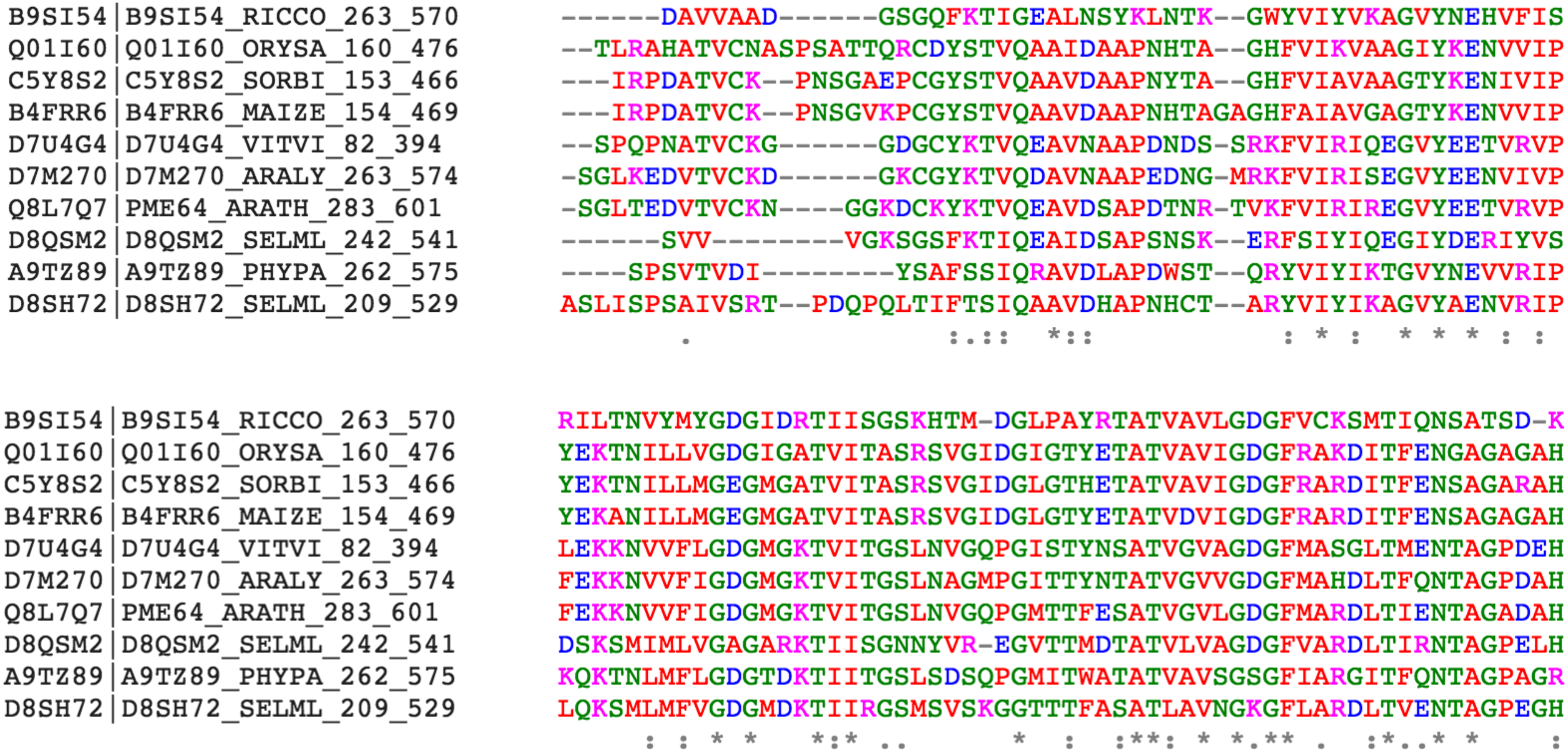
Multiple sequence alignments and phylogenetic trees

