACATTTAGCCGGAATTTCTTCAAT
Amnosia101 1 ATGAACAGTTTACCCCCCCCTTTCATCTAATATTGCTCATGGAGGAACTTCAGTTGATTTGGCAATTTTTTCATTACACTTAGCTGGAATTTCCTCTAT
Anartia36 3  ATGAACAGTTTACCCCCCACTTTCATCAAATATTGCTCACGGAGGTGCATCAGTAGATCTTGCAATTTTTTCTTTACATTTAGCTGGTATCTCATCAAT
Antanartia65 ATGAACAGTTTATCCCCCACTTTCATCTAATATTGCTCATAGAGGTTCATCTGTAGATTTAGCAATTTTCTCTCTTCATTTAGCTGGAATTTCTTCAAT
Anthanassa12 ATGAACGGTTTACCCCCCCCTTTCATCTAATATTGCTCATAGAGGAGCATCTGTTGATTTAGCTATTTTTTCTCTACATTTAGCTGGAATTTCGTCAAT
Antirrhea109 ATGAACTGTTTATCCCCCCCTTTCTTCAAATATTGCTCATAGAGGTTCCTCAGTTGACTTAGCAATTTTTTCTTTACATTTAGCAGGTATTTCTTCTAT
Araschnia39  ATGAACAGTTTATCCCCCACTTTCATCTAATATCGCACATAGAGGTTCATCTGTAGATTTAGCAATTTTCTCTCTTCATTTAGCAGGAATTTCTTCAAT
Archaeoprepo ATGAACAGTGTACCCCCCACTTTCATCAAATATTGCCCATGGAGGTTCTTCAGTGGATTTAGCTATTTTTTCCTTACATTTAGCGGGTATTTCCTCTAT
Asteroca82 1 ATGAACAGTATACCCACCACTTTCTTCTAATATTGCTCACGGAGGATCTTCTGTTGACTTAGCAATTTTCTCATTACATTTAGCTGGAATTTCTTCAAT
Caligo70 10  ATGAACAGTGTACCCCCCACTATCATCTAATATTGCTCATGGAGGCTCATCTGTTGATTTAGCTATTTTTTCCCTTCATTTAGCAGGTATTTCTTCAAT
Calinaga64 3 ATGAACAGTTTACCCCCCACTTTCATCTAATATTGCCCATAGTGGATCTTCAGTTGACTTAGCAATTTTTTCCCTTCATCTAGCTGGAATTTCCTCAAT
Castilia76 2 ATGAACGGTCTACCCCCCCCTTTCATCTAATATTGCCCATAATGGAGCATCTGTGGACTTAGCAATTTTTTCCCTTCATTTAGCTGGAATTTCATCAAT
Catacropte88 ATGAACAGTGTACCCCCCACTATCATCTAATATTGCTCACGGCGGATCTTCCGTTGACTTAGCAATTTTTTCCCTTCACTTAGCTGGAATTTCCTCTAT
Catonephele6 ATGAACTGTTTACCCCCCTCTTTCTTCTAATATTGCTCACGGTGGATCTTCCGTAGATTTAGCAATTTTTTCATTACATTTAGCTGGAATTTCTTCAAT
Cercyonis8 1 ATGAACTGTTTATCCCCCCTTATCCTCAAATATTGCCCATGGGGGAGCTTCTGTAGATTTAGCTATTTTTTCTTTACACTTAGCTGGTATTTCCTCAAT
Chersonesia1 GTGAACTGTATATCCCCCCCTATCTTCTAATATTGCTCATAGAGGATCTTCAGTTGATTTAGCAATTTTTTCTCTGCATCTAGCTGGTATTTCATCAAT
Chlosyne62 1 ATGAACAGTGTACCCCCCACTTTCTTCTAATATTGGTCATAGAGGATCATCTGTTGATTTAGCAATTTTCTCCTTACATTTAGCTGGAATTTCATCAAT
Clossiana76  ATGAACAGTCTACCCACCCCTTTCATCTAATCTTGCCCATAGAGGAACTTCAGTAGATCTAGCAATTTTTTCCTTACATTTAGCAGGTATTTCCTCCAT
Colobura68 1 ATGAACAGTTTATCCTCCTTTATCATCTAATATCGCTCATGGAGGATCTTCTGTAGATTTAGCAATTTTTTCATTACATTTAGCAGGAATTTCGTCAAT
Cyr thyodama GTGAACAGTTTATCCACCCCTATCCTCTAATATTGCCCATGGTGGGTCTTCAGTTGATTTAGCTATTTTTTCCTTACATTTAGCTGGTATTTCATCTAT
Danaus108 21 ATGAACAGTTTACCCCCCACTTTCATCAAATATTGCTCATAGAGGATCTTCTGTAGATCTAGCTATTTTTTCTTTACATTTAGCTGGAATTTCATCTAT
Dichorragia1 ATGAACAGTTTATCCCCCACTTTCTTCTAATATTGCTCACAGAGGTACATCAGTTGATTTAGCAATTTTTTCTTTACATTTAGCTGGGATTTCTTCTAT
Doleschallia ATGAACAGTGTACCCCCCACTTTCTTCTAATATTGCTCATGGAGGTTCTTCTGTAGATTTAGCAATTTTTTCACTACACTTAGCAGGAATTTCATCTAT
Doxocopa lau ATGAACAGTTTATCCCCCACTTTCATCTAATATTGCCCATGGAGGATCATCAGTTGATTTAGCTATTTTTTCCCTTCACTTAGCTGGTATTTCATCAAT
Dynamine115  ATGAACAGTATATCCCCCTCTTTCATCTAATATCGCTCATGGAGGATCATCAGTAGATTTAGCAATTTTTTCTTTACATTTAGCTGGTATTTCATCAAT
Eresia92 5   ATGAACGGTTTACCCCCCTCTTTCATCTAATATTGCACATAACGGAGCATCTGTTGATTTAGCAATTTTTTCTCTTCATTTAGCTGGAATTTCCTCAAT
Eueides proc ATGAACAGTTTACCCCCCTTTATCTTCTAATATTGCTCATGGAGGTTCCTCTGTTGATTTAACTATTTTCTCATTACATTTAGCTGGTATTTCCTCAAT
Euphydryas13 ATGAACAGTTTATCCCCCACTTTCATCTAATATTGCTCACAGAGGATCTTCTGTTGATTTAGCGATCTTTTCTCTACATTTAGCCGGAATCTCATCTAT
Euploea70 8  ATGAACTGTTTATCCCCCTCTTTCTTCTAATATTGCTCATAGTGGATCTTCAGTAGATTTAGCAATTTTTTCTCTTCATATTGCAGGTATTTCTTCTAT
Euptoieta94  ATGAACAGTTTACCCTCCTTTATCCTCTAATATTGCCCATGGGGGATCTTCAGTAGACTTAGCCATTTTTTCTCTTCATTTAGCTGGAATTTCATCAAT
Gnathotric89 ATGAACAGTTTATCCCCCCCTTTCATCTAATATCGCCCATAGAGGATCATCTGTTGATCTAGCAATTTTCTCCCTTCACTTAGCTGGAATTTCTTCTAT
Greta70 9    TTGAACAGTGTACCCCCCACTTTCATCCAATATTGCACATGGAGGATCTTCTGTTGATTTAGCTATTTTTTCTTTACATTTAGCTGGAATTTCATCAAT
Hamadryas62  ATGAACAGTATATCCCCCCCTTTCATCAAATATTGCCCATGGGGGATCTTCAGTAGATTTAGCTATTTTTTCCTTACATTTAGCTGGAATTTCATCAAT
```

…or difficult



BioEdit Sequence Alignment Editor - [C:\Documents and Settings\Niklas Wahlberg\My Documents\Winclada\Geometridae\Geometrid_D2b.fst]

41 total sequences

shade threshold 63%

Mode: Select / Slide

Selection:0
Position:

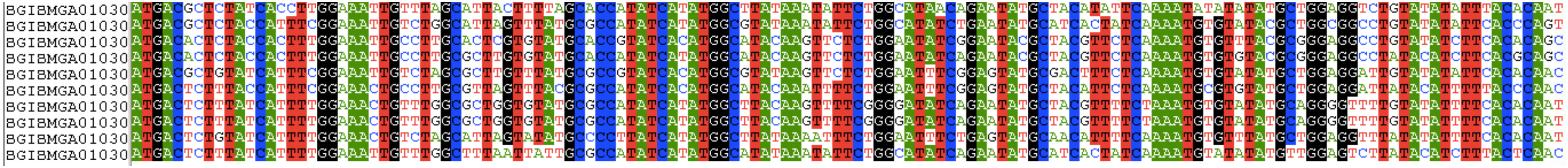
Sequence Mask: None
Numbering Mask:None

Start ruler at: 1

```
Agrotis sege
Spodoptera l
Drepana lace
Cilix glauca
Drepana falc
Drepana curv
Scopula immo
Scopula orna
Idaea bisela
Idaea strami
Idaea aversa
Operophtera
Operophtera
Ecliptopera
Hydriomena i
Hydriomena f
Jodis putata
Archiearis p
Colotois pen
Erannis defo
Agriopis aur
Alsophila ae
Lycia lappon
Bupalus pini
Lb1
Ea11
Tr1
Tf1
Sc1
Pp1
Pm1
Pf1
Pd1
Palc1
Pa1
Ob1
Mr1
Lh1
I11
Ib1
Hp1
```

Homology: Definition

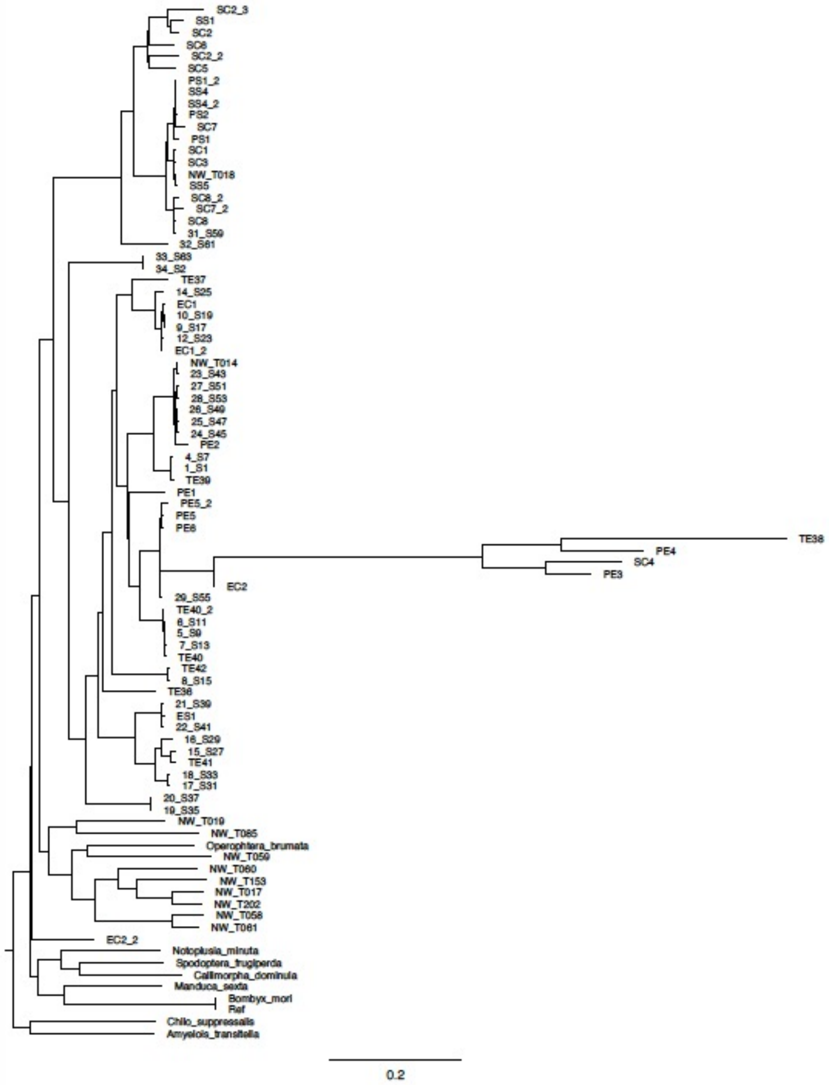
- Homology: similarity that is the result of inheritance from a common ancestor - identification and analysis of homologies is central to phylogenetic systematics
- An alignment is a hypothesis of positional homology between bases/amino acids



```
BGIBMGA01030 ATGACGCTCTATCACCTTGGAAATTGTTTAGCATTACTTTTAGCACCATATCATATGGCTTATAAATATTCTGGCATAACAGAATATGCTACATATTCAAAATATATATATGCTGGAGGTCTGTATATATTTACACAAT
BGIBMGA01030 ATGACGCTCTACCATTTCGGAAATTGTCTGGCATTAGTTTATGCGCCATATCATATGGCGTATAAATATTCTGGCATATCTGAATATGCATCACTATCAAAATGTGTATACGCTGGCGGCCTGTATATATTCACCCAGT
BGIBMGA01030 ATGACACTCTACCACTTTGGAAATTGCCTTGCACTCGTGTATGCACCGTATCACATGGCATACAAGTTCTCTGGAATATCGGAATACGCTACGTTCTCAAAATGTGTTTACGCGGGAGGCCTGTATATCTTCACACAGC
BGIBMGA01030 ATGACACTCTACCACTTTGGAAATTGCCTTGCGCTTGTGTATGCACCATATCATATGGCATACAAGTTCTCTGGAATATCAGAATACGCTACGTTCTCAAAATGTGTGTACGCGGGAGGCCTATACATCTTCACGCAGC
BGIBMGA01030 ATGACGCTGTATCATTTCGGAAATTGTCTAGCGCTTGTTTATGCGCCGTATCACATGGCGTATAAGTTCTCTGGAATTTCGGAGTATGCGACTTTCTCAAAATGTGTATATGCTGGAGGATTGTATATATTCACACAAC
BGIBMGA01030 ATGACTCTTTACCATTTCGGAAACTGCCTTGCGTTAGTTTACGCGCCATATCACATGGCATACAAATTTTCTGGAATTTCGGAGTATGCTACATTCTCAAAATGCGTGTATGCTGGAGGATTATACATTTTTACCCAAC
BGIBMGA01030 ATGACTCTTTATCATTTTGGAAACTGTTTGGCGCTGGTGTATGCGCCATATCATATGGCTTACAAGTTTTCGGGGATATCAGAATATGCTACGTTTTCTAAATGTGTATATGCAGGGGTTTTGTATATTTTCACACAAT
BGIBMGA01030 ATGACTCTTTATCATTTTGGAAACTGTTTGGCGCTGGTGTATGCGCCATATCATATGGCTTACAAGTTTTCGGGGATATCAGAATATGCTACGTTTTCTAAATGTGTATATGCAGGGGTTTTGTATATTTTCACACAAT
BGIBMGA01030 ATGACTCTGTATCATTTTGGAAACTGTCTAGCATTAGTATATGCCCCTTATCATATGGCTTATAAAATTTCTGGAATTTCTGAGTATGCAACATTTTCAAAATGTGTTTATGCTGGAGGTTTATATATTTTCACACAAT
BGIBMGA01030 ATGACTCTTTATCATTTTGGAAATTGTTTGGCTTTAATTATTGCGCCATATCATATGGCATATAAATATTCTGGCATATCAGAATATGCATCACTATCAAAATGTATATATGTTGGAGTCTTATACATCTTTACTCAAC
```

Multiple sequence alignment- goals

- To generate a concise, information-rich summary of sequence data
- Alignments can be treated as models that can be used to test hypotheses
- Does this model of events accurately reflect known biological evidence?



