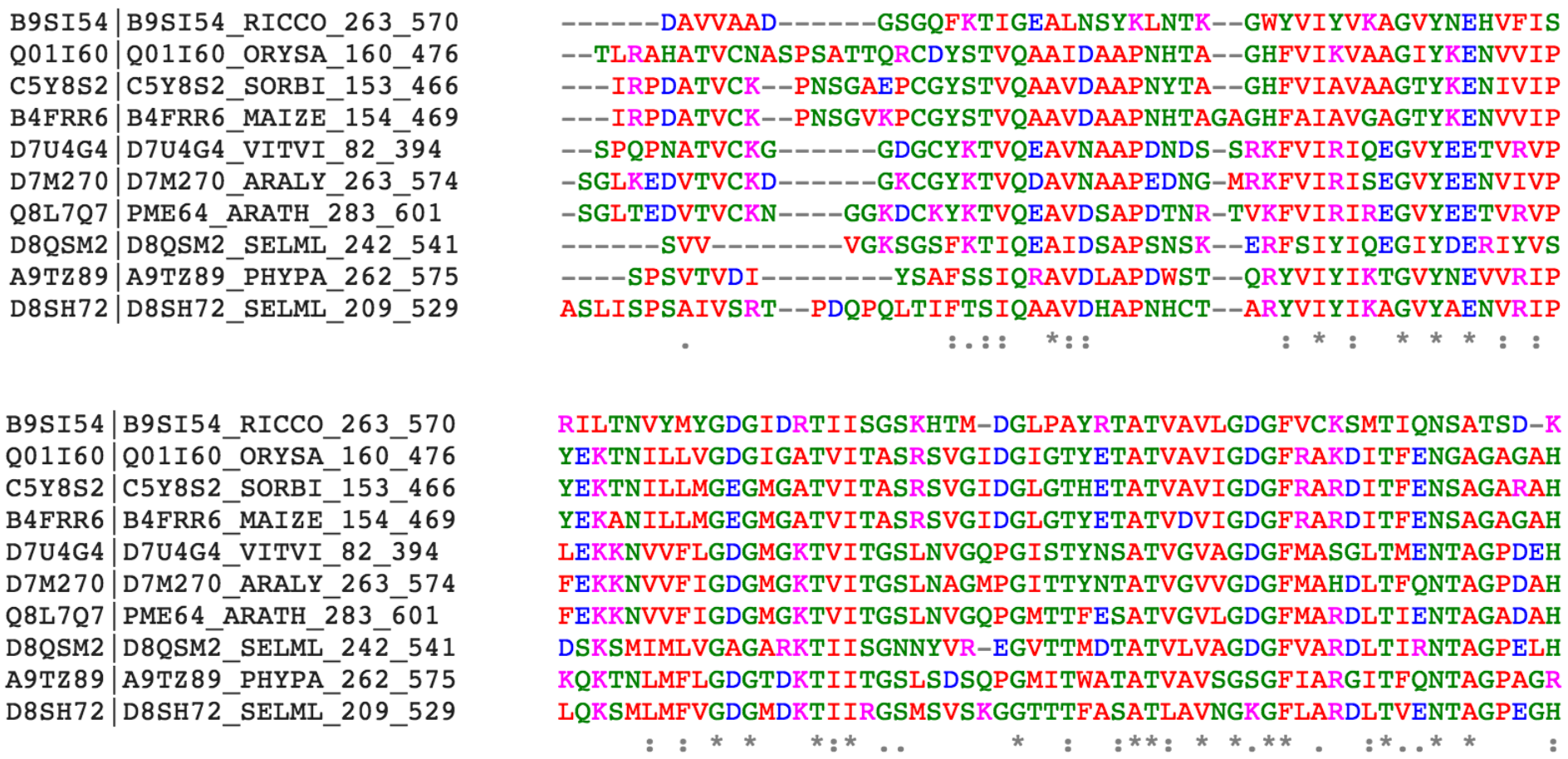
# Multiple sequence alignments and phylogenetic trees

# Multiple sequence alignment (MSA)



# Software to generate MSAs

- MAFFT
  (very good, very fast)
  http://mafft.cbrc.jp/alignment/software/
- Clustal Omega
  (very good, very fast)
  http://www.ebi.ac.uk/Tools/msa/clustalo/
- PRANK
  (extremely good, very slow)
  http://wasabiapp.org/software/prank/