Multiple sequence alignment (MSA)



```
B9SI54|B9SI54_RICCO_263_570      ------DAVVAAD------GSGQFKTIGEALNSYKLNTK--GWYVIYVKAGVYNEHVFIS
Q01I60|Q01I60_ORYSA_160_476      --TLRAHATVCNASPSATTQRCDYSTVQAAIDAAPNHTA--GHFVIKVAAGIYKENVVIP
C5Y8S2|C5Y8S2_SORBI_153_466      ---IRPDATVCK--PNSGAEPCGYSTVQAAVDAAPNYTA--GHFVIAVAAGTYKENIVIP
B4FRR6|B4FRR6_MAIZE_154_469      ---IRPDATVCK--PNSGVKPCGYSTVQAAVDAAPNHTAGAGHFAIAVGAGTYKENVVIP
D7U4G4|D7U4G4_VITVI_82_394       --SPQPNATVCKG------GDGCYKTVQEAVNAAPDNDS-SRKFVIRIQEGVYEETVRVP
D7M270|D7M270_ARALY_263_574      -SGLKEDVTVCKD------GKCGYKTVQDAVNAAPEDNG-MRKFVIRISEGVYEENVIVP
Q8L7Q7|PME64_ARATH_283_601       -SGLTEDVTVCKN----GGKDCKYKTVQEAVDSAPDTNR-TVKFVIRIREGVYEETVRVP
D8QSM2|D8QSM2_SELML_242_541      ------SVV--------VGKSGSFKTIQEAIDSAPSNSK--ERFSIYIQEGIYDERIYVS
A9TZ89|A9TZ89_PHYPA_262_575      ----SPSVTVDI--------YSAFSSIQRAVDLAPDWST--QRYVIYIKTGVYNEVVRIP
D8SH72|D8SH72_SELML_209_529      ASLISPSAIVSRT--PDQPQLTIFTSIQAAVDHAPNHCT--ARYVIYIKAGVYAENVRIP
                                        .               :.::  *::           : * :  * * * : :

B9SI54|B9SI54_RICCO_263_570      RILTNVYMYGDGIDRTIISGSKHTM-DGLPAYRTATVAVLGDGFVCKSMTIQNSATSD-K
Q01I60|Q01I60_ORYSA_160_476      YEKTNILLVGDGIGATVITASRSVGIDGIGTYETATVAVIGDGFRAKDITFENGAGAGAH
C5Y8S2|C5Y8S2_SORBI_153_466      YEKTNILLMGEGMGATVITASRSVGIDGLGTHETATVAVIGDGFRARDITFENSAGARAH
B4FRR6|B4FRR6_MAIZE_154_469      YEKANILLMGEGMGATVITASRSVGIDGLGTYETATVDVIGDGFRARDITFENSAGAGAH
D7U4G4|D7U4G4_VITVI_82_394       LEKKNVVFLGDGMGKTVITGSLNVGQPGISTYNSATVGVAGDGFMASGLTMENTAGPDEH
D7M270|D7M270_ARALY_263_574      FEKKNVVFIGDGMGKTVITGSLNAGMPGITTYNTATVGVVGDGFMAHDLTFQNTAGPDAH
Q8L7Q7|PME64_ARATH_283_601       FEKKNVVFIGDGMGKTVITGSLNVGQPGMTTFESATVGVLGDGFMARDLTIENTAGADAH
D8QSM2|D8QSM2_SELML_242_541      DSKSMIMLVGAGARKTIISGNNYVR-EGVTTMDTATVLVAGDGFVARDLTIRNTAGPELH
A9TZ89|A9TZ89_PHYPA_262_575      KQKTNLMFLGDGTDKTIITGSLSDSQPGMITWATATVAVSGSGFIARGITFQNTAGPAGR
D8SH72|D8SH72_SELML_209_529      LQKSMLMFVGDGMDKTIIRGSMSVSKGGTTTFASATLAVNGKGFLARDLTVENTAGPEGH
                                      : : * *   *:* ..      *  :  :**: * *.** .  :*..* *    :
```