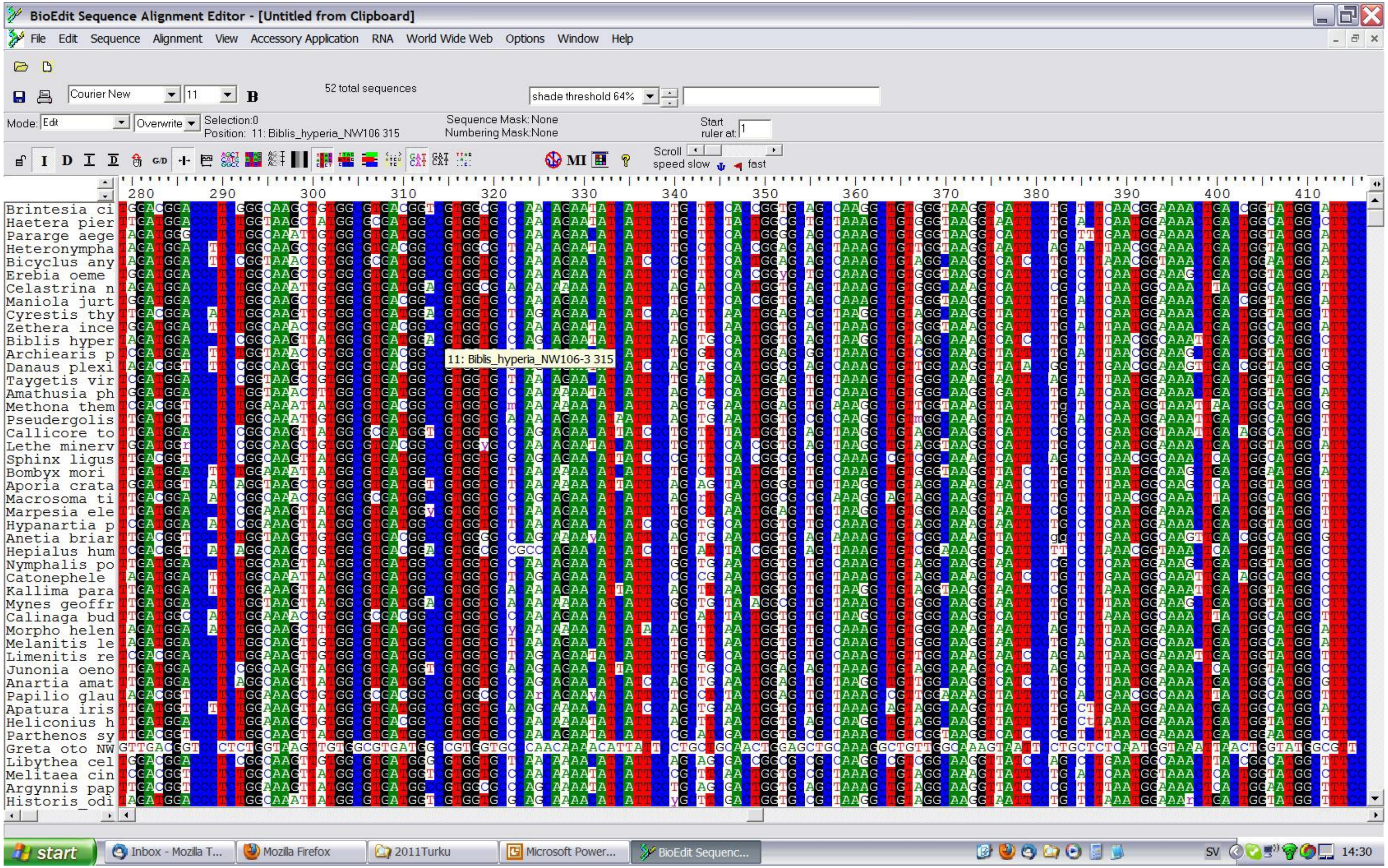
Multiple sequence alignment

- Manual
- Dynamic programming
- Heuristic methods
  - Progressive alignment
  - Consistency-based scoring
  - Iterative refinement methods

Manual alignment - reasons

- Might be carried out because:
- Alignment is easy
- There is some extraneous information (structural)
- Automated alignment methods have encountered a local minimum problem
- An automated alignment method can be "improved"

Protein-coding genes can often be manually aligned



BioEdit Sequence Alignment Editor - [Untitled from Clipboard]

52 total sequences

Courier New 11

shade threshold 64%

Mode: Edit Overwrite

Selection:0
Position: 11: Biblis_hyperia_NW106 315

Sequence Mask: None
Numbering Mask: None

Start ruler at: 1

11: Biblis_hyperia_NW106-3 315

How to align these sequences:

```
AGGGCTTTAA
AGGCTA
AATGGCTCTAA
GGAGCCCTAA
```

How to align these sequences:

```
A-GGGCTTTAA
A--GGCT--A-
AATGGCTCTAA
GGAG-CCCTAA
```

How to align these sequences:

```
-AGGGCTTTAA
-A-GGC--TA-
AATGGCTCTAA
-GGAGCCCTAA
```

Multiple sequence alignment

- Is not easy! How to be objective?