# File formats: FASTA (holds any sequence data)


```
>human
MNGTEGPNFYVPFSNATGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVLGGFTSTLYTSLHGYFVFGPTGCNLEGFFATLG
YNPVIYIMMNKQFRNCMLTTICCGKNPLGDDEASATVSKTETSQVAPA
>domestic_cat
MNGTEGPNFYVPFSNKTGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
YNPVIYIMMNKQFRNCMLTTLCCGKNPLGDDEASTTASKTETSQVAPA
>chimpanzee
MNGTEGPNFYVPFSNATGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVLGGFTSTLYTSLHGYFVFGPTGCNLEGFFATLG
YNPVIYIMMNKQFRNCMLTTICCGKNPLGDDEASATVSKTETSQVAPA
```

# File formats: Clustal (holds an alignment)

```
CLUSTAL O(1.2.1) multiple sequence alignment

human           MNGTEGPNFYVPFSNATGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLY
chimpanzee      MNGTEGPNFYVPFSNATGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLY
domestic_cat    MNGTEGPNFYVPFSNKTGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLY
                *************** ************************************

human           VTVQHKKLRTPLNYILLNLAVADLFMVLGGFTSTLYTSLHGYFVFGPTGCNLEGFFATLG
chimpanzee      VTVQHKKLRTPLNYILLNLAVADLFMVLGGFTSTLYTSLHGYFVFGPTGCNLEGFFATLG
domestic_cat    VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
                ****************************:****:****************************

human           YNPVIYIMMNKQFRNCMLTTICCGKNPLGDDEASATVSKTETSQVAPA
chimpanzee      YNPVIYIMMNKQFRNCMLTTICCGKNPLGDDEASATVSKTETSQVAPA
domestic_cat    YNPVIYIMMNKQFRNCMLTTLCCGKNPLGDDEASTTASKTETSQVAPA
                ************************:***********:*.**********
```

labels

sequences

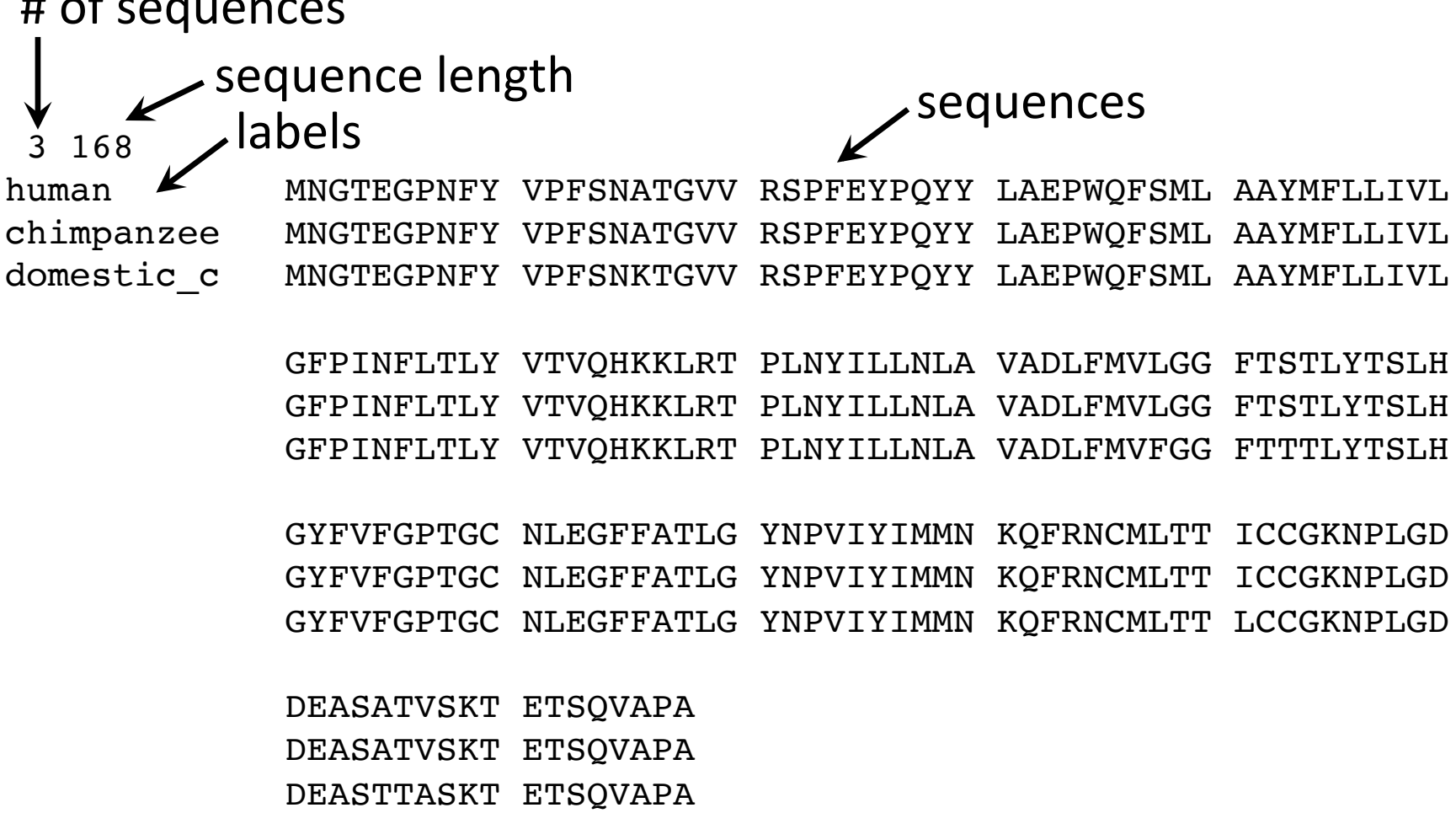
consensus indicators:

