

# Multiple sequence alignments and phylogenetic trees

# Multiple sequence alignment (MSA)

B9SI54	B9SI54_RICCO_263_570
Q01I60	Q01I60_ORYSA_160_476
C5Y8S2	C5Y8S2_SORBI_153_466
B4FRR6	B4FRR6_MAIZE_154_469
D7U4G4	D7U4G4_VITVI_82_394
D7M270	D7M270_ARALY_263_574
Q8L7Q7	PME64_ARATH_283_601
D8QSM2	D8QSM2_SELML_242_541
A9TZ89	A9TZ89_PHYPA_262_575
D8SH72	D8SH72_SELML_209_529

```
-----DAVVAAD-----GSGQFKTIGEALNSYKLNTK--GWYVIYVKAGVYNEHVFIS
--TLRAHATVCNAPSPATTQRCDYSTVQAAIDAAPNHTA--GHFVIKVAAGIYKENVVIP
---IRPDATVCK--PNSGAEP CGYSTVQAAVDAAPNYTA--GHFVIAVAAGTYKENIVIP
---IRPDATVCK--PNSGVKPCGYSTVQAAVDAAPNHTAGAGHFAIAVGAGTYKENVVIP
--SPQPNATVCKG-----GDGCYKTVQEA VNAAPDNDS--SRKFVIRIQEGVYEETVRVP
-SGLKEDVTVC KD-----GKCGYKTVQDAVNAAPEDNG--MRKFVIRISEGVYEENVVIP
-SGLTEDVTVC KN----GGKDCYKTVQEA VDSAPDTNR--TVKFVIRIREGVYEETVRVP
-----SVV-----VGKSGSF KTIQE AID SAPSNSK--ERFSIYIQEGIYDERIYVS
---SPSVTVDI-----YSAFSSIQRAVDLAPDWST--QRYVIYIKTGVYNEVVRIP
ASLISPSAIVSRT--PDQPQLTIFTSIQAAVDHAPNHCT--ARYVIYIKAGVYAENVVRIP
      .                : . : :      * : :                : * :      * * * : :
```

B9SI54	B9SI54_RICCO_263_570
Q01I60	Q01I60_ORYSA_160_476
C5Y8S2	C5Y8S2_SORBI_153_466
B4FRR6	B4FRR6_MAIZE_154_469
D7U4G4	D7U4G4_VITVI_82_394
D7M270	D7M270_ARALY_263_574
Q8L7Q7	PME64_ARATH_283_601
D8QSM2	D8QSM2_SELML_242_541
A9TZ89	A9TZ89_PHYPA_262_575
D8SH72	D8SH72_SELML_209_529

```
RILT NVMYGDGIDRTIISGSKHTM-DGLPAYRTATVAVLGDGFVCKSMTIQNSATSD-K
YEKTNILLVGDGIGATVITASRSVGIDGIGTYETATVAVIGDGFRAKDITFENGAGAGAH