- Dynamic programming
- Heuristic methods
    - Progressive alignment
    - Consistency-based scoring
    - Iterative refinement methods

Dynamic programming

- For two sequences, the best alignment can be found by scoring all possible pairs of aligned nucleotides and penalizing gaps
- An optimality criterion
- Time and computer memory needed grows exponentially with number of sequences
- Becomes impossible to align more than 4 sequences of modest length
- Fails to fully exploit phylogeny and does not incorporate an evolutionary model

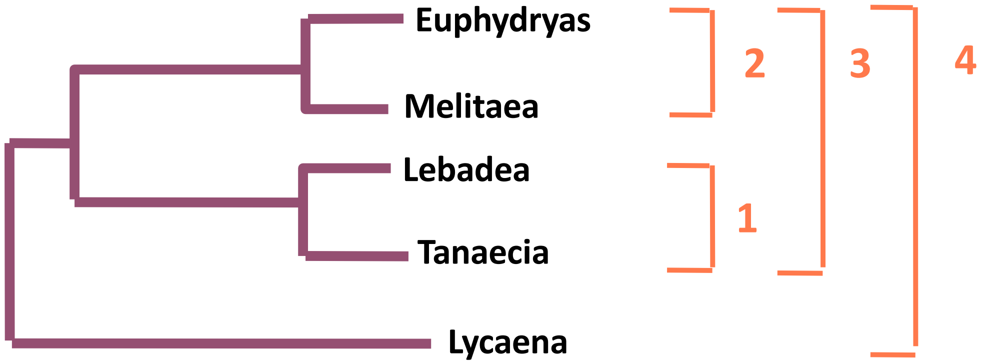
Heuristics: Progressive alignment

- Devised by Feng and Doolittle in 1987
- A heuristic method and as such is not guaranteed to find the ‘optimal’ alignment
- Requires *n-1+n-2+n-3...n-n+1* pairwise alignments as a starting point
- Most successful implementation is Clustal
    - ClustalW
    - ClustalX

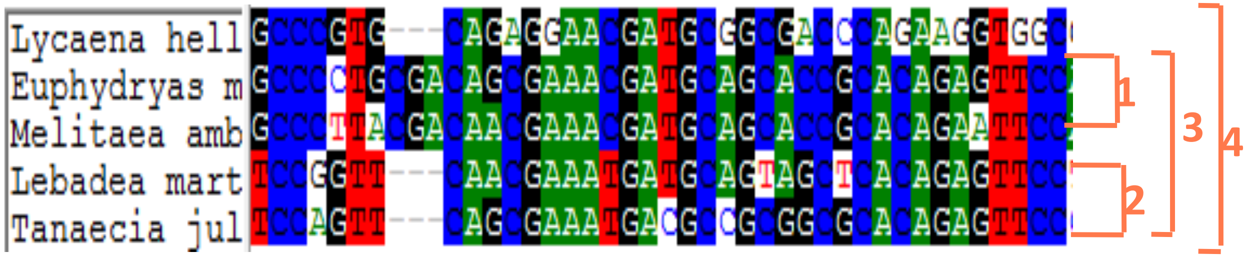
Overview of Clustal procedure

Euphydryas	1	-				
Melitaea	2	.17	-			
Lebadea	3	.59	.60	-		
Tanaecia	4	.59	.59	.13	-	
Lycaena	5	.77	.77	.75	.75	-

Quick pairwise alignment: calculate distance matrix



Neighbour-joining tree (guide tree)



```
Lycaena hell GCCCGTG---CAGAGGAACGATGCGGCGACCCAGAAGGTGGC
Euphydryas m GCCCCTGCGACAGCGAAACGATGCAGCACCGCACAGAGTTCC
Melitaea amb GCCCTTACGACAACGAAACGATGCAGCACCGCACAGAATTCC
Lebadea mart TCCGGTT---CAACGAAATGATGCAGTAGCTCACAGAGTTCC
Tanaecia jul TCCAGTT---CAGCGAAATGACGCCGCGGCGCACAGAGTTCC
```

Progressive alignment following guide tree

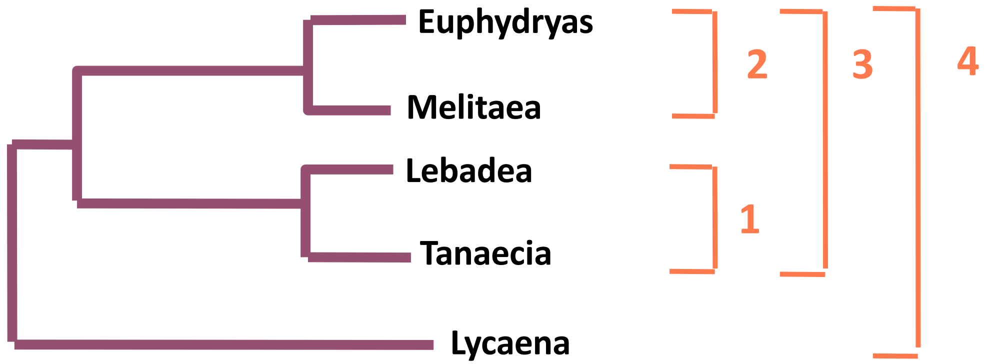
Clustal - pairwise alignments

- First perform all possible pairwise alignments between each pair of sequences
- Calculate the ‘distance’ between each pair of sequences based on these isolated pairwise alignments
- Generate a distance matrix

Taxon	***Euphydryas***	***Melitaea***	***Lebadea***	***Tanaecia***	***Lycaena***
Euphydryas	-				
Melitaea	0.17	-			
Lebadea	0.59	0.60	-		
Tanaecia	0.59	0.59	0.13	-	
Lycaena	0.77	0.77	0.75	0.75	-

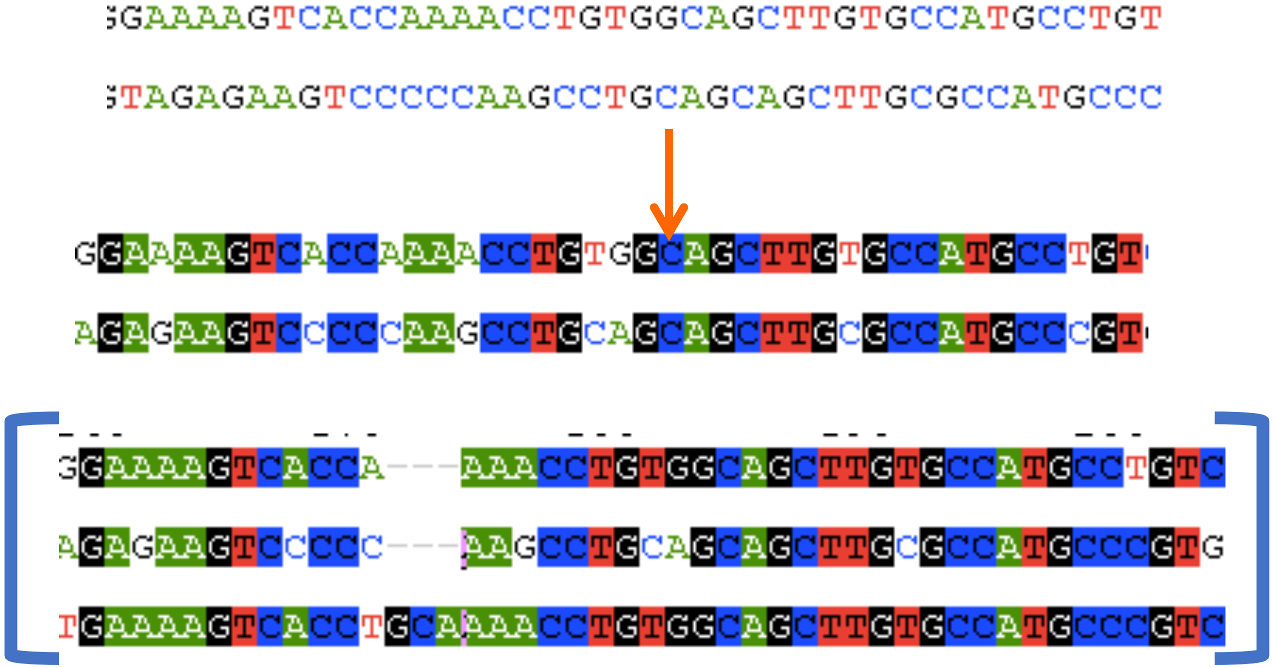
Clustal - guide tree

- Generate a Neighbour-Joining ‘guide tree’ from these pairwise distances
- This guide tree gives the order in which the progressive alignment will be carried out



Multiple alignment- first pair

- Align the two most closely-related sequences first
- This alignment is then ‘fixed’ and will never change
- If a gap is to be introduced subsequently, then it will be introduced in the same place in both sequences, but their relative alignment remains unchanged

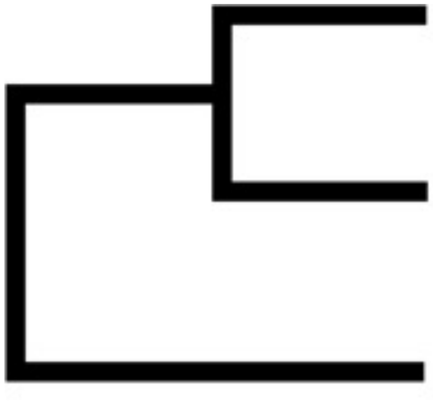

GGAAAAGTCACCAAAACCTGTGGCAGCTTGTGCCATGCCTGT
GTAGAGAAGTCCCCCAAGCCTGCAGCAGCTTGCGCCATGCCC
GGAAAAGTCACCAAAACCTGTGGCAGCTTGTGCCATGCCTGT
AGAGAAGTCCCCCAAGCCTGCAGCAGCTTGCGCCATGCCCGT
GGAAAAGTCACCA---AAACCTGTGGCAGCTTGTGCCATGCCTGTC
AGAGAAGTCCCCC---AAGCCTGCAGCAGCTTGCGCCATGCCCGTG
TGAAAAGTCACCTGCAAAACCTGTGGCAGCTTGTGCCATGCCCGTC

Clustal - decision time

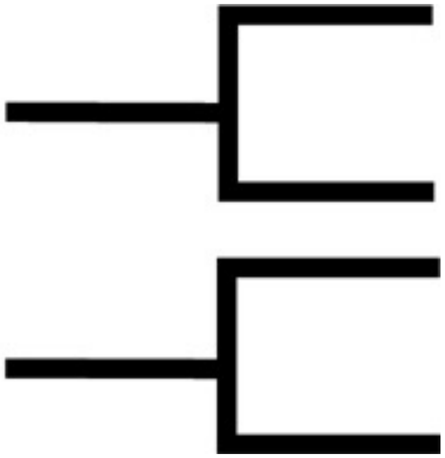
- Consult the guide tree to see what alignment is performed next.
  - Align a third sequence to the first two

  Or

  - Align two entirely different sequences to each other.



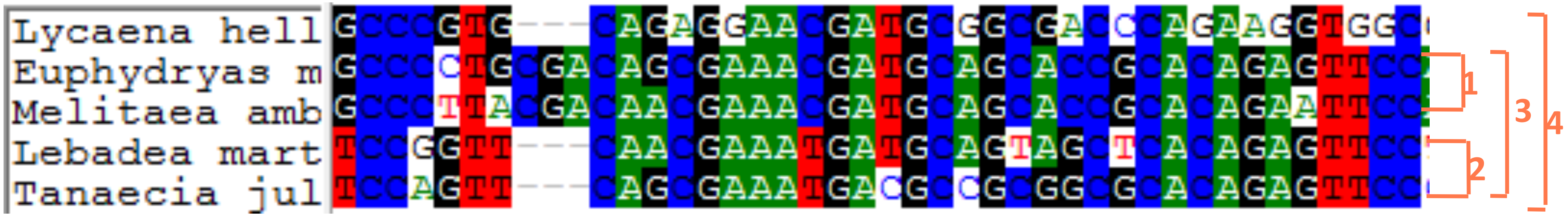
Option 1



Option 2

Clustal - progression

- The alignment is progressively built up in this way, with each step being treated as a pairwise alignment, sometimes with each member of a 'pair' having more than one sequence



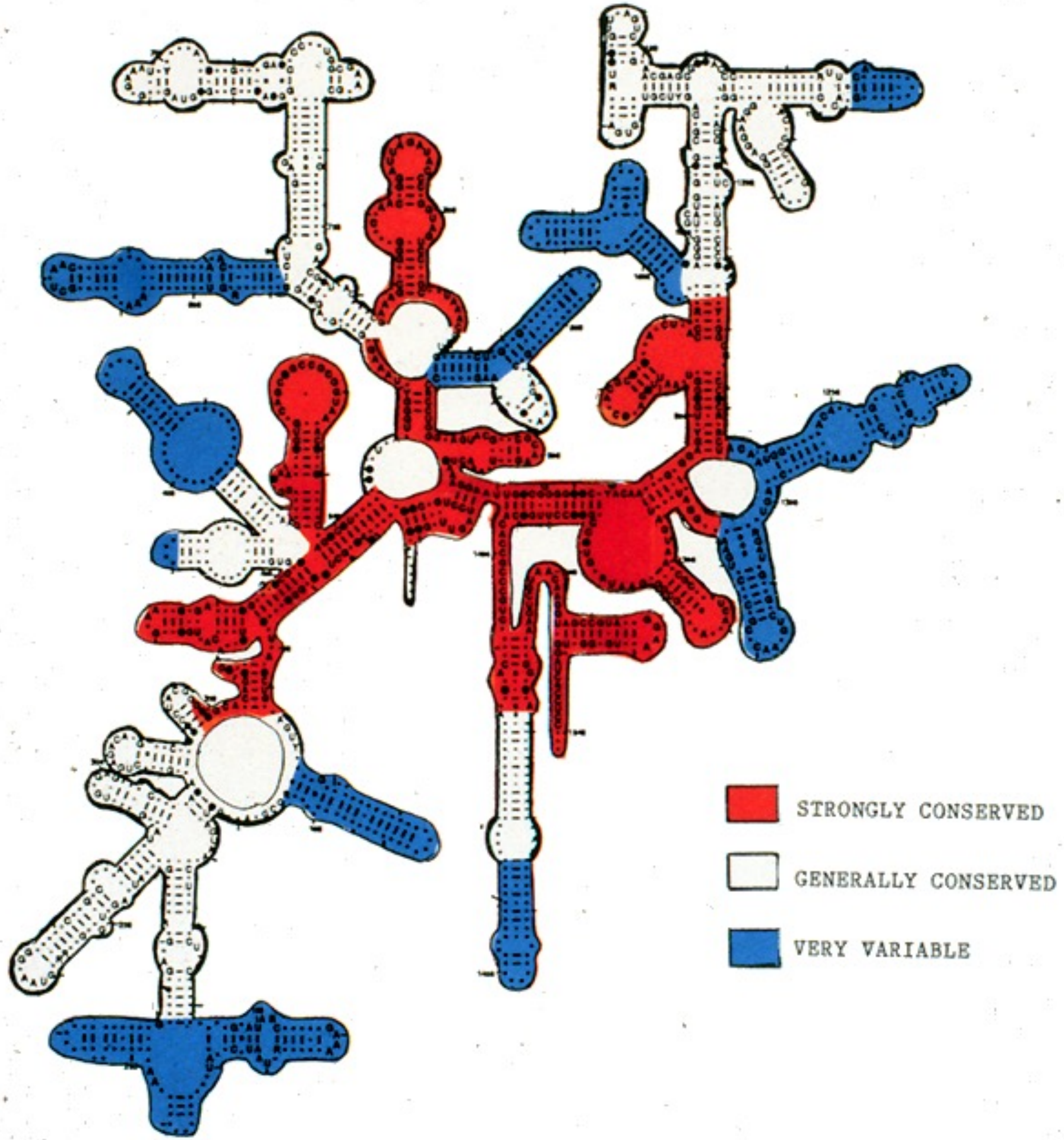
```
Lycaena hell GCCCGTG---CAGAGGAACGATGCGGCGACCCAGAAGGTGGC
Euphydryas m GCCCCTGCGACAGCGAAACGATGCAGCACCGCACAGAGTTCC
Melitaea amb GCCCTTACGACAACGAAACGATGCAGCACCGCACAGAATTCC
Lebadea mart TCCGGTT---CAACGAAATGATGCAGTAGCTCACAGAGTTCC
Tanaecia jul TCCAGTT---CAGCGAAATGACGCCGCGGCGCACAGAGTTCC
```

Clustal - good points/bad points

- Advantages:
  - Speed
- Disadvantages:
  - Hierarchic structure introduced that is not necessarily phylogenetic
  - No way of quantifying whether or not the alignment is good
  - No way of knowing if the alignment is ‘correct’
  - Local minimum problem. If an error is introduced early in the alignment process, it is impossible to correct this later in the procedure
  - Arbitrary alignment

Increasing the sophistication of the alignment process

- Should we treat all the sequences in the same way?