* = no variation

: = highly similar amino acids

. = somewhat similar amino acids

# File formats: Phylip (holds an alignment)

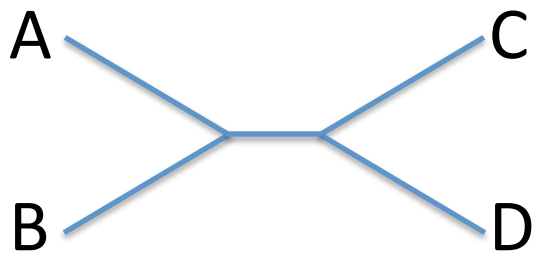


# Tools exist to convert from one sequence format to another

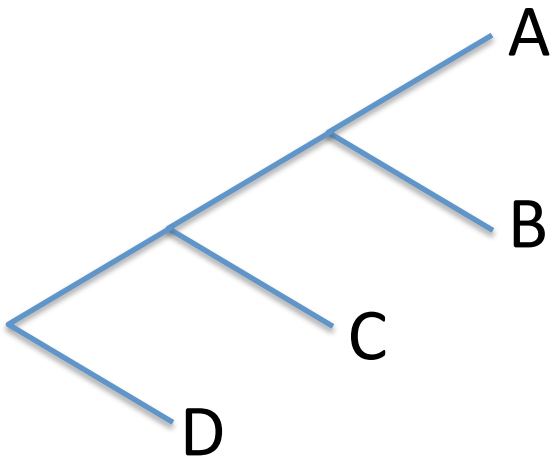
- Online:
  http://www.ebi.ac.uk/Tools/sfc/readseq/
- In a script:
  Use biopython SeqIO

# Storing trees: The Newick format

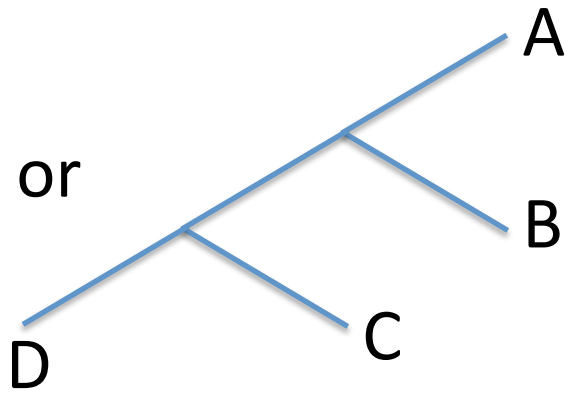
((A,B),(C,D))




