Multiple Sequence Alignments

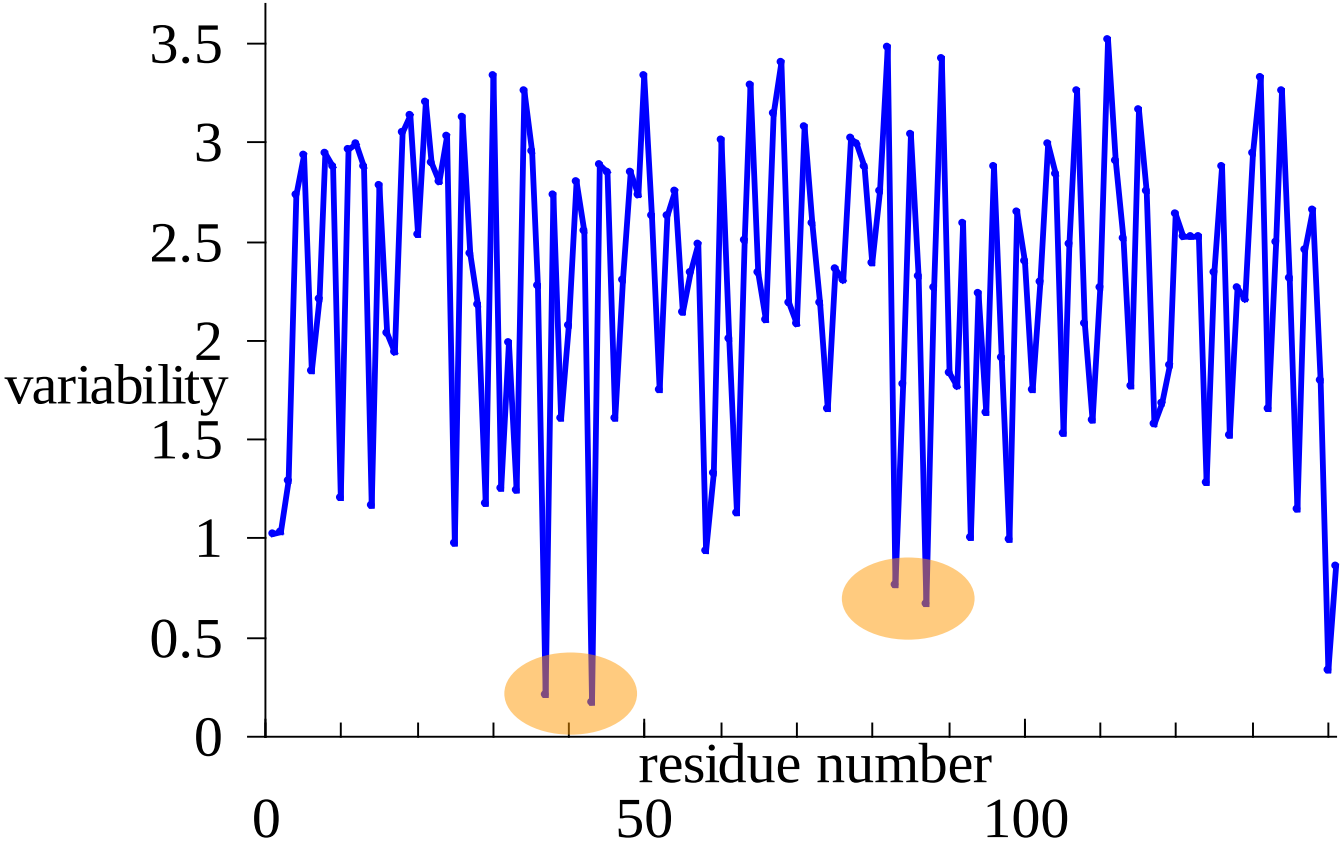
- for biology / for phylogeny
- Here – mostly proteins

```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALEKMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPDDKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTHVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEAWERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEAWERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
VLSADDKANIKAAWGKIGGHGAEYGAEALERMFCSFPTTKTYFPHFDVSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEAFERMFLSFPTTKTYFPHFDLSHGSAQVKGQG
VLSPADKTNVKAAWGKVGAHAGEYGAEAFERMFLSFPTTKTYFPHFDLSHGSAQVKGQA
VLSAADKSNVKAAWGKVGGNAGAYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKSNVKATWDKIGSHAGEYGGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHG
VLSPADKSNVKAAWGKVGGHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTGTYFPHFDLSHGSAQVKGHG
VLSSADKNNVKACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFDLSHGSAQVQAHG
VLSAADKSNVKAAWGKVGGNAGAYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSANDKSNVKAAWGKVGNHAPEYGAEALERMFLSFPTTKTYFPHFDLSHGSSQVKAHG
VLSPADKSNVKAAWGKVGGHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
    …         …          … …
```