  - some sequences are closely related and some sequences are distant relatives.
- Should we treat all positions in the sequences as though they were the same?
  - they might have different functions and different locations in the 3-dimensional structure.
  - codon structure – how to retain this?


🟥 STRONGLY CONSERVED
⬜ GENERALLY CONSERVED
🟦 VERY VARIABLE

Consistency-based scoring

- One way to avoid the problems of getting stuck in local minima or fixed gaps
- Based on optimizing a multiple alignment using information from all pairwise alignments
- Identifies those nucleotides that are aligned most consistently across the different alignments
- Used in e.g. T-Coffee

Iterative refinement methods

- Initial alignments split into two groups randomly
- Within groups the alignment is kept fixed
- Dynamic programming used to align the two groups to each other
- This is repeated until score converges
- Used in e.g. Muscle and MAFFT

Using models in alignment

- New methods are being developed all the time
- Latest methods include using a Bayesian statistic framework, DNA evolutionary models and alignment concomitantly with estimation of phylogentic relationships
- Still not feasible with a moderately sized dataset

Bottom line

- Alignments are extremely important in phylogenetics
- A bad alignment means many wrong statements of homology, which means pure rubbish as output
- A good alignment can be hard to attain

The Tree

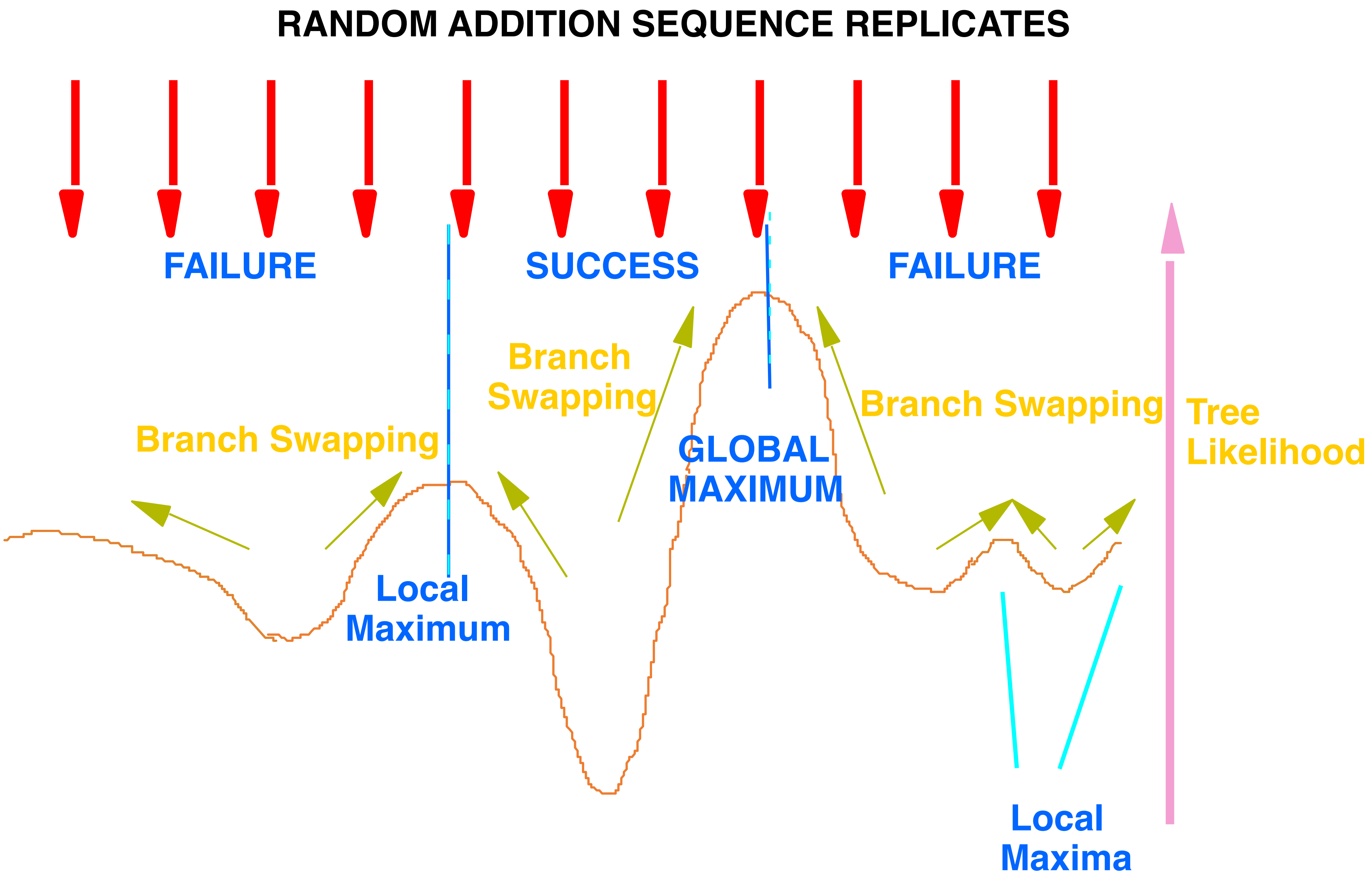
Finding the optimal trees

Numbers of possible trees for N taxa

1	1
2	1
3	1
4	3
5	15
6	105
7	945
8	10395
9	135135
10	2027025
11	34459425
12	654729075
13	13749310575
14	316234143225
15	7905853580625
16	213458046676875
17	6190283353629370
18	191898783962510625
19	6332659870762850625
20	221643095476699771875 (2 x 10^{20})
50	3 x 10^{74}

How can we find the most optimal tree?

Tree space may be populated by local optima and islands of optimal trees



Finding optimal trees - exact solutions

- Exact solutions can only be used for small numbers of taxa

- Exhaustive search examines all possible trees
- Branch and bound does not examine all trees, but will find optimal tree(s)

- Typically used for problems with 10 -20 taxa

Finding optimal trees - heuristics

- The number of possible trees increases faster than exponentially with the number of taxa making exhaustive searches impractical for many data sets (an NP-complete problem)

- Heuristic methods are used to search tree space for optimal trees by building or selecting an initial tree and swapping branches to search for better ones