(((A,B),C),D)




or

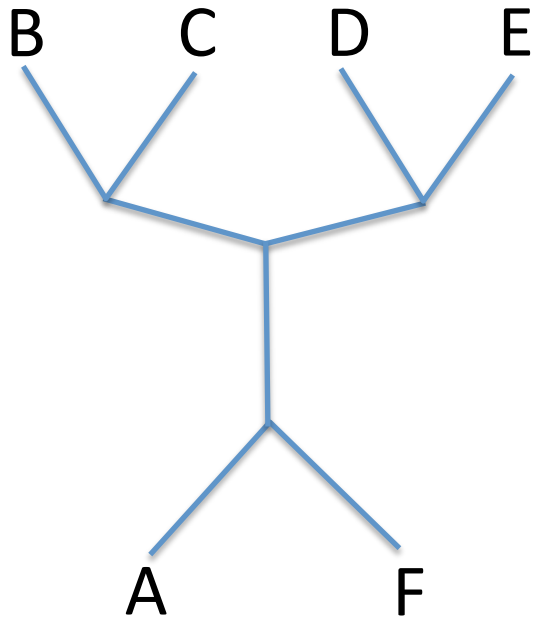


# What does this tree look like?

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

# What does this tree look like?

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



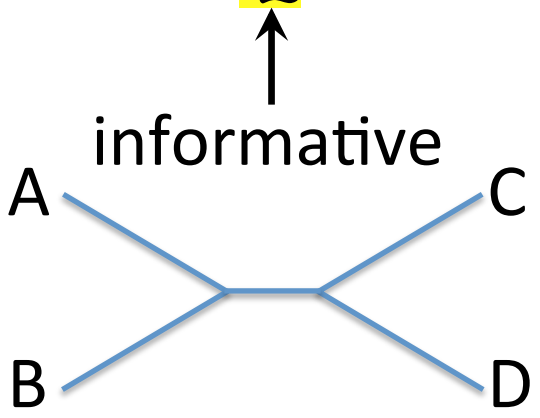
# Not all sites in an alignment contain information about the tree topology

```
A     MNGTEG
B     MNGYER
C     MQGYDK
D     MQGTDI
```

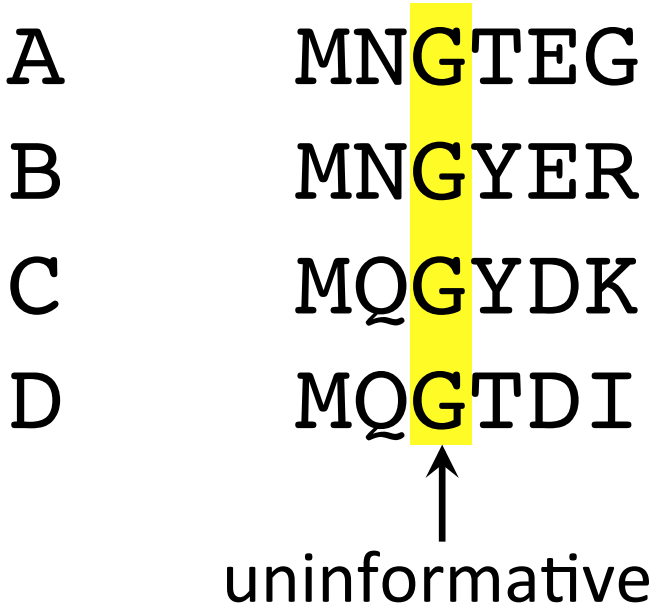
uninformative

# Not all sites in an alignment contain information about the tree topology

```
A     MNGTEG
B     MNGYER
C     MQGYDK
D     MQGTDI
```