- data for a haemoglobin
- summarise this data

Conservation / variability

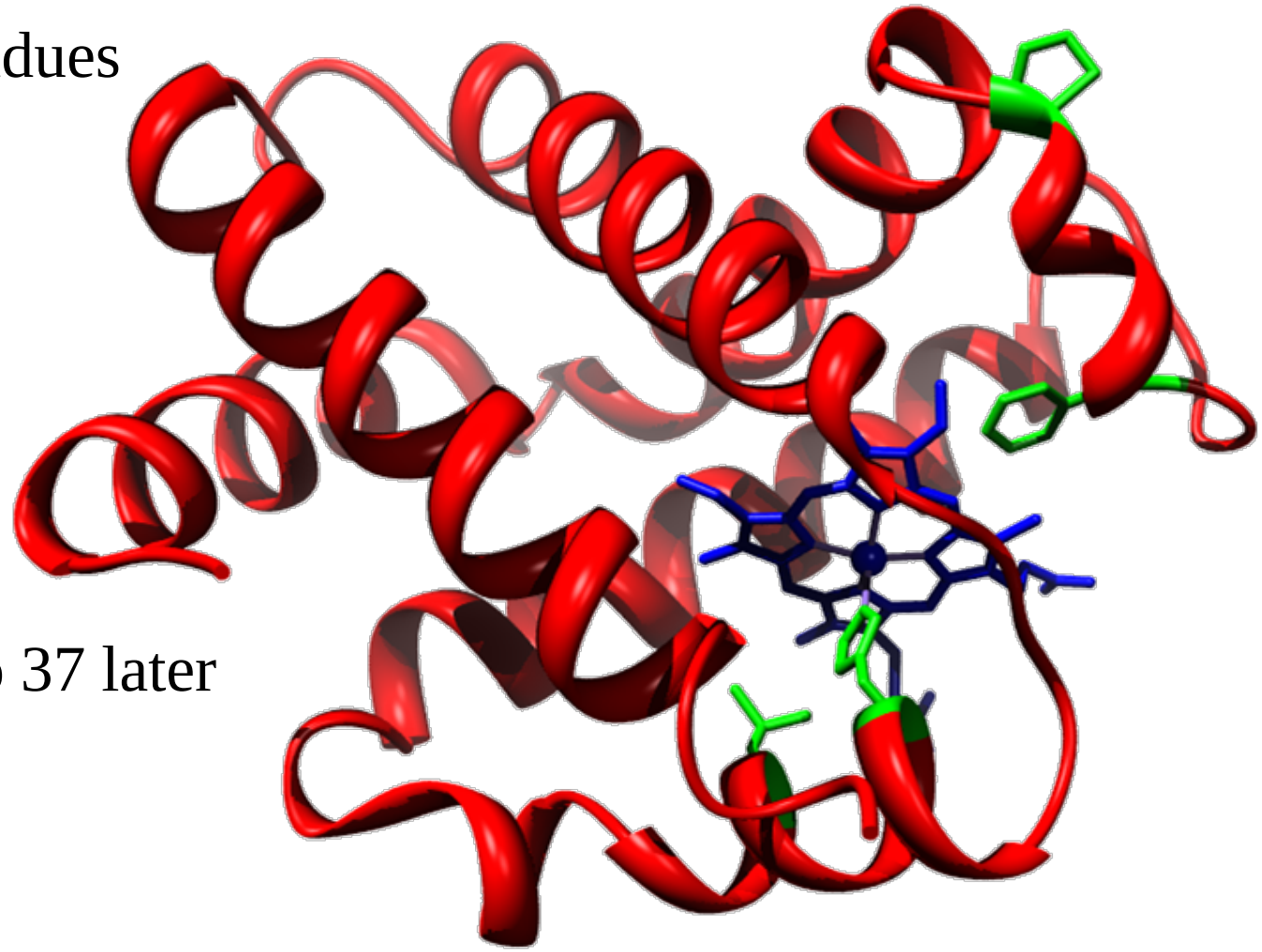
- look at residues 37, 43, 83 and 87



- how do we get these and what does it mean ?
- what does it mean for this protein ?

Conserved residues

- proximity to haem group
  - green residues



- more on pro 37 later

Beliefs – multiple alignments

Most proteins found in many organisms

- rarely identical
- how much they vary will reflect evolution (phylogeny)
- where they vary ? importance of residues

How many homologues might you have ?

- many
  - some DNA replication proteins – every form of life
  - some glycolysis proteins – from bacteria to man
  - ..
- few
  - some exotic viral proteins
  - some messengers exclusively in human biochemistry
  - …

Costs

- two sequence alignment
  - optimal path through $n \times m$ matrix
- three sequence alignment
  - optimal path through $n \times m \times p$ matrix
- four sequence alignment
  - …
- m sequence alignment of n residues…. $O(n^m)$

- excuse to use lots of approximations
  - no guarantee of perfect answer

- reasonable starting point
  - begin with pairs of proteins

progressive multiple sequence alignment

- align two sequences
- while more sequences
  - align next to existing alignment
- tools
  - align a sequence to existing alignment – easy
  - list of all pairwise alignments – coming

aligning to existing alignment