YEKTNILLMGE MGATVITASRSVGIDGLGTHETATVAVIGDGFRAKDITFENSAGARAH
YEKANILLMGE MGATVITASRSVGIDGLGTYETATVDVIGDGFRAKDITFENSAGAGAH
LEKKNVVF LIGDGMGKT VITGSLNVGQPGISTYNSATVGVAGDGFMA SGLTMENTAGPDEH
FEKKNVVF LIGDGMGKT VITGSLNAGMPGITTYNTATVGVVGDGFMA HDLTFQNTAGPDAH
FEKKNVVF LIGDGMGKT VITGSLNVGQPGMTTFESATVGVLDGDFMARDLTIENTAGADAH
DSKSMIMLVGAGARKTIIISGNNYVR-EGVTTMDTATVLVAGDGFVARDLTIRNTAGPELH
KQKTNLMFLGDGTDKTIITGSLSDSQPGMITWATATVAVSGSGFIARGITFQNTAGPAGR
LQKSMLMFVGDGMDKTIIRGSM SVSKGGTTTFASATLAVNGKGF LARDLTVENTAGPEGH
      : : * *      * : * . .      * :      : * : * * . .      : * . . * *      :
```

# 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
MNGTEGPNFYVPFSNATGVVRSPFEYPQYYLAEPWQFSMLAAYMFL LIVLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMV LGGFTSTLYTSLHGYFVFGPTGCNLEGFFATLG
YNPVIYIMMNKQFRNCMLTTICCGKNPLGDDEASATVSKTETSQVAPA
>domestic_cat
MNGTEGPNFYVPFSNKTGVVRSPFEYPQYYLAEPWQFSMLAAYMFL LIVLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMV FGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
YNPVIYIMMNKQFRNCMLTTLCCGKNPLGDDEASTTASKTETSQVAPA
>chimpanzee
MNGTEGPNFYVPFSNATGVVRSPFEYPQYYLAEPWQFSMLAAYMFL LIVLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMV LGGFTSTLYTSLHGYFVFGPTGCNLEGFFATLG
YNPVIYIMMNKQFRNCMLTTICCGKNPLGDDEASATVSKTETSQVAPA
```

# File formats: Clustal (holds an alignment)

CLUSTAL O(1.2.1) multiple sequence alignment

human  
chimpanzee  
domestic\_cat

sequences

MNGTEGPNFYVPFSNATGVVRSPFEYPQYYLAEPWQFSMLAAYMFLIIVLGFPINFLTLY  
MNGTEGPNFYVPFSNATGVVRSPFEYPQYYLAEPWQFSMLAAYMFLIIVLGFPINFLTLY  