- The trees found are not guaranteed to be optimal - they are best guesses

Finding optimal trees - heuristics

- Stepwise addition

  Asis - the order in the data matrix

  Closest -starts with shortest 3-taxon tree, adds taxa in order that produces the least increase in tree length (greedy heuristic)

  Simple - the first taxon in the matrix is taken as a reference - taxa are added to it in the order of their decreasing similarity to the reference

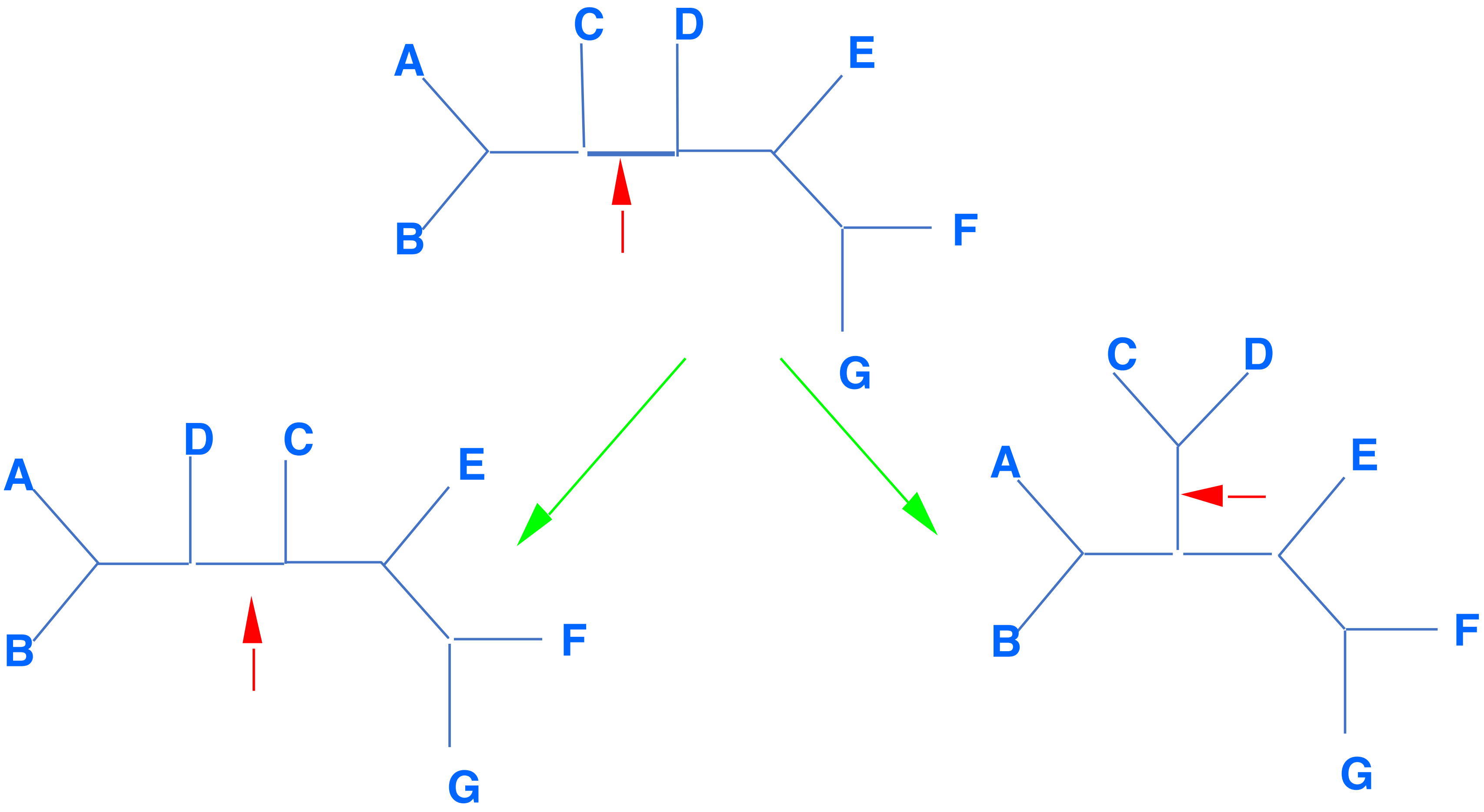
  Random - taxa are added in a random sequence, many different sequences can be used

Finding optimal trees – branch swapping

- Nearest neighbor interchange (NNI)
- Subtree pruning and regrafting (SPR)
- Tree bisection and reconnection (TBR)

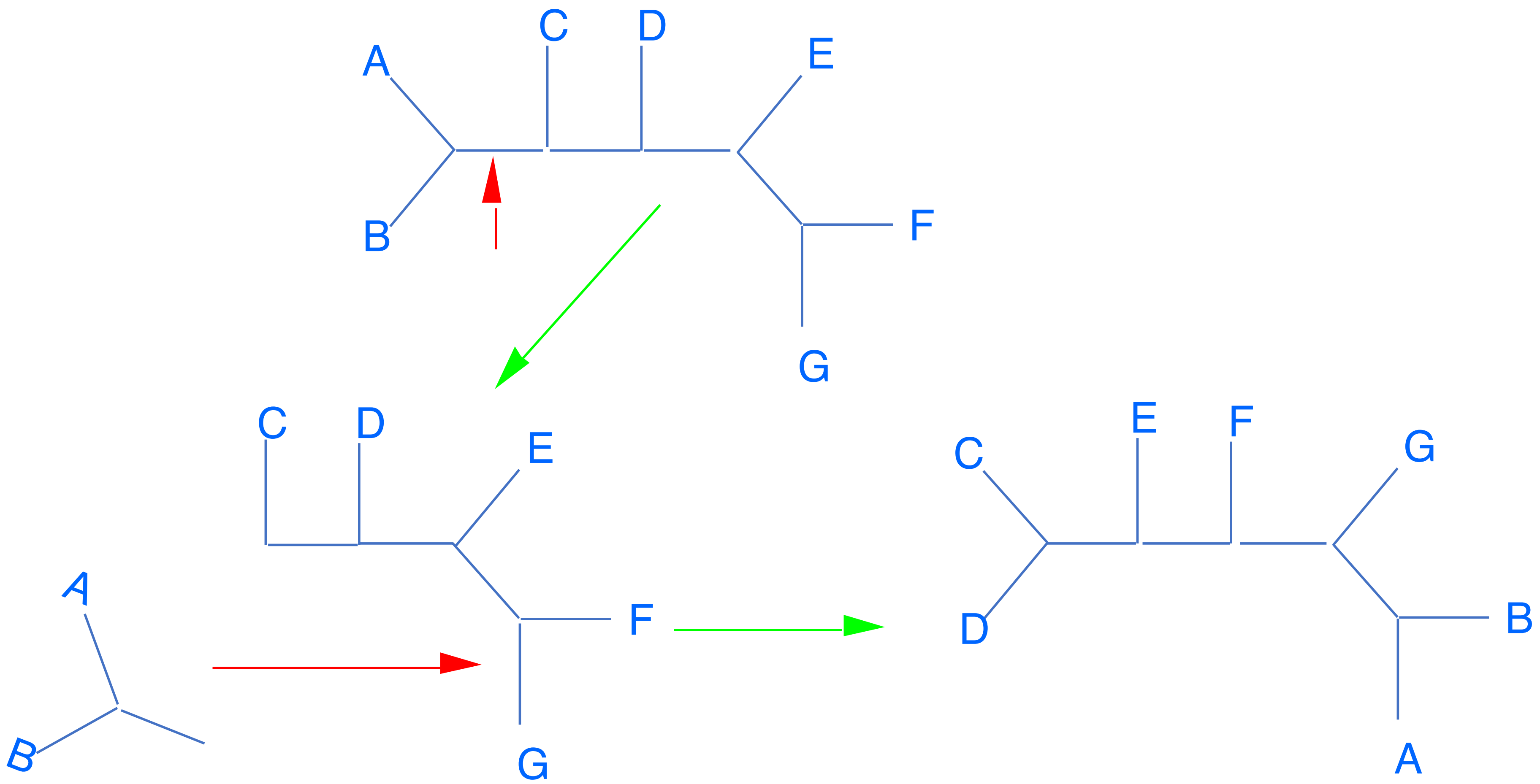
Finding optimal trees - heuristics

Nearest neighbor interchange (NNI)



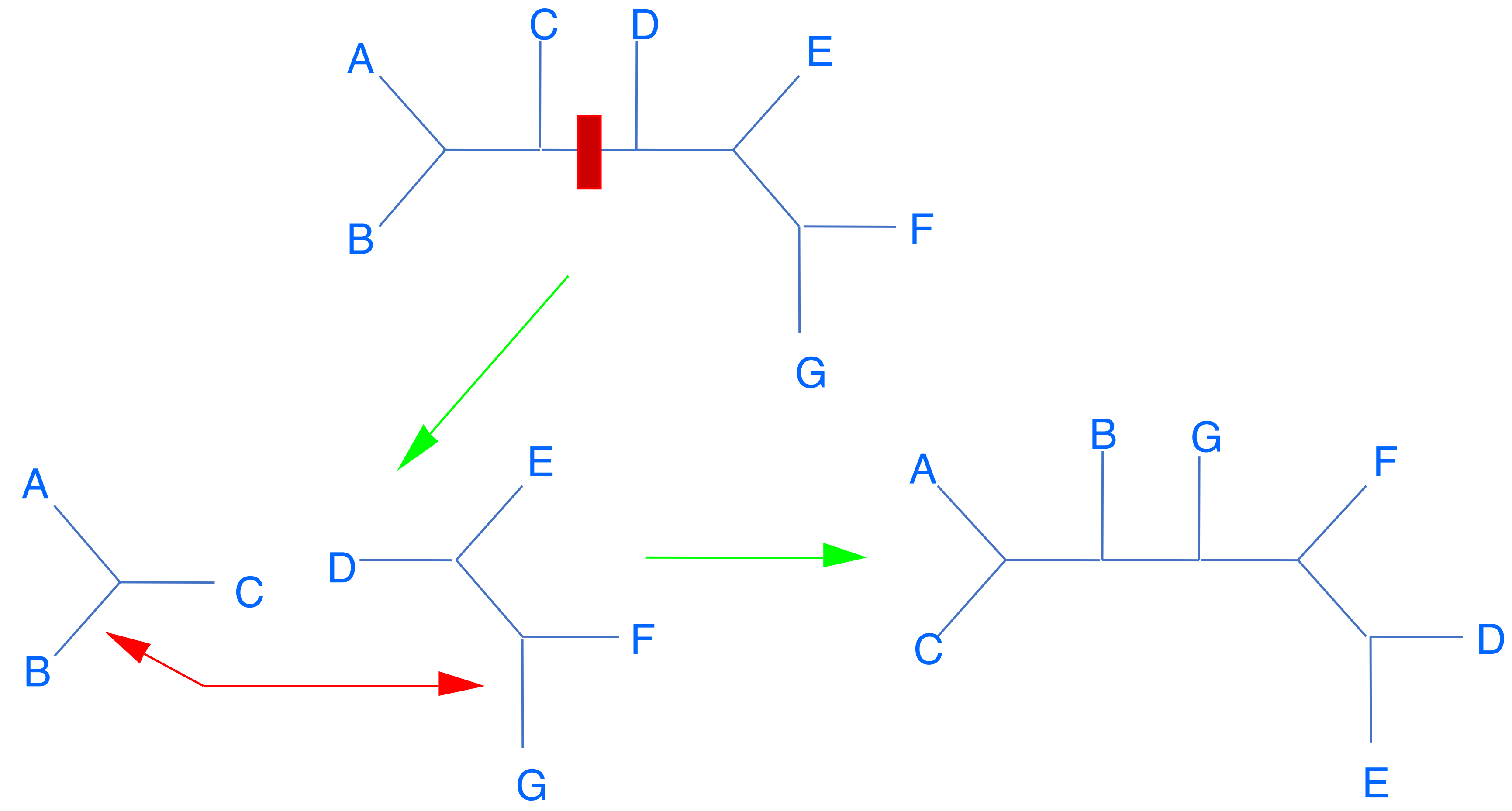
Finding optimal trees - heuristics

Subtree pruning and regrafting (SPR)



Finding optimal trees - heuristics

Tree bisection and reconnection (TBR)

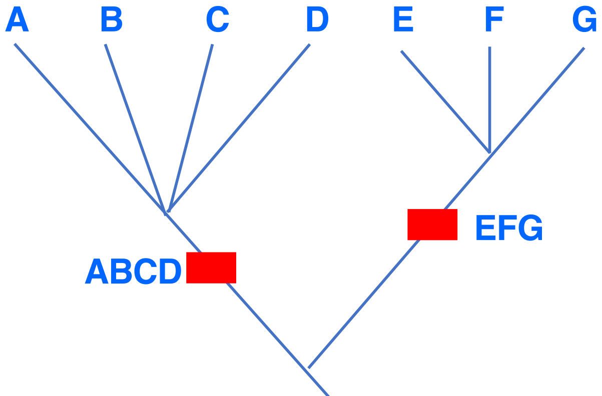


Searching with topological constraints

- Topological constraints are user-defined phylogenetic hypotheses
- Can be used to find optimal trees that either:

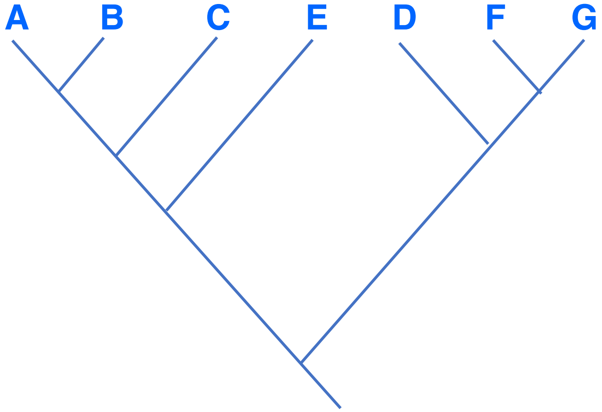
  1. include a specified clade or set of relationships
  2. exclude a specified clade or set of relationships (reverse constraint)

Searching with topological constraints



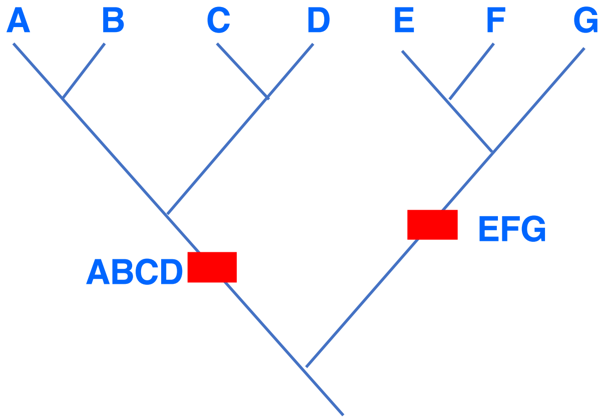
CONSTRAINT TREE

((A,B,C,D)(E,F,G))



Compatible with reverse constraint tree

Incompatible with constraint tree

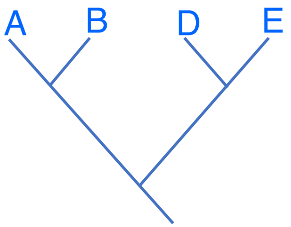


Compatible with constraint tree

Incompatible with reverse constraint tree

Searching with topological constraints
backbone constraints

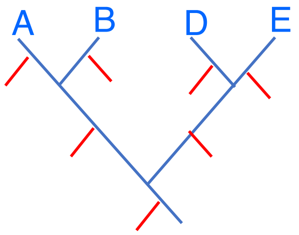
- Backbone constraints specify relationships among a subset of the taxa



BACKBONE CONSTRAINT

((A,B)(D,E))

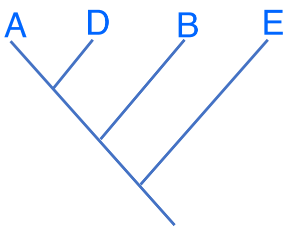
relationships of taxon C are not specified



/ possible positions of taxon C

Compatible with backbone constraint

Incompatible with reverse constraint



Incompatible with backbone constraint

Compatible with reverse constraint

Consensus methods

Multiple optimal trees

- Many methods can yield multiple equally optimal trees
- We can further select among these trees with additional criteria, but
- Typically, relationships common to all the optimal trees are summarised with *consensus trees*

Consensus methods

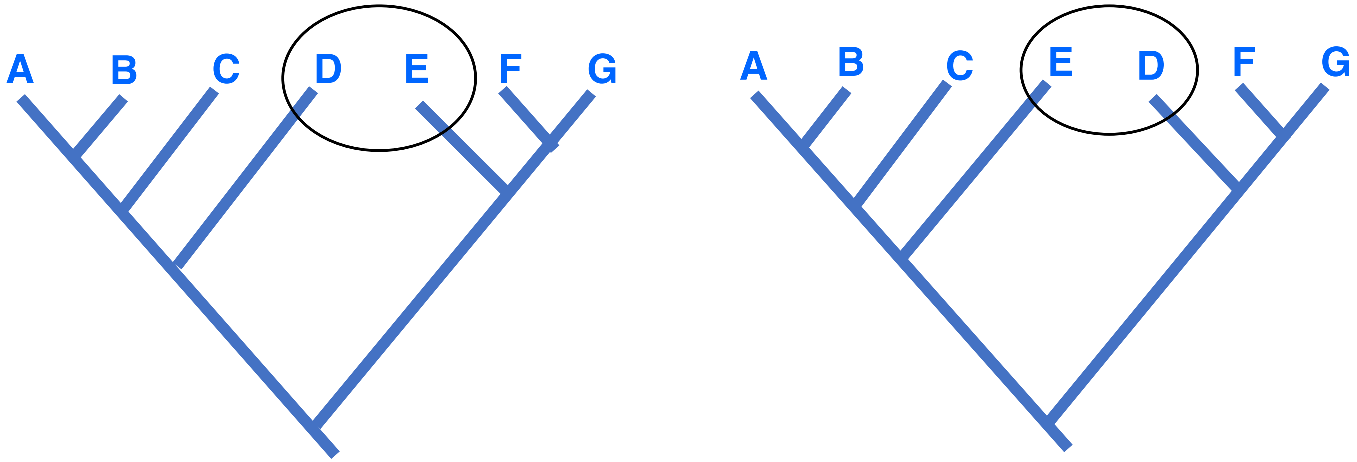
- A consensus tree is a summary of the agreement among a set of fundamental trees