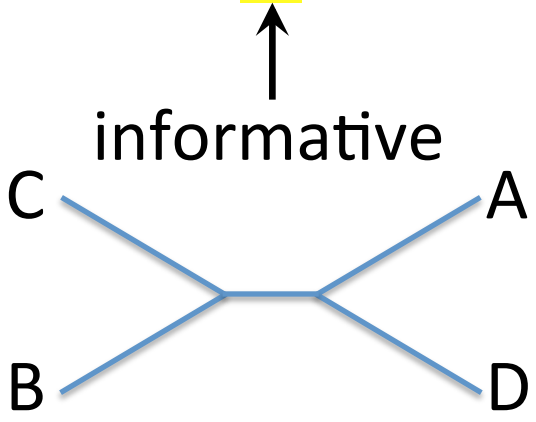


# Not all sites in an alignment contain information about the tree topology



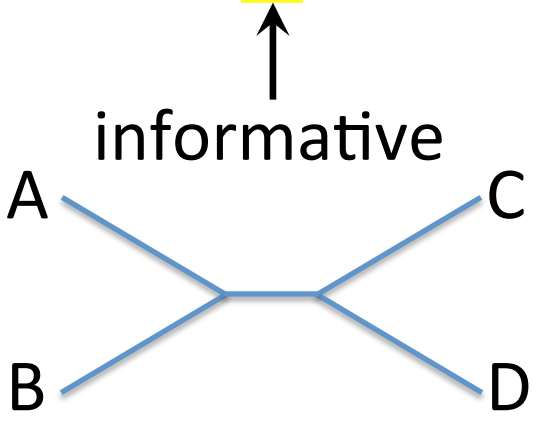
# Not all sites in an alignment contain information about the tree topology

```
A    MNGTEG
B    MNGYER
C    MQGYDK
D    MQGTDI
```



# Not all sites in an alignment contain information about the tree topology

```
A    MNGTEG
B    MNGYER
C    MQGYDK
D    MQGTDI
```



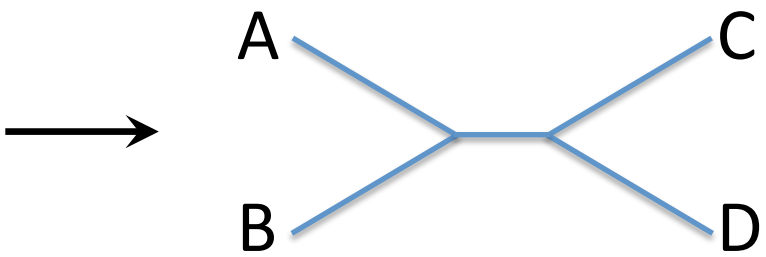
# Not all sites in an alignment contain information about the tree topology

```
A     MNGTEG
B     MNGYER
C     MQGYDK
D     MQGTDI
```

uninformative (in simplest model)

# Not all sites in an alignment contain information about the tree topology

```
A    MNGTEG
B    MNGYER
C    MQGYDK
D    MQGTDI
```



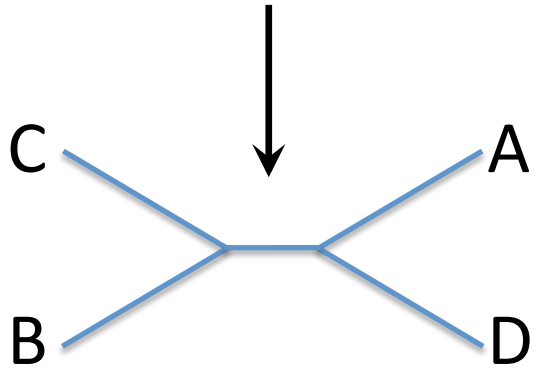

by majority rule

How confident are we in a given tree topology?

# Bootstrap: a method to assess confidence in tree toplogy

Randomly re-sample columns from the alignment, count frequency of topologies

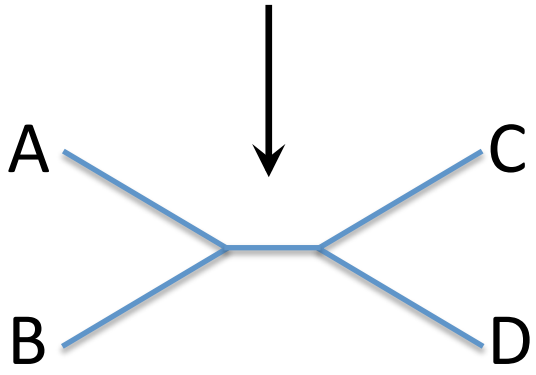
| | | | | |
|---|---|---|---|---|
| A | MNGTEG | → | A | GMGTMG |
| B | MNGYER | | B | GMRYMR |
| C | MQGYDK | | C | GMKYMK |
| D | MQGTDI | | D | GMITMI |



# Bootstrap: a method to assess confidence in tree toplogy

Randomly re-sample columns from the alignment, count frequency of topologies

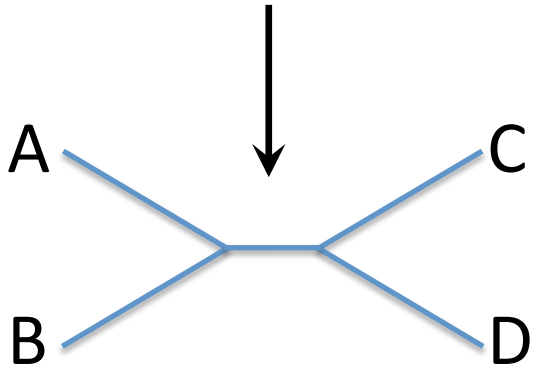
| | | | | |
|---|---|---|---|---|
| A | MNGTEG | → | A | NMNTMG |
| B | MNGYER | | B | NMNYMG |
| C | MQGYDK | | C | QMQYMG |
| D | MQGTDI | | D | QMQTMG |