MNGTEGPNFYVPFSNKTGVVRSPFEYPQYYLAEPWQFSMLAAYMFLIIVLGFPINFLTLY  
\*\*\*\*\*

human  
chimpanzee  
domestic\_cat

VTVQHKKLRTPLNYILLNLAVADLFMVLGGFTSTLYTSLHGYFVFGPTGCNLEGFFATLG  
VTVQHKKLRTPLNYILLNLAVADLFMVLGGFTSTLYTSLHGYFVFGPTGCNLEGFFATLG  
VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG  
\*\*\*\*\*:\*\*\*\*\*

human  
chimpanzee  
domestic\_cat

YNPVIYIMMNKQFRNCMLTTICCGKNPLGDDEASATVSKTETSQVAPA  
YNPVIYIMMNKQFRNCMLTTICCGKNPLGDDEASATVSKTETSQVAPA  
YNPVIYIMMNKQFRNCMLTTICCGKNPLGDDEASTTASKTETSQVAPA  
\*\*\*\*\*:\*\*\*\*\*

consensus indicators:

\* = no variation

: = highly similar amino acids

. = somewhat similar amino acids

# File formats: Phylip (holds an alignment)

# of sequences  
↓  
3 168

sequence length  
↓  
168

labels  
↓  
human  
chimpanzee  
domestic\_c

sequences  
↓

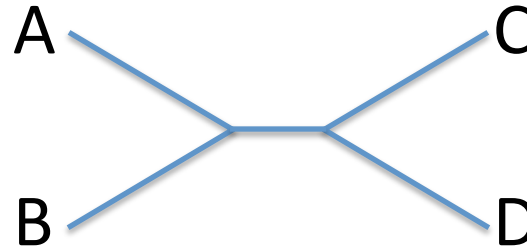
MNGTEGPNFY	VPFSNATGVV	RSPFEYPQYY	LAEPWQFSML	AAYMFLLIVL
MNGTEGPNFY	VPFSNATGVV	RSPFEYPQYY	LAEPWQFSML	AAYMFLLIVL
MNGTEGPNFY	VPFSNKTGVV	RSPFEYPQYY	LAEPWQFSML	AAYMFLLIVL
GFPINFLTLY	VTVQHKKLRT	PLNYILLNLA	VADLFMVLGG	FTSTLYTSLH
GFPINFLTLY	VTVQHKKLRT	PLNYILLNLA	VADLFMVLGG	FTSTLYTSLH
GFPINFLTLY	VTVQHKKLRT	PLNYILLNLA	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

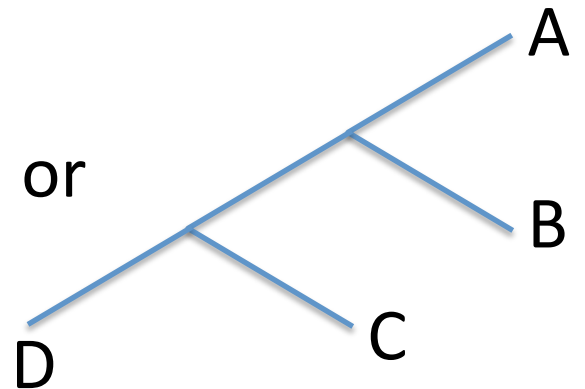
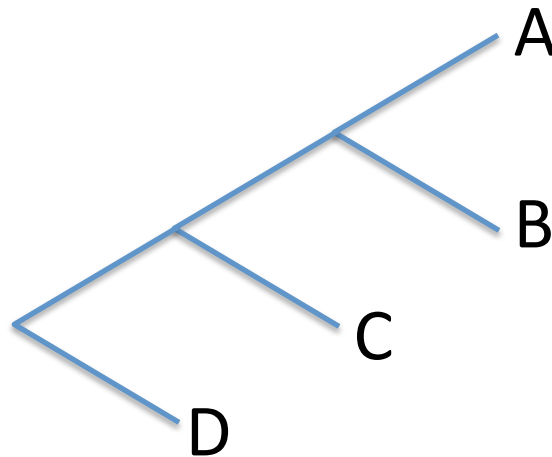
- Online:  
<http://www.ebi.ac.uk/Tools/sfc/readseq/>
- In a script:  
Use biopython SeqIO

# Storing trees: The Newick format

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