- There are many consensus methods that differ in:
  1. the kind of agreement
  2. the level of agreement
- Consensus methods can be used with multiple trees from a single analysis or from multiple analyses

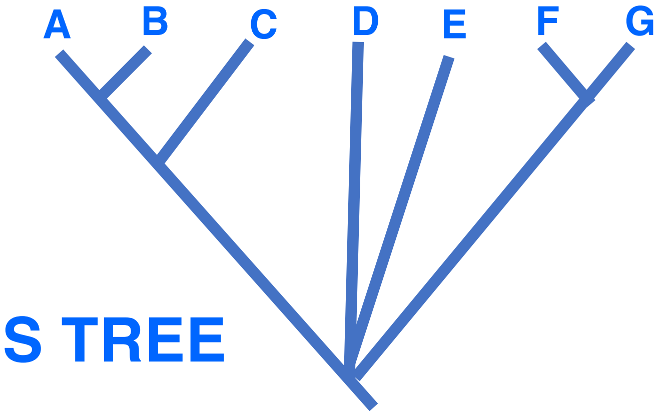
Strict consensus methods

- Strict consensus methods require agreement across all the fundamental trees
- They show only those relationships that are unambiguously supported by the parsimonious interpretation of the data
- The commonest method (*strict component consensus*) focuses on clades/components/full splits
- This method produces a consensus tree that includes all and only those full splits found in all the fundamental trees
- Other relationships (those in which the fundamental trees disagree) are shown as unresolved polytomies

Strict consensus methods

TWO FUNDAMENTAL TREES





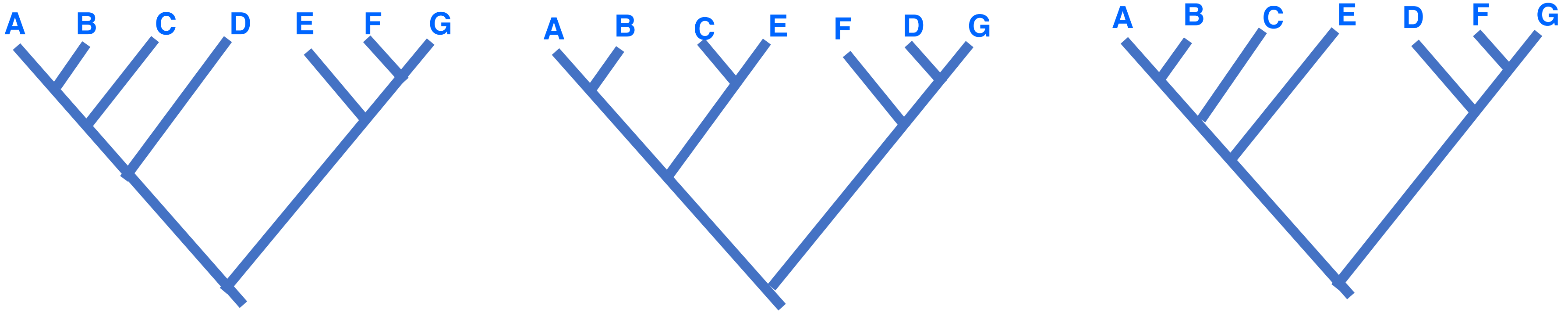
STRICT CONSENSUS TREE

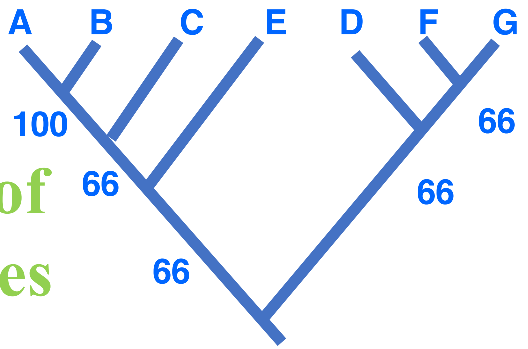
Majority-rule consensus methods

- Majority-rule consensus methods require agreement across a majority of the fundamental trees
- May include relationships that are not supported by the most parsimonious interpretation of the data
- The commonest method focuses on clades/components/full splits
- This method produces a consensus tree that includes all and only those full splits found in a majority (>50%) of the fundamental trees
- Other relationships are shown as unresolved polytomies
- Of particular use in bootstrapping

Majority rule consensus

THREE FUNDAMENTAL TREES





Numbers indicate frequency of clades in the fundamental trees

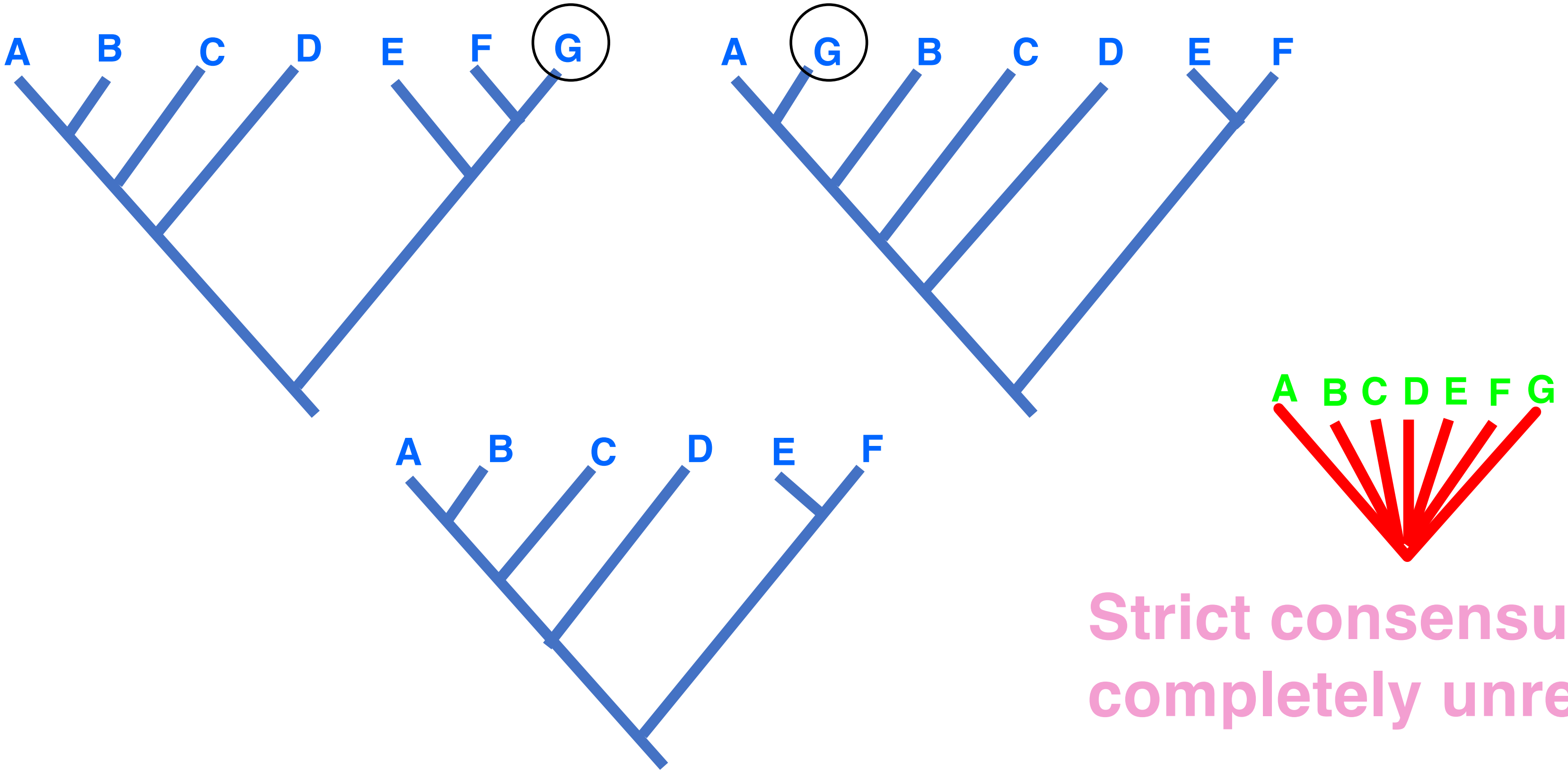
MAJORITY-RULE CONSENSUS TREE

Reduced consensus methods

- Focuses upon any relationships (not just full splits)
- Reduced consensus methods occur in strict and majority-rule varieties
- Other relationships are shown as unresolved polytomies
- May be more sensitive than methods focusing only on clades/components/full splits

Reduced consensus methods

TWO FUNDAMENTAL TREES



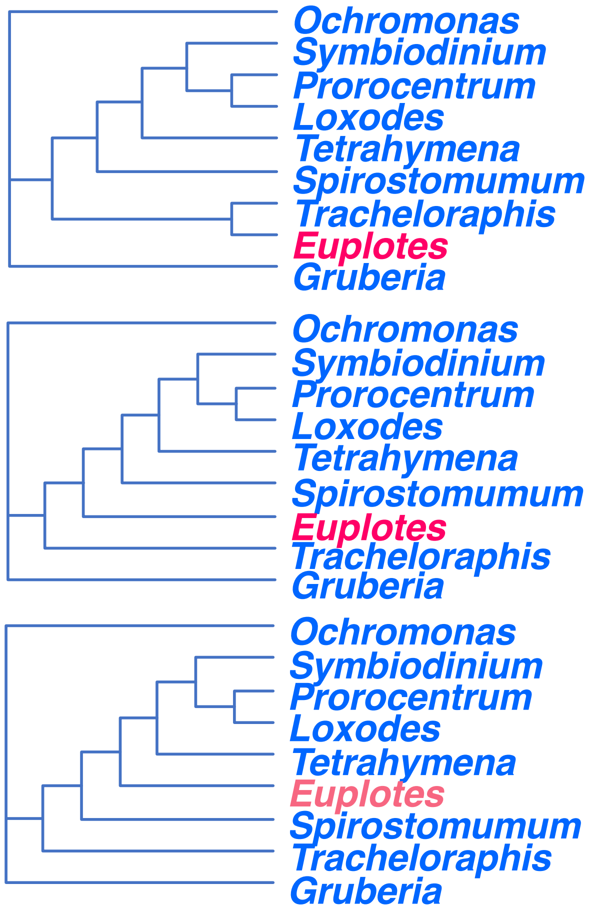
Strict consensus completely unresolved

STRICT REDUCED CONSENSUS TREE

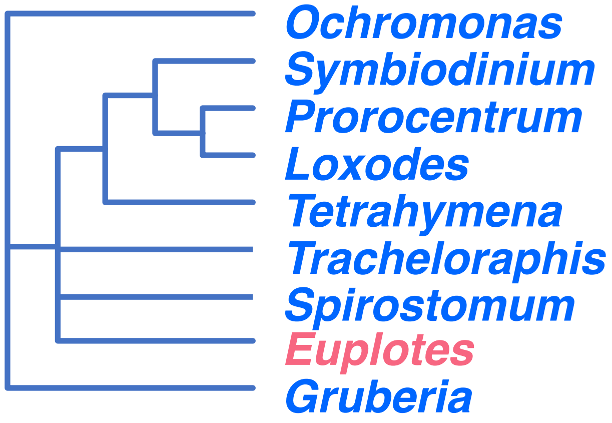
Taxon G is excluded

Consensus methods

Three fundamental trees

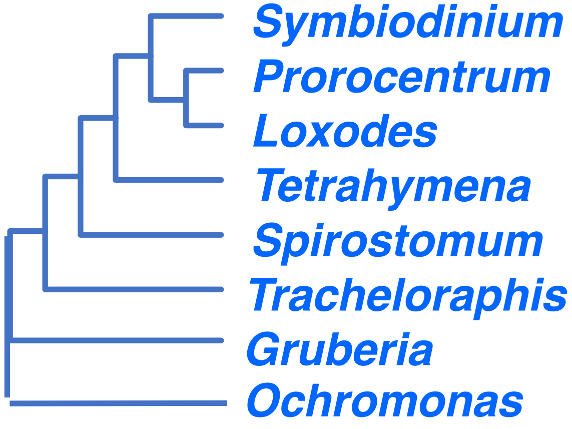


Strict (component)

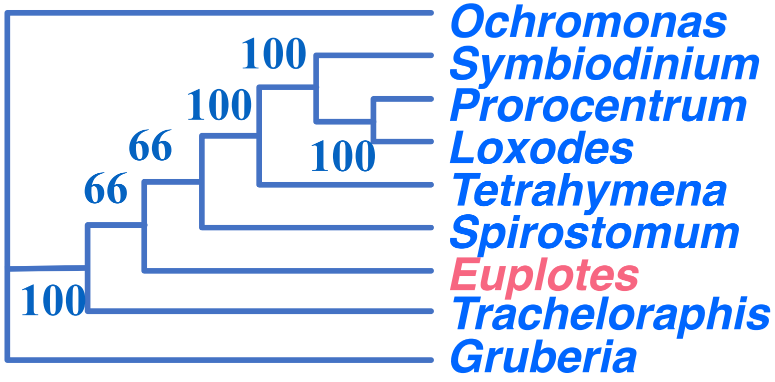


Strict reduced cladistic

Euplotes excluded



Majority-rule



Consensus methods – use

- Currently majority-rule methods mainly used
  - bootstrapping
  - Bayesian methods
- Reduced methods can be useful to identify problem taxa
  - E.g. RogueNaRok