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)

label (1 line)

sequence (multiple lines)

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

File formats: Clustal (holds an alignment)

labels

sequences

```
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
                  ********************:*************:*.***********
```

consensus indicators:

* = no variation

: = highly similar amino acids

. = somewhat similar amino acids

File formats: Phylip (holds an alignment)

# of sequences

sequence length

labels

sequences

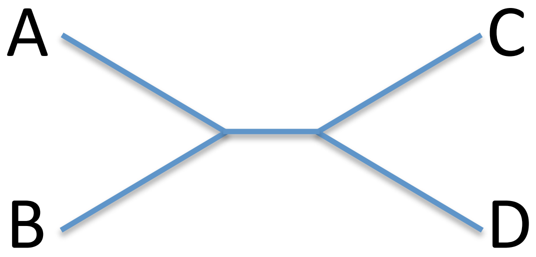
```
 3 168
human        MNGTEGPNFY VPFSNATGVV RSPFEYPQYY LAEPWQFSML AAYMFLLIVL
chimpanzee   MNGTEGPNFY VPFSNATGVV RSPFEYPQYY LAEPWQFSML AAYMFLLIVL
domestic_c   MNGTEGPNFY VPFSNKTGVV RSPFEYPQYY LAEPWQFSML AAYMFLLIVL

             GFPINFLTLY VTVQHKKLRT PLNYILLNLA VADLFMVLGG FTSTLYTSLH
             GFPINFLTLY VTVQHKKLRT PLNYILLNLA VADLFMVLGG FTSTLYTSLH
             GFPINFLTLY VTVQHKKLRT PLNYILLNLA VADLFMVFGG FTTTLYTSLH

             GYFVFGPTGC NLEGFFATLG YNPVIYIMMN KQFRNCMLTT ICCGKNPLGD
             GYFVFGPTGC NLEGFFATLG YNPVIYIMMN KQFRNCMLTT ICCGKNPLGD
             GYFVFGPTGC NLEGFFATLG YNPVIYIMMN KQFRNCMLTT LCCGKNPLGD

             DEASATVSKT ETSQVAPA
             DEASATVSKT ETSQVAPA
             DEASTTASKT ETSQVAPA
```

Tools exist to convert from one sequence format to another

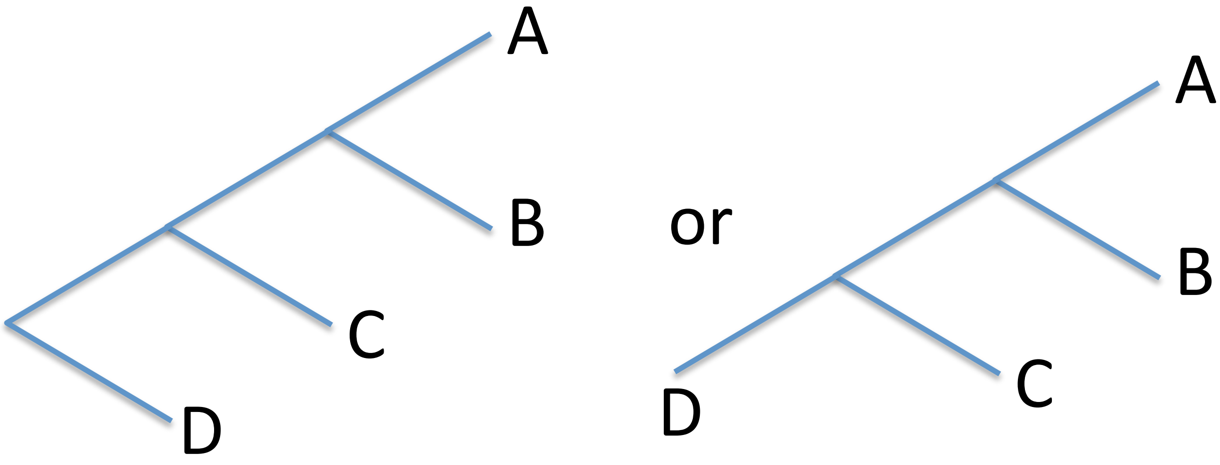
- Online:
  http://www.ebi.ac.uk/Tools/sfc/readseq/
- In a script:
  Use biopython SeqIO