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





# What does this tree look like?

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

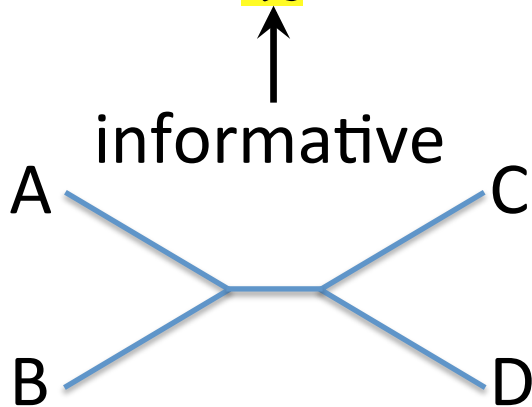
# Not all sites in an alignment contain information about the tree topology

A	M	N	G	T	E	G
B	M	N	G	Y	E	R
C	M	Q	G	Y	D	K
D	M	Q	G	T	D	I

↑  
uninformative

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

A	M	N	G	T	E	G
B	M	N	G	Y	E	R
C	M	Q	G	Y	D	K
D	M	Q	G	T	D	I



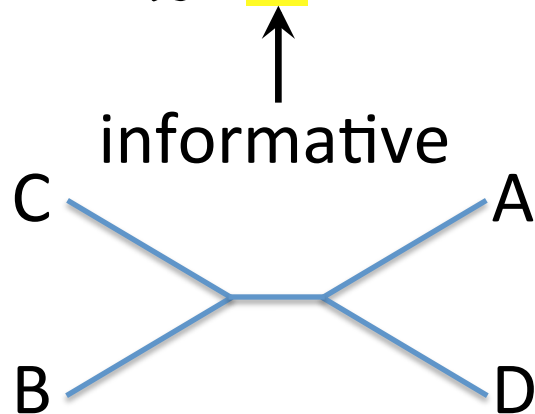
# Not all sites in an alignment contain information about the tree topology

A	M	N	G	T	E	G
B	M	N	G	Y	E	R
C	M	Q	G	Y	D	K
D	M	Q	G	T	D	I

↑  
uninformative

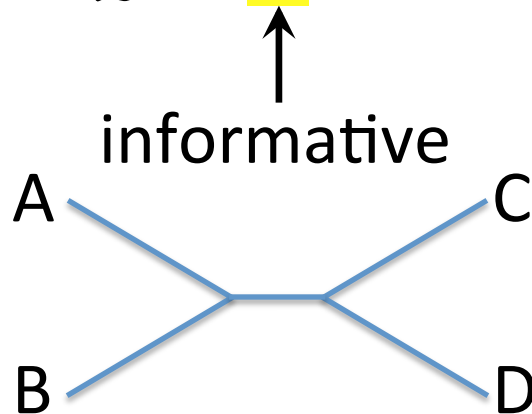
# Not all sites in an alignment contain information about the tree topology

A	MNGTEG
B	MNGYER
C	MQGYDK
D	MQGTDI



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

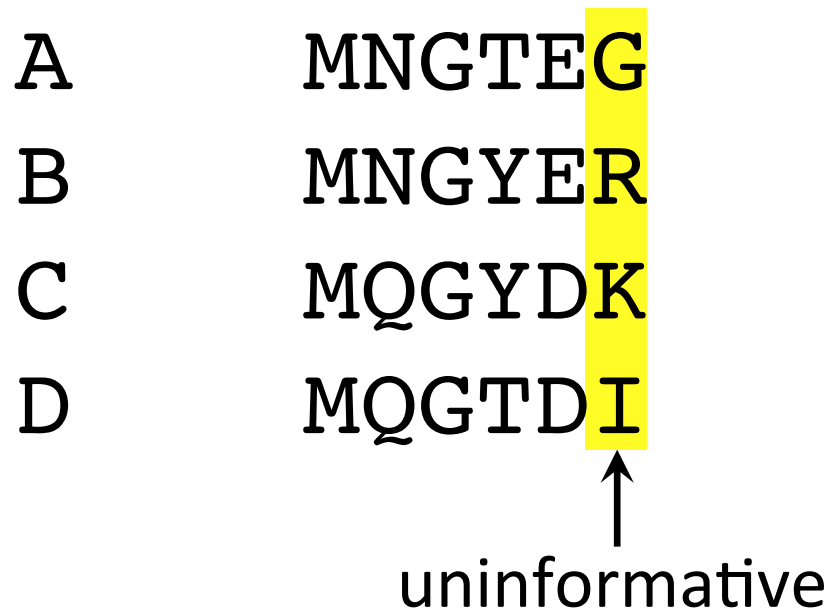
A	MNGTEG
B	MNGYER
C	MQGYDK
D	MQGTDI



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

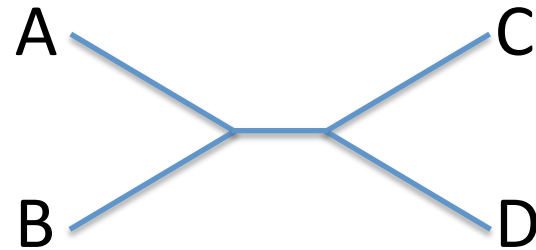
A	MNGTEG
B	MNGYER
C	MQGYDK
D	MQGTDI

↑  
uninformative

The diagram shows a sequence alignment with four rows labeled A, B, C, and D. The sequences are MNGTEG, MNGYER, MQGYDK, and MQGTDI. A vertical yellow highlight covers the sixth column, which contains the characters G, R, K, and I. An upward-pointing arrow is positioned below the 'I' in row D, pointing to the highlighted column. The word 'uninformative' is written below the arrow.