- Strict methods mainly used in parsimony analyses
  - rarely used with molecular data

Take home messages from today

- Statements of homology are the basis of phylogenetics
- Alignments of molecular sequences are very strong statements of positional homology
- Finding an optimal tree is not a trivial task

The Data

File formats and alignments

Computer programs

- Multitude of programs available for free!
- Most have their own input format
- Many are ”black box” programs
- Input files are always simple text files!!!

No good online resource available

http://evolution.gs.washington.edu/phylip/software.html

was an attempt but not updated for a long time

Computer programs - ML

- IQ-TREE (recommended)
- RAxML (recommended)
- PHYML
- GARLI

Computer programs- Bayesian inference

- MrBayes (recommended)
- BEAST (recommended)
- BAMBE
- BayesPhylogenies

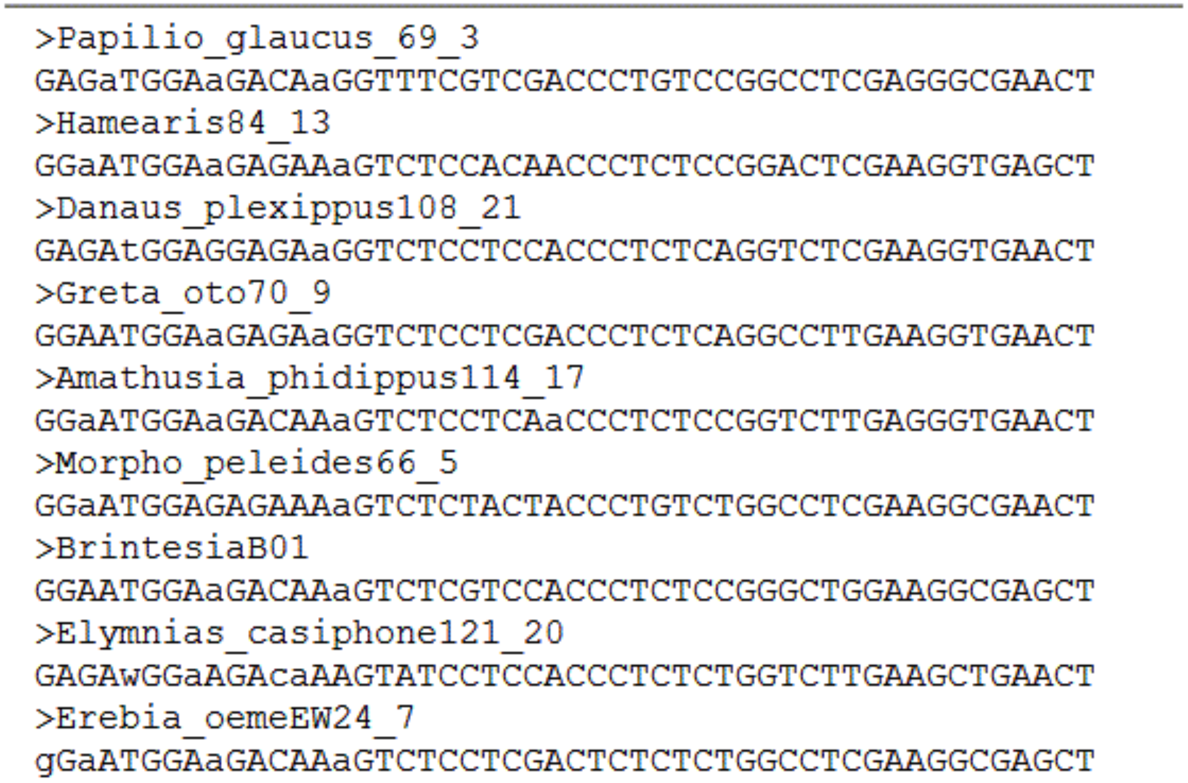
Viewing trees

- FigTree (recommended)
- TreeView
- Winclada
- Dendroscope (for large trees >200 taxa)

Three most common data formats

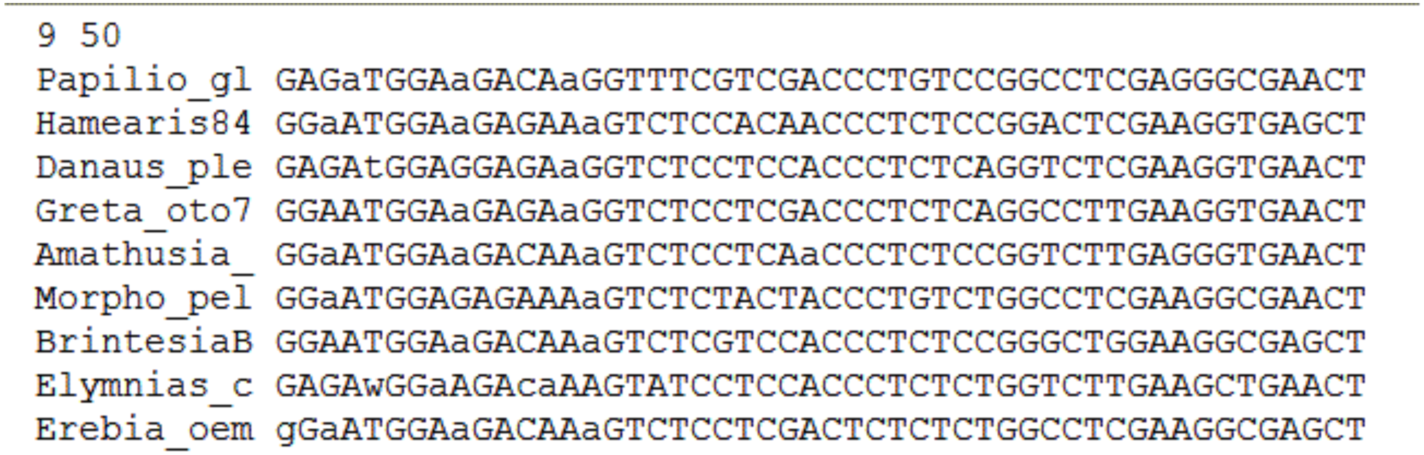
- FASTA
- Phylip
- Nexus

Input format - FASTA



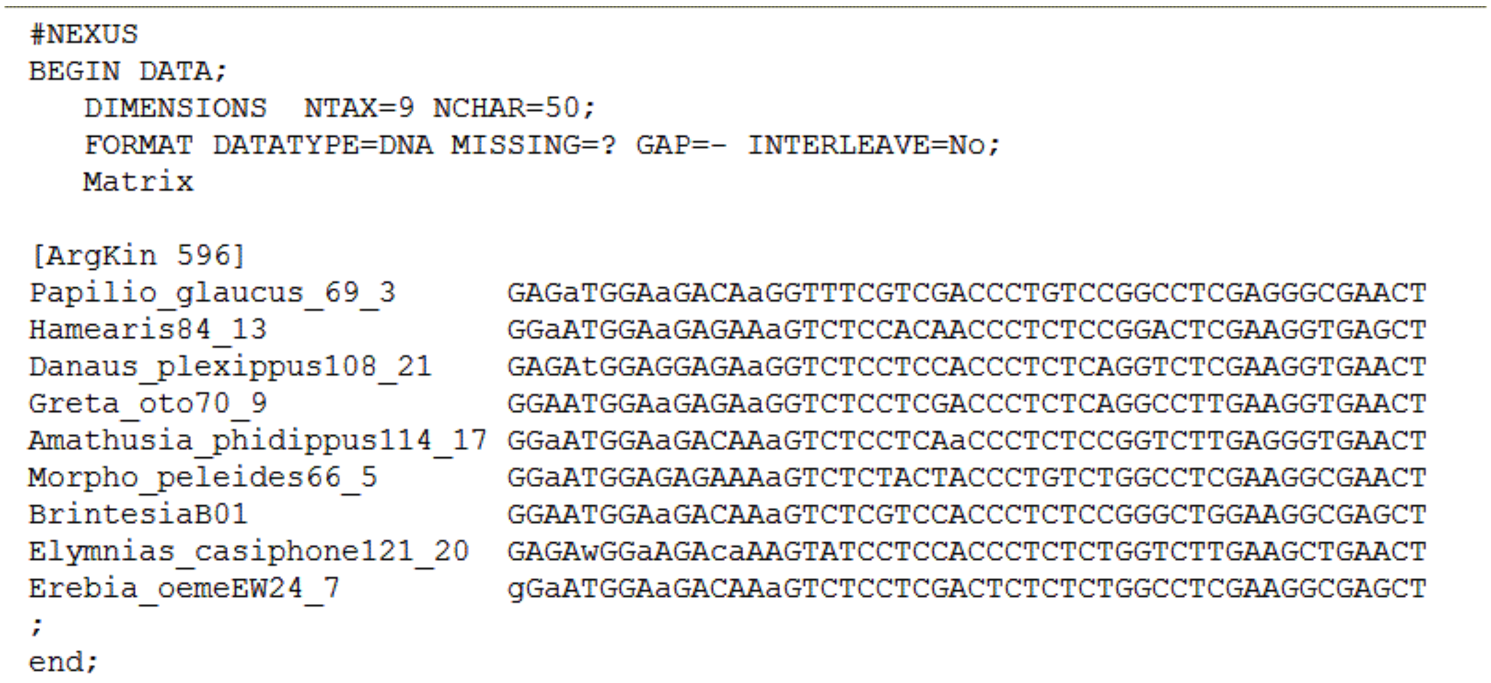
```
>Papilio_glaucus_69_3
GAGaTGGAaGACAaGGTTTCGTCGACCCTGTCCGGCCTCGAGGGCGAACT
>Hamearis84_13
GGaATGGAaGAGAAaGTCTCCACAACCCTCTCCGGACTCGAAGGTGAGCT
>Danaus_plexippus108_21
GAGAtGGAGGAGAaGGTCTCCTCCACCCTCTCAGGTCTCGAAGGTGAACT
>Greta_oto70_9
GGAATGGAaGAGAaGGTCTCCTCGACCCTCTCAGGCCTTGAAGGTGAACT
>Amathusia_phidippus114_17
GGaATGGAaGACAAaGTCTCCTCAaCCCTCTCCGGTCTTGAGGGTGAACT
>Morpho_peleides66_5
GGaATGGAGAGAAAaGTCTCTACTACCCTGTCTGGCCTCGAAGGCGAACT
>BrintesiaB01
GGAATGGAaGACAAaGTCTCGTCCACCCTCTCCGGGCTGGAAGGCGAGCT
>Elymnias_casiphone121_20
GAGAwGGaAGAcaAAGTATCCTCCACCCTCTCTGGTCTTGAAGCTGAACT
>Erebia_oemeEW24_7
gGaATGGAaGACAAaGTCTCCTCGACTCTCTCTGGCCTCGAAGGCGAGCT
```

Input format – PHYLIP



```
9 50
Papilio_gl GAGaTGGAaGACAaGGTTTCGTCGACCCTGTCCGGCCTCGAGGGCGAACT
Hamearis84 GGaATGGAaGAGAAaGTCTCCACAACCCTCTCCGGACTCGAAGGTGAGCT
Danaus_ple GAGAtGGAGGAGAaGGTCTCCTCCACCCTCTCAGGTCTCGAAGGTGAACT
Greta_oto7 GGAATGGAaGAGAaGGTCTCCTCGACCCTCTCAGGCCTTGAAGGTGAACT
Amathusia_ GGaATGGAaGACAAaGTCTCCTCAaCCCTCTCCGGTCTTGAGGGTGAACT
Morpho_pel GGaATGGAGAGAAAaGTCTCTACTACCCTGTCTGGCCTCGAAGGCGAACT
BrintesiaB GGAATGGAaGACAAaGTCTCGTCCACCCTCTCCGGGCTGGAAGGCGAGCT
Elymnias_c GAGAwGGaAGAcaAAGTATCCTCCACCCTCTCTGGTCTTGAAGCTGAACT
Erebia_oem gGaATGGAaGACAAaGTCTCCTCGACTCTCTCTGGCCTCGAAGGCGAGCT
```

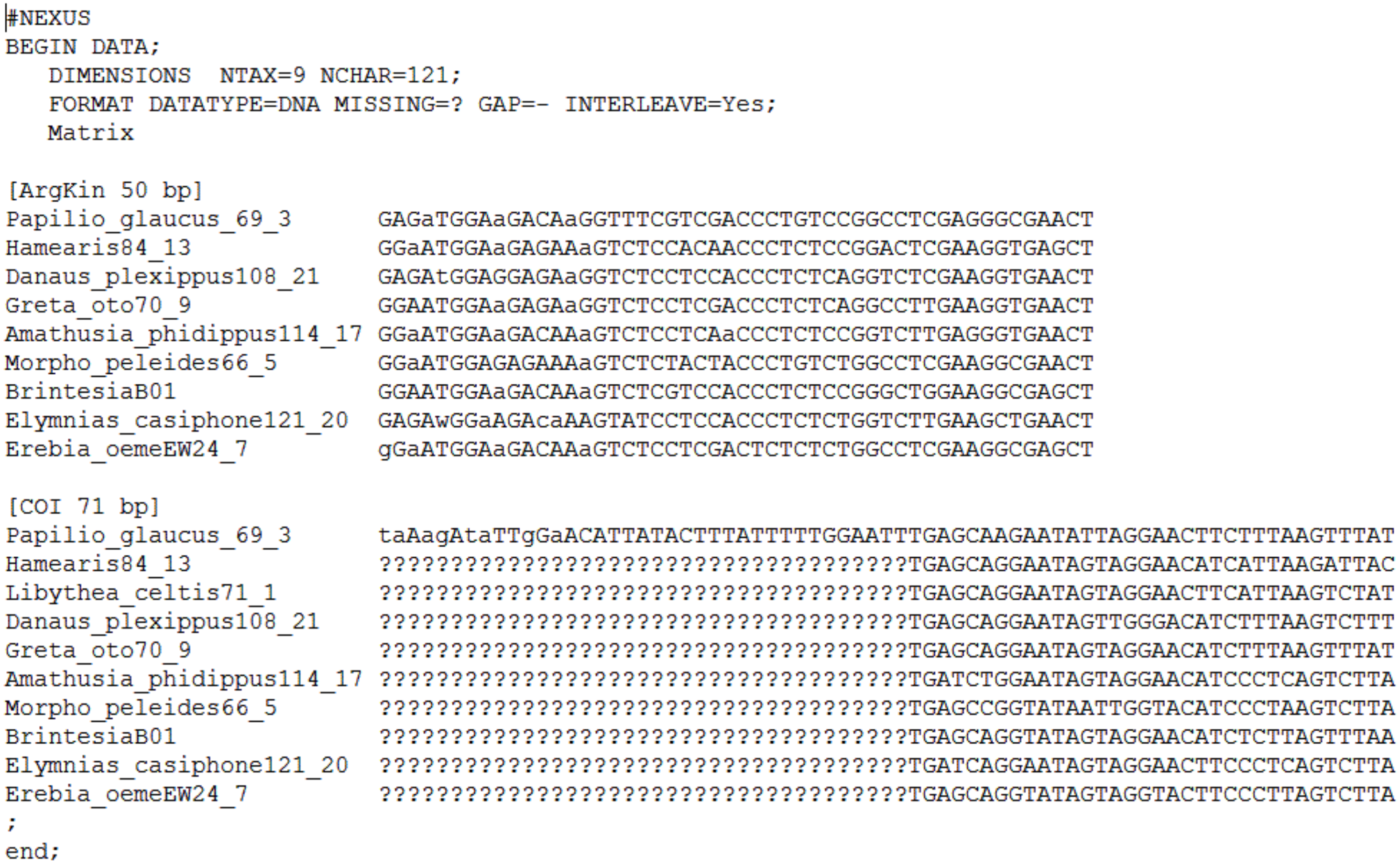
Input format - NEXUS



```
#NEXUS
BEGIN DATA;
   DIMENSIONS  NTAX=9 NCHAR=50;
   FORMAT DATATYPE=DNA MISSING=? GAP=- INTERLEAVE=No;