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)

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

A	MNGTEG
B	MNGYER
C	MQGYDK
D	MQGTDI

↑

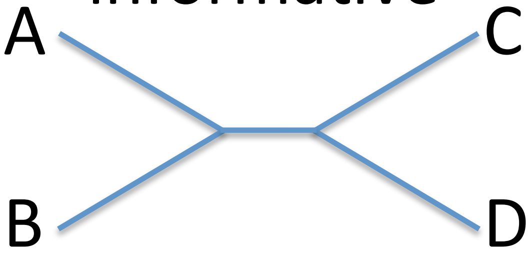
uninformative

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

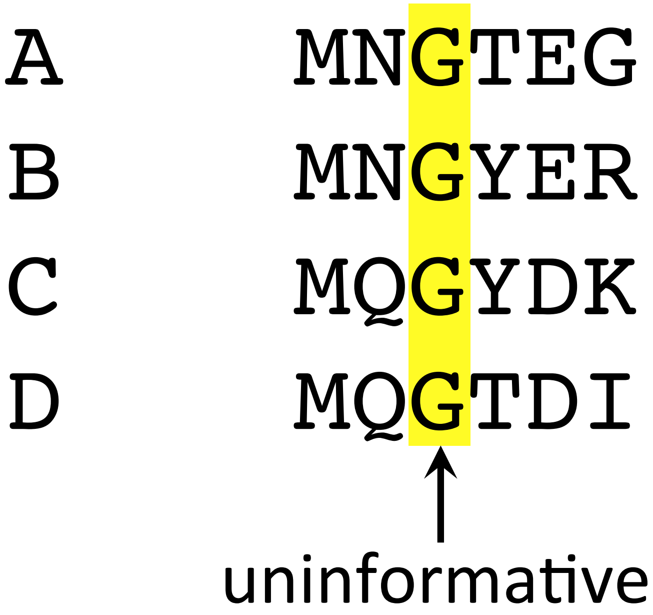
A	MNGTEG
B	MNGYER
C	MQGYDK
D	MQGTDI

↑

informative



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



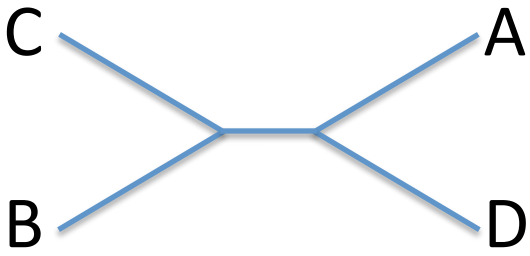
A	MNGTEG
B	MNGYER
C	MQGYDK
D	MQGTDI

uninformative

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

A	MNGTEG
B	MNGYER
C	MQGYDK
D	MQGTDI

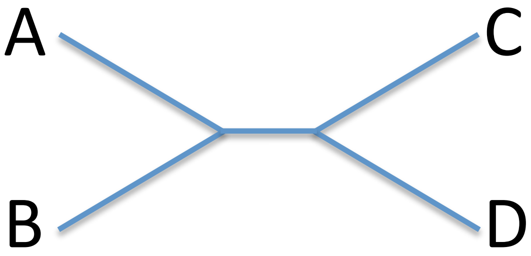
informative



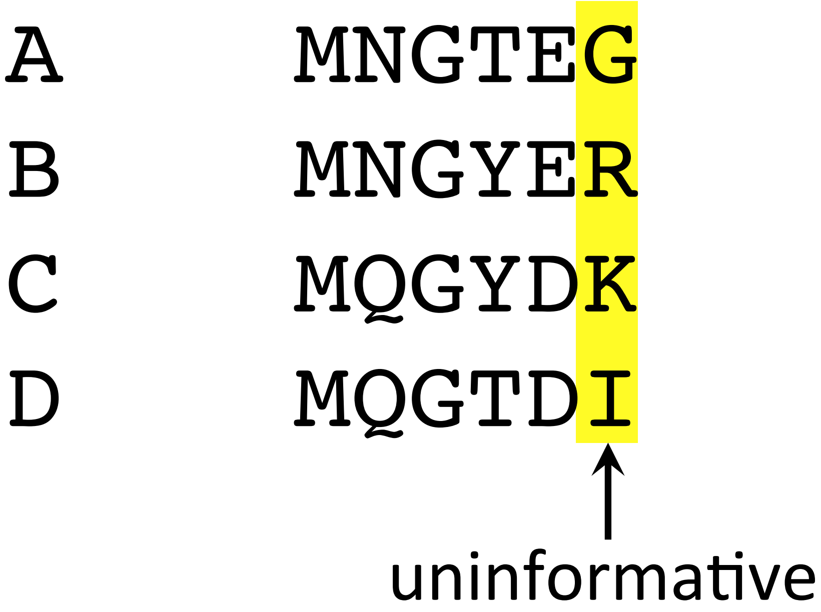