# Bootstrap: a method to assess confidence in tree toplogy

Randomly re-sample columns from the alignment, count frequency of topologies

| | | | | |
|---|---|---|---|---|
| A | MNGTEG | → | A | MTNGEG |
| B | MNGYER | | B | MYNREG |
| C | MQGYDK | | C | MYQKDG |
| D | MQGTDI | | D | MTQIDG |

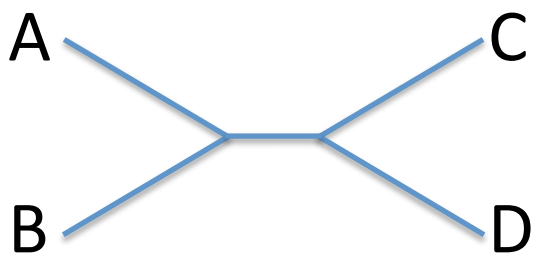


# Bootstrap: a method to assess confidence in tree toplogy

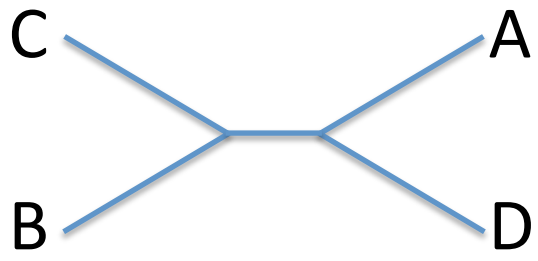
Randomly re-sample columns from the alignment, count frequency of topologies

Bootstrapped trees (100 x):

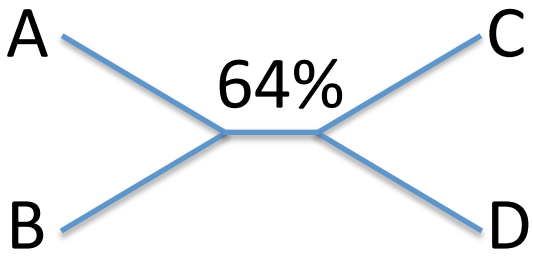
64 x




36 x




Final result:



# Tree-building methods:
# 1. Neighbor-joining

- Calculate all pair-wise distances
- Join two closest taxa, replace by new node
- Repeat