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

A	MNGTEG
B	MNGYER
C	MQGYDK
D	MQGTDI



by majority rule

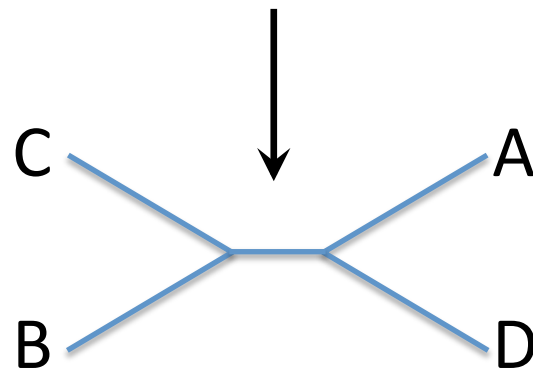
How confident are we in a given tree topology?



# Bootstrap: a method to assess confidence in tree topology

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



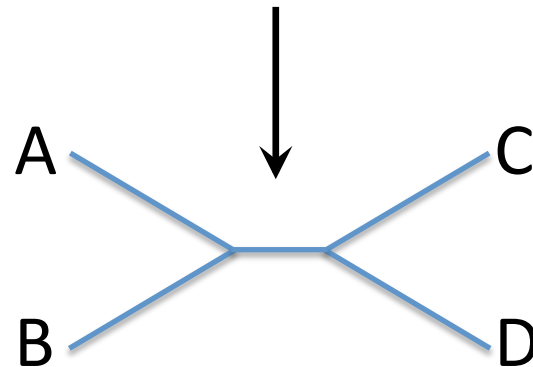
# Bootstrap: a method to assess confidence in tree topology

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

A	MNGTEG
B	MNGYER
C	MQGYDK
D	MQGTDI



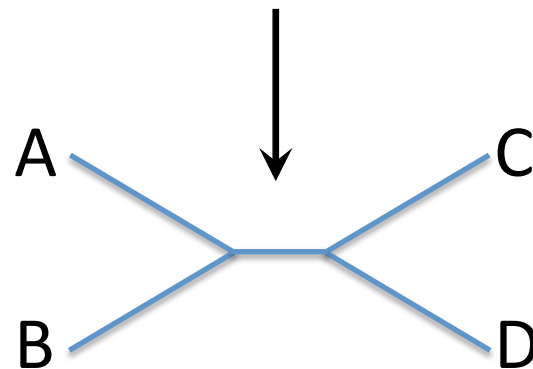
A	NMNTMG
B	NMNYMG
C	QMQYMG
D	QMQTMG