Not all sites in an alignment contain information about the tree topology

A	MNGTEG
B	MNGYER
C	MQGYDK
D	MQGTDI

informative



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



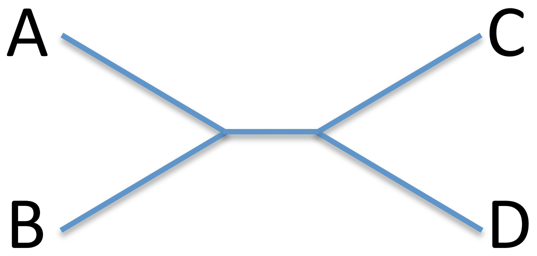
A	MNGTEG
B	MNGYER
C	MQGYDK
D	MQGTDI

uninformative

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

A	MNGTEG
B	MNGYER
C	MQGYDK
D	MQGTDI

→



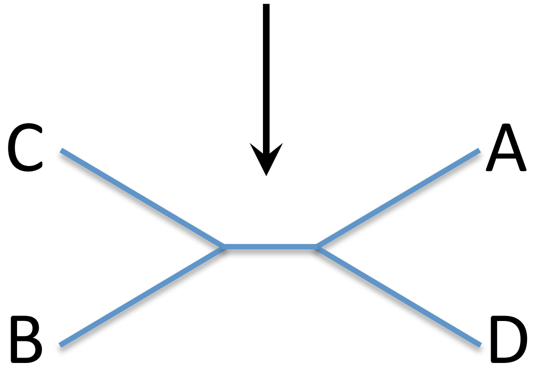
by majority rule

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

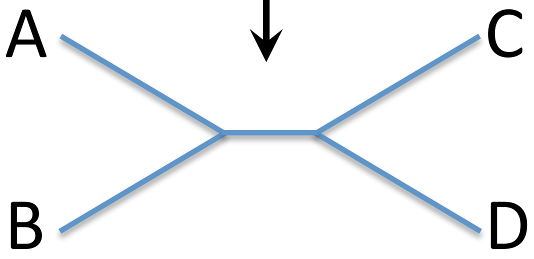


Bootstrap: a method to assess confidence in tree toplogy

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
```

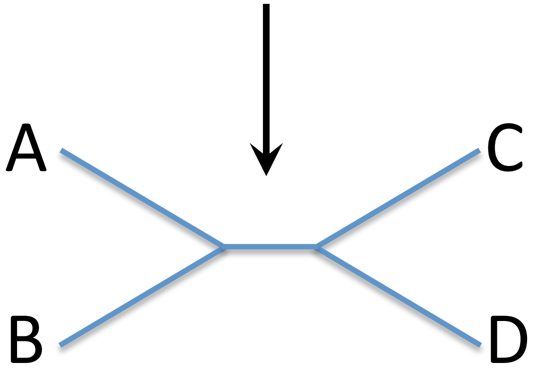
↓



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
```