- sequence 1 normal sequence
- sequence 2 combination/average/profile

ACDEFG

ACEEFG align first to get AA CC DE EE FF GG

AAEEQG

	AA	CC	DE	EE	FF	GG
A	1	0	0	0	0	0
A	1	0	0	0	0	0
E	0	0	0.5	1	0	0
E	0	0	0.5	1	0	0
Q	0	0	0	0	0	0
G	0	0	0	0	0	1

- an average might have *n* sequences

order of alignments

- GG, DGG, DGD
- D G D is as good as D G D
  G G G G

  but consider optimal alignment

```
D G D        D G D
  G G        G G
D G G        D G G
```

- choice at first step determines quality of final alignment
- can one guarantee final best answer ? No
  - can one reduce the likelihood of wrong answers ? yes

progressive alignments

- more similar sequences
  - less problems in alignments, fewer errors

- for all *n* sequences
  - calculate pair-wise alignment

sort similarities
while sequence left
    select next sequence
    add next sequence to aligned set

use some kind of average similarity

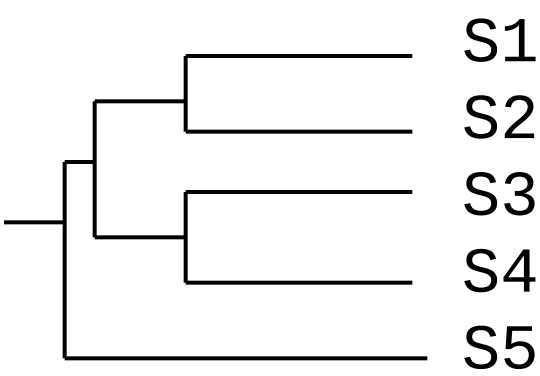
Progressive alignment - tree

S1	ATCTCGAGA
S2	ATCCGAGA
S3	ATGTCGACGA
S4	ATGTCGACAGA
S5	ATTCAACGA

Compute pairwise alignments, calculate the distance matrix

	S1	S2	S3	S4	S5
S1	-				
S2	.11	-			
S3	.20	.30	-		
S4	.27	.36	.09	-	
S5	.30	.33	.23	.27	-

calculate guide tree



Multiple alignment from guide tree

- av(S1,S2) is average of S1 and S2

- gaps at early stages remain
- problems..
- S1/S2 and S3/S4 good
  - no guarantee of S1/S4 or S2/S3