# Bootstrap: a method to assess confidence in tree topology

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

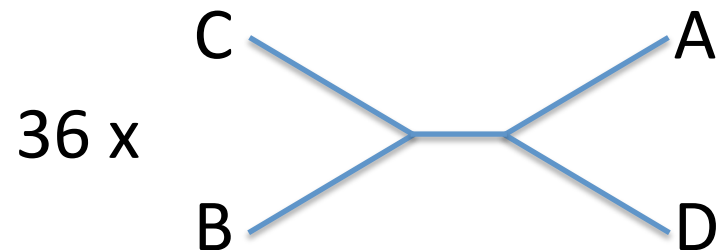
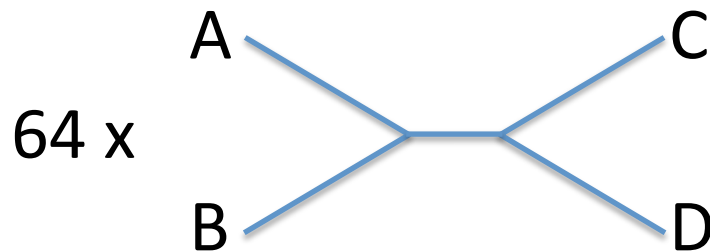
A	MNGTEG		A	MTNGEG
B	MNGYER	→	B	MYNREG
C	MQGYDK		C	MYQKDG
D	MQGTDI		D	MTQIDG



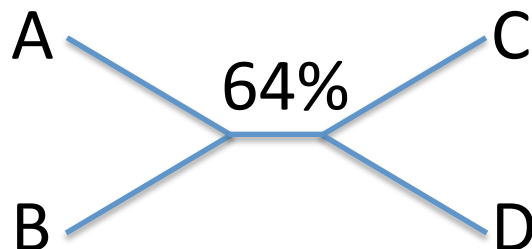
# Bootstrap: a method to assess confidence in tree topology

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

Bootstrapped trees (100 x):



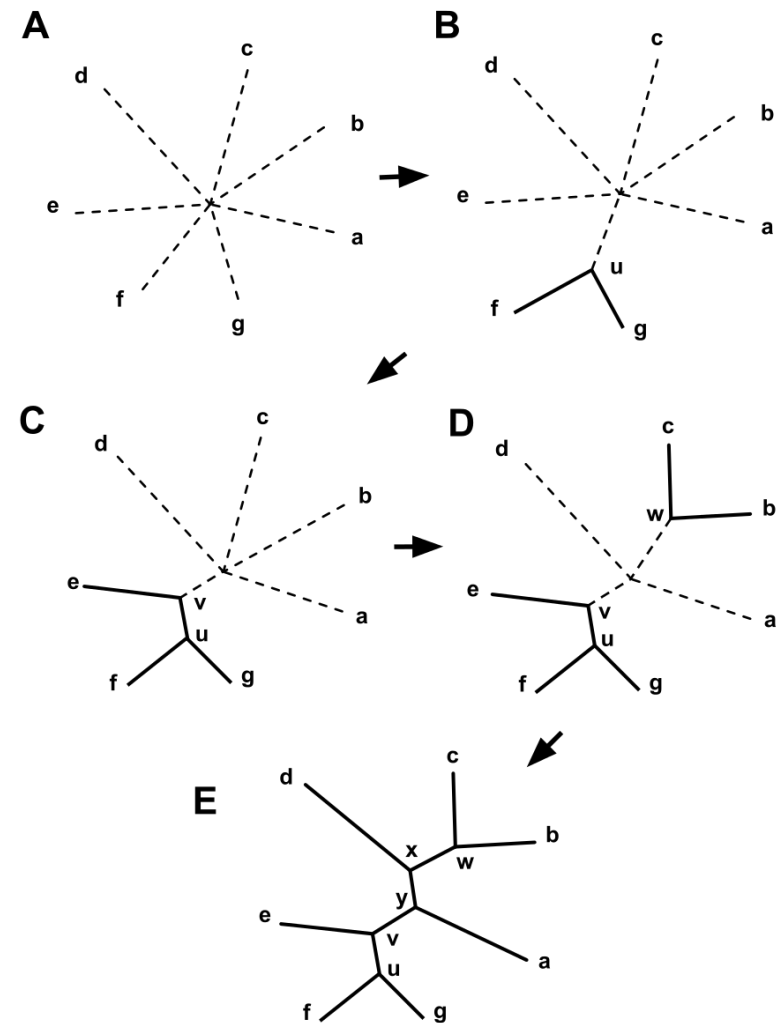
Final result:



# Tree-building methods:

## 1. Neighbor-joining

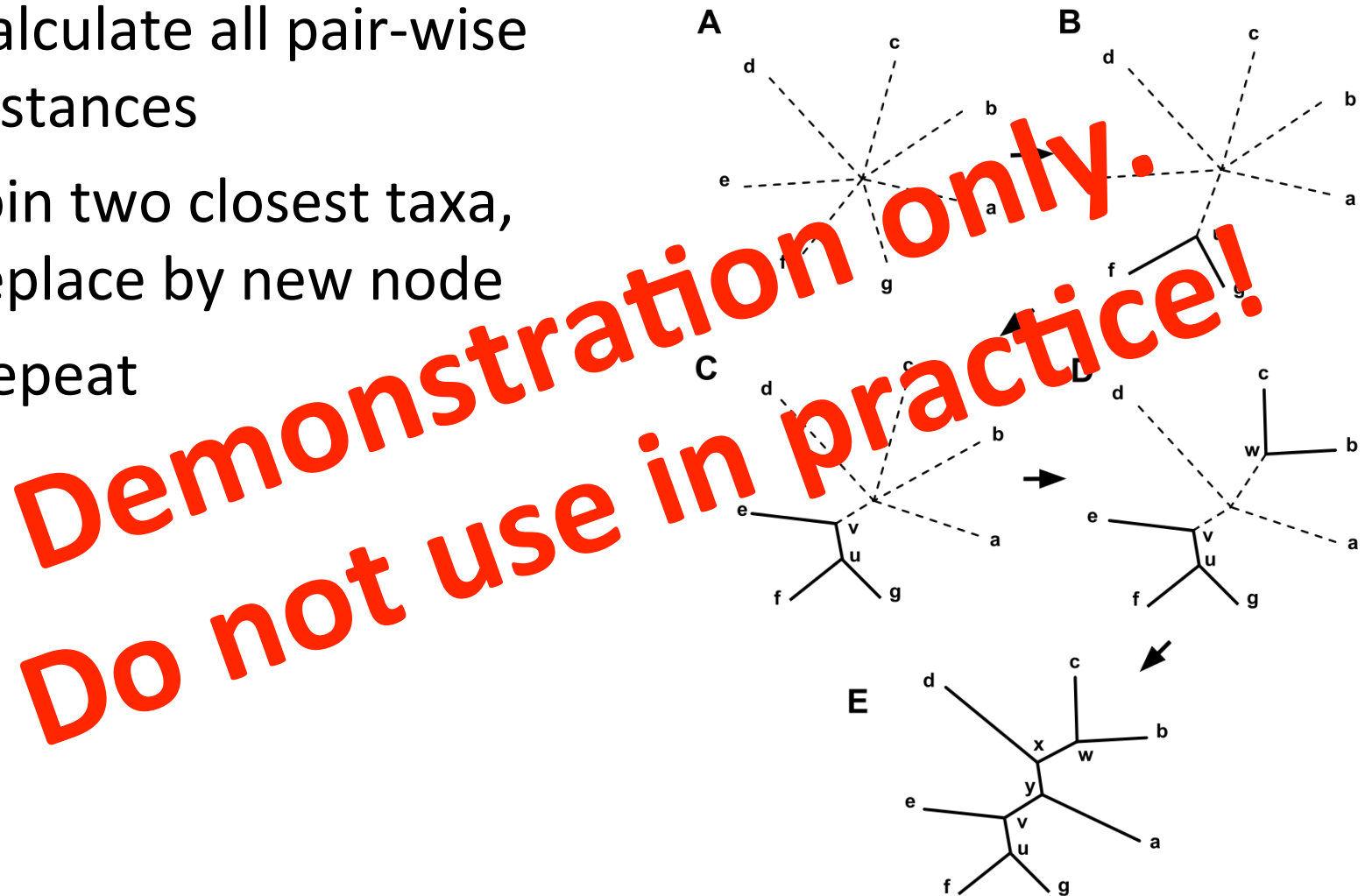
- Calculate all pair-wise distances
- Join two closest taxa, replace by new node
- Repeat



# Tree-building methods:

## 1. Neighbor-joining

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



# Tree-building methods:

## 2. Maximum likelihood

- Builds likelihood model of molecular evolution
- Maximizes:  
 $\text{Pr}(\text{tree} \mid \text{sequence data})$
- Commonly used software:  
RAxML, FastTree2

# Tree-building methods:

## 3. Bayesian

- Builds likelihood model of molecular evolution
- Calculates:  
 $\text{Pr}(\text{sequence data} \mid \text{tree})$
- Commonly used software:  
MrBayes, BEAST