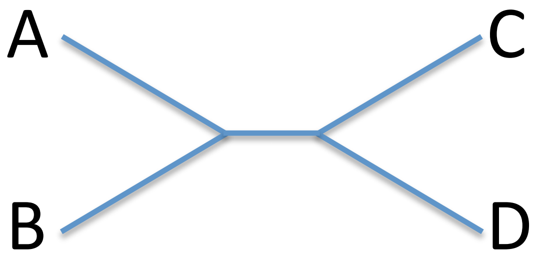


Bootstrap: a method to assess confidence in tree toplogy

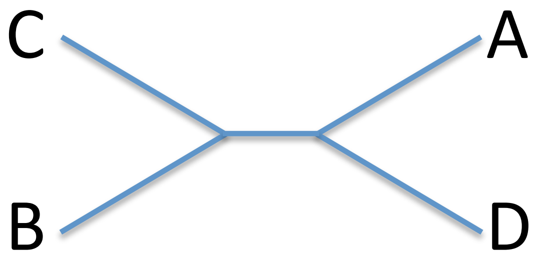
Randomly re-sample columns from the alignment, count frequency of topologies

Bootstrapped trees (100 x):

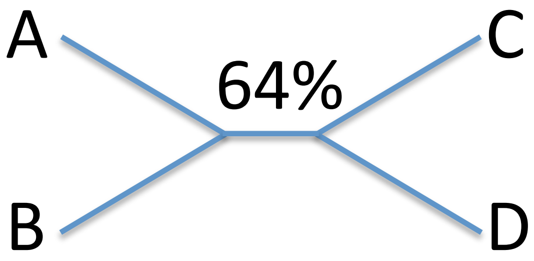
64 x



36 x



Final result:



Tree-building methods:
1. Neighbor-joining

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

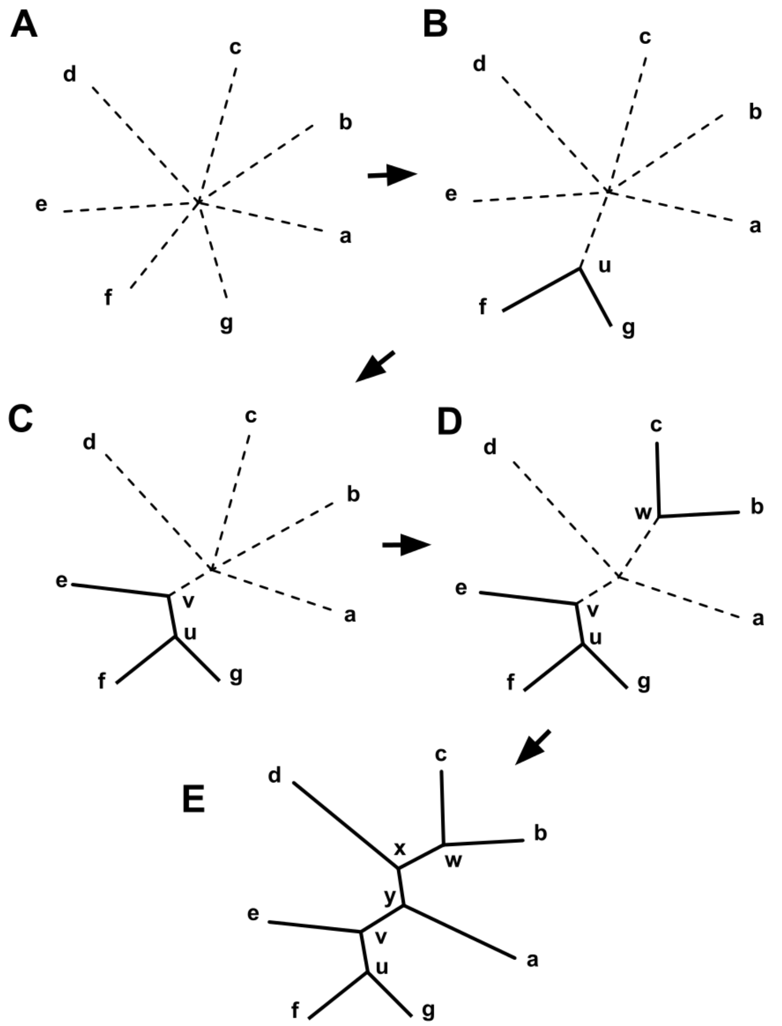
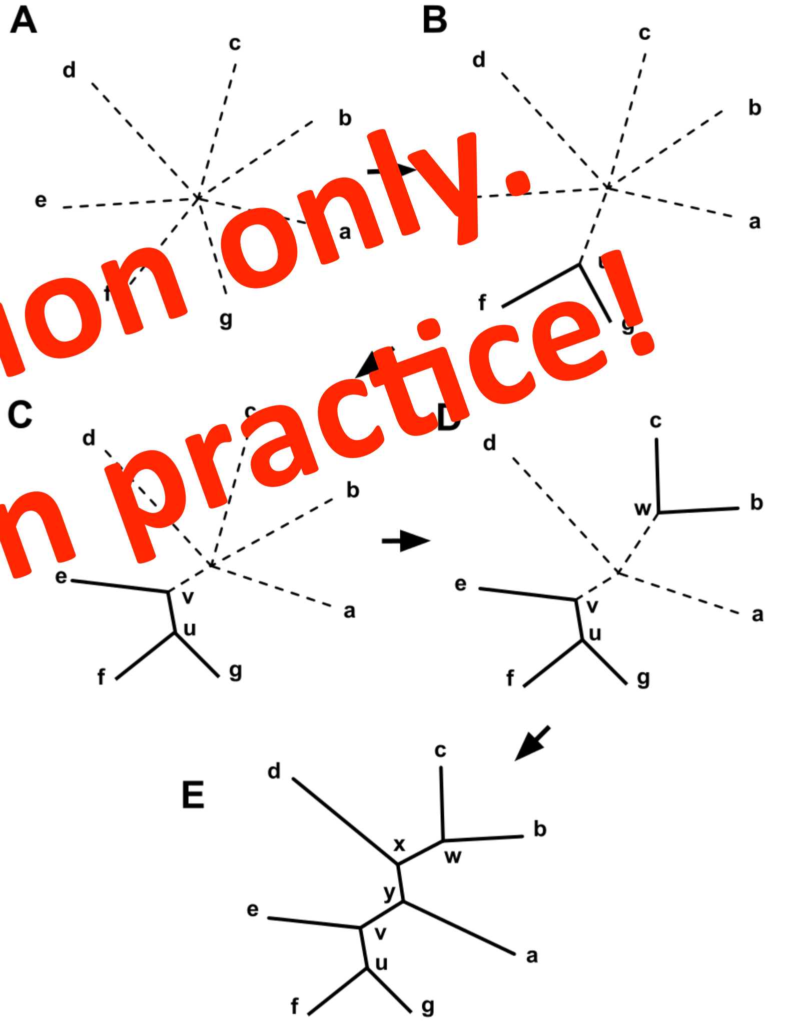


Image: http://en.wikipedia.org/wiki/File:Neighbor_joining_7_taxa_start_to_finish_diagram.svg

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