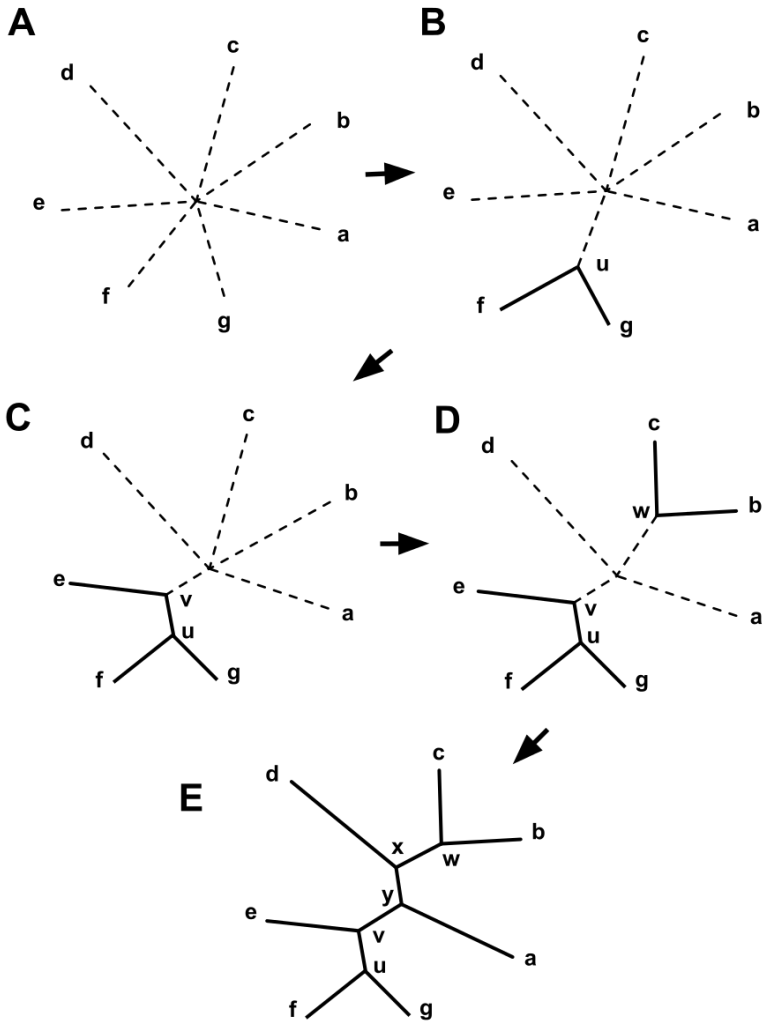
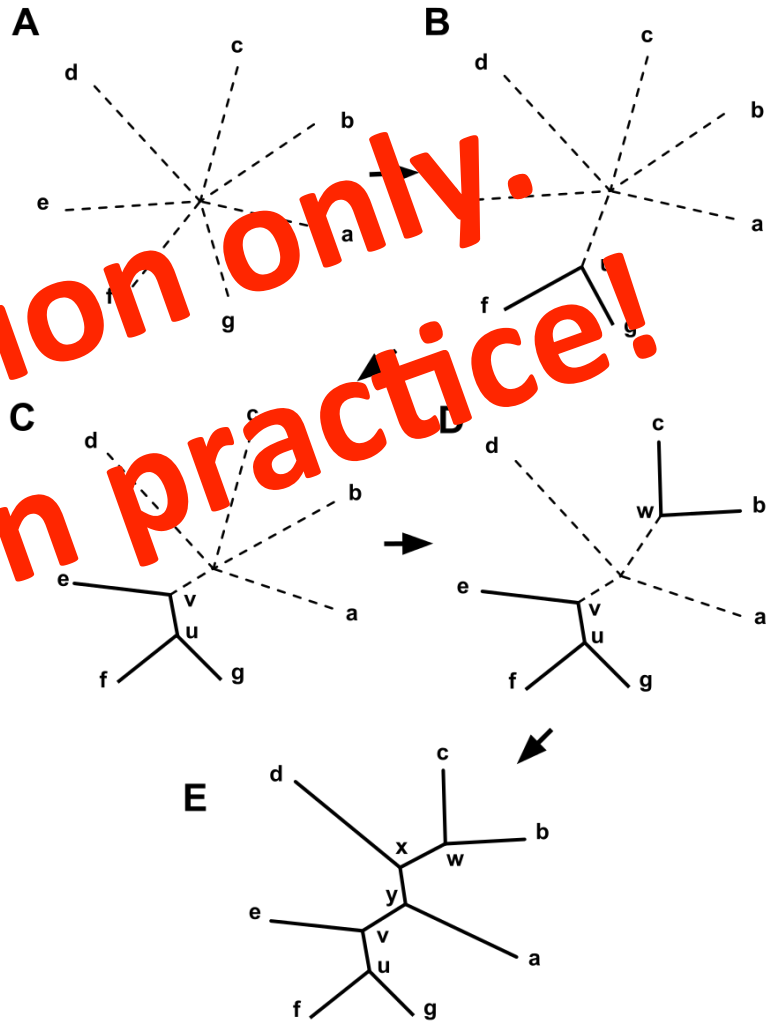



Image: http://en.wikipedia.org/wiki/File:Neighbor_joining_7_taxa_start_to_finish_diagram.svg

# Tree-building methods:
# 1. Neighbor-joining

- Calculate all pair-wise distances
- Join two closest taxa, replace by new node
- Repeat




**Demonstration only.**
**Do not use in practice!**

Image: http://en.wikipedia.org/wiki/File:Neighbor_joining_7_taxa_start_to_finish_diagram.svg

# Tree-building methods:
# 2. Maximum likelihood

- Builds likelihood model of molecular evolution
- Finds tree that maximizes:
  Pr(sequence data | tree)
- Commonly used software:
  RAxML, FastTree2

# Tree-building methods:
# 3. Bayesian

- Builds likelihood model of molecular evolution
- Calculates:
  Pr(tree | sequence data)
- Commonly used software:
  MrBayes, BEAST