align S1 with S2

```
S1      ATCTCGAGA
S2      ATC-CGAGA
```

align S3 with S4

```
S3      ATGTCGAC-GA
S4      ATGTCGACAGA
```

align av(S1,S2) with av(S3,S4)

```
S1      ATCTCGA--GA
S2      ATC-CGA--GA
S3      ATGTCGAC-GA
S4      ATGTCGACAGA
```

align av(S1,S2,S3,S4) with S5

```
S1      ATCTCGA--GA
S2      ATC-CGA--GA
S3      ATGTCGAC-GA
S4      ATGTCGACAGA
S5      AT-TCAAC-GA
```

Properties of alignment

- my scheme ? bit simple
- any scheme – impossible to guarantee optimality
- reflection of evolution ? not quite – more later
- how useful ? very - conservation

```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VITP-EQSNVKAAWGKVGAHAGEYGAEAIEQMFLSYPTTKTYFP-FDLSHGSAQIKGHG
MLSPGDKTQVQAGFGRVGAHAG--GAEAVDRMFLSFPTTKSFFPYFELTHGSAQVKGHG
VLSPAEKTNIKAAWGKVGAHAGEYGAEAAEKMF-SYPSTKTYFPHFDISHATAQ-KGHG
–VTPGDKTNLQAGW-KIGAHAGEYGAEALDRMFLSFPTTK-YFPHYNLSHGSAQVKGHG
VLSPAEKTNVKAAWGRVGAHAGDYGAEAGERMFLSFPSTQTYFPHFDLS-GSAQVQAHA
VLSPDDKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
```

conserved

not conserved

- quantify this

quantifying conservation

- Gibbs entropy
  - how much disorder do I have ?
  - in how many states may I find the system ?

$$S = -k \sum_{i=1}^{N_{states}} p_i \ln p_i$$

- Our question
  - look at a column – how much disorder is there ?

```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VITP-EQSNVKAAWGKVGAHAGEYGAEAIEQMFLSYPTTKTYFP-FDLSHGSAQIKGHG
MLSPGDKTQVQAGFGRVGAHAG--GAEAVDRMFLSFPTTKSFFPYFELTHGSAQVKGHG
VLSPAEKTNIKAAWGKVGAHAGEYGAEAAEKMF-SYPSTKTYFPHFDISHATAQ-KGHG
–VTPGDKTNLQAGW-KIGAHAGEYGAEALDRMFLSFPTTK-YFPHYNLSHGSAQVKGHG
VLSPAEKTNVKAAWGRVGAHAGDYGAEAGERMFLSFPSTQTYFPHFDLS-GSAQVQAHA
VLSPDDKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
```

much order

little order

- Calculate an "entropy" for each column

toy alignment entropy

- first column is boring
- second
  - $p_D = 5/7$
  - $p_E = 1/7$
  - $p_N = 1/7$