   Matrix

[ArgKin 596]
Papilio_glaucus_69_3      GAGaTGGAaGACAaGGTTTCGTCGACCCTGTCCGGCCTCGAGGGCGAACT
Hamearis84_13             GGaATGGAaGAGAAaGTCTCCACAACCCTCTCCGGACTCGAAGGTGAGCT
Danaus_plexippus108_21    GAGAtGGAGGAGAaGGTCTCCTCCACCCTCTCAGGTCTCGAAGGTGAACT
Greta_oto70_9             GGAATGGAaGAGAaGGTCTCCTCGACCCTCTCAGGCCTTGAAGGTGAACT
Amathusia_phidippus114_17 GGaATGGAaGACAAaGTCTCCTCAaCCCTCTCCGGTCTTGAGGGTGAACT
Morpho_peleides66_5       GGaATGGAGAGAAAaGTCTCTACTACCCTGTCTGGCCTCGAAGGCGAACT
BrintesiaB01              GGAATGGAaGACAAaGTCTCGTCCACCCTCTCCGGGCTGGAAGGCGAGCT
Elymnias_casiphone121_20  GAGAwGGaAGAcaAAGTATCCTCCACCCTCTCTGGTCTTGAAGCTGAACT
Erebia_oemeEW24_7         gGaATGGAaGACAAaGTCTCCTCGACTCTCTCTGGCCTCGAAGGCGAGCT
;
end;
```

Input format – NEXUS interleaved



```
#NEXUS
BEGIN DATA;
   DIMENSIONS  NTAX=9 NCHAR=121;
   FORMAT DATATYPE=DNA MISSING=? GAP=- INTERLEAVE=Yes;
   Matrix

[ArgKin 50 bp]
Papilio_glaucus_69_3      GAGaTGGAaGACAaGGTTTCGTCGACCCTGTCCGGCCTCGAGGGCGAACT
Hamearis84_13             GGaATGGAaGAGAAaGTCTCCACAACCCTCTCCGGACTCGAAGGTGAGCT
Danaus_plexippus108_21    GAGAtGGAGGAGAaGGTCTCCTCCACCCTCTCAGGTCTCGAAGGTGAACT
Greta_oto70_9             GGAATGGAaGAGAaGGTCTCCTCGACCCTCTCAGGCCTTGAAGGTGAACT
Amathusia_phidippus114_17 GGaATGGAaGACAAaGTCTCCTCAaCCCTCTCCGGTCTTGAGGGTGAACT
Morpho_peleides66_5       GGaATGGAGAGAAAaGTCTCTACTACCCTGTCTGGCCTCGAAGGCGAACT
BrintesiaB01              GGAATGGAaGACAAaGTCTCGTCCACCCTCTCCGGGCTGGAAGGCGAGCT
Elymnias_casiphone121_20  GAGAwGGaAGAcaAAGTATCCTCCACCCTCTCTGGTCTTGAAGCTGAACT
Erebia_oemeEW24_7         gGaATGGAaGACAAaGTCTCCTCGACTCTCTCTGGCCTCGAAGGCGAGCT

[COI 71 bp]
Papilio_glaucus_69_3      taAagAtaTTgGaACATTATACTTTATTTTTGGAATTTGAGCAAGAATATTAGGAACTTCTTTAAGTTTAT
Hamearis84_13             ?????????????????????????????????????TGAGCAGGAATAGTAGGAACATCATTAAGATTAC
Libythea_celtis71_1       ?????????????????????????????????????TGAGCAGGAATAGTAGGAACTTCATTAAGTCTAT
Danaus_plexippus108_21    ?????????????????????????????????????TGAGCAGGAATAGTTGGGACATCTTTAAGTCTTT
Greta_oto70_9             ?????????????????????????????????????TGAGCAGGAATAGTAGGAACATCTTTAAGTTTAT
Amathusia_phidippus114_17 ?????????????????????????????????????TGATCTGGAATAGTAGGAACATCCCTCAGTCTTA
Morpho_peleides66_5       ?????????????????????????????????????TGAGCCGGTATAATTGGTACATCCCTAAGTCTTA
BrintesiaB01              ?????????????????????????????????????TGAGCAGGTATAGTAGGAACATCTCTTAGTTTAA
Elymnias_casiphone121_20  ?????????????????????????????????????TGATCAGGAATAGTAGGAACTTCCCTCAGTCTTA
Erebia_oemeEW24_7         ?????????????????????????????????????TGAGCAGGTATAGTAGGTACTTCCCTTAGTCTTA
;
end;
```