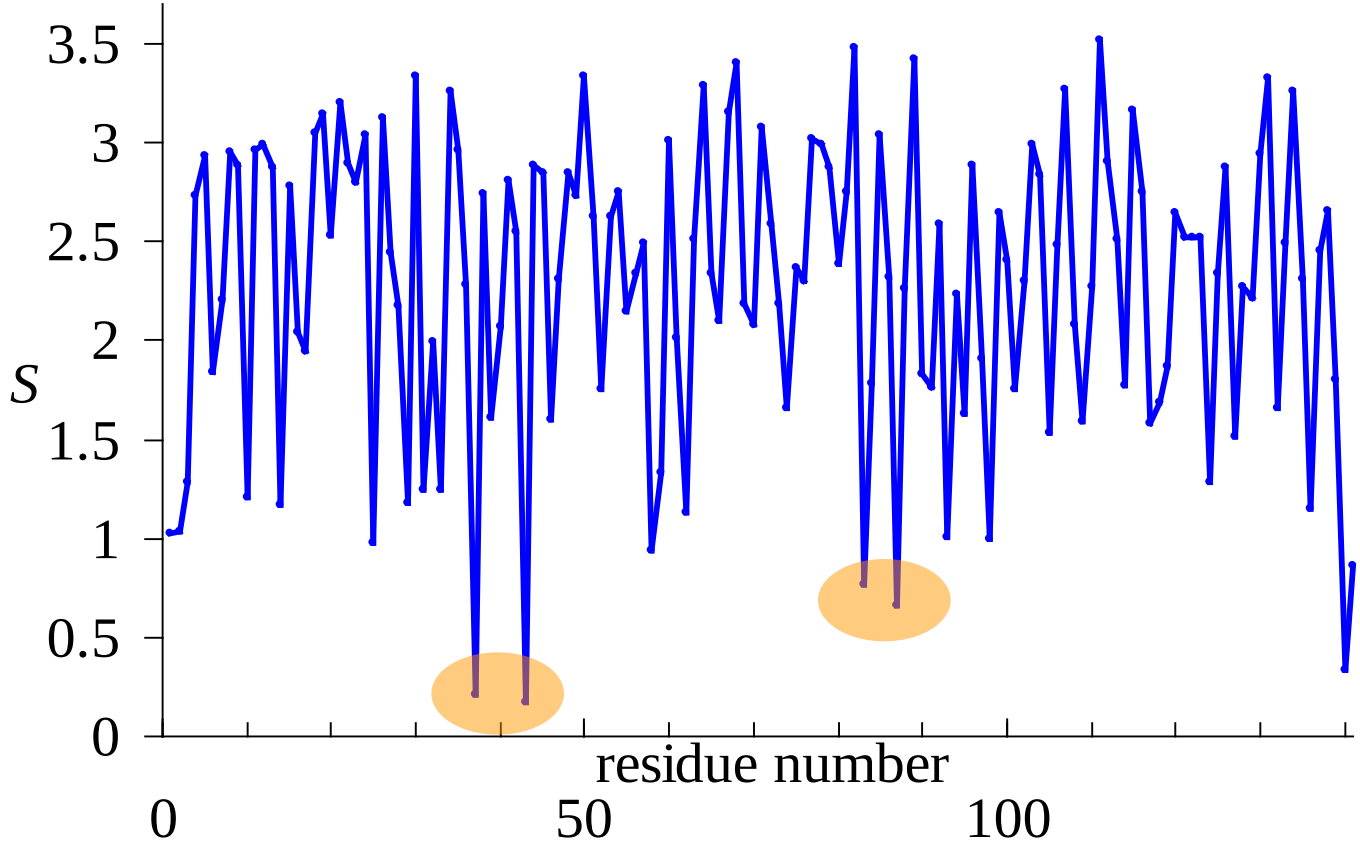
```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VITP-EQSNVKAAWGKVGAHAGEYGAEAIEQMFLSYPTTKTYFP-FDLSHGSAQIKGHG
MLSPGDKTQVQAGFGRVGAHAG--GAEAVDRMFLSFPTTKSFFPYFELTHGSAQVKGHG
VLSPAEKTNIKAAWGKVGAHAGEYGAEAAEKMF-SYPSTKTYFPHFDISHATAQ-KGHG
–VTPGDKTNLQAGW-KIGAHAGEYGAEALDRMFLSFPTTK-YFPHYNLSHGSAQVKGHG
VLSPAEKTNVKAAWGRVGAHAGDYGAEAGERMFLSFPSTQTYFPHFDLS-GSAQVQAHA
VLSPDDKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
```

$$S = -\left(\frac{5}{7}\ln\frac{5}{7} + \frac{1}{7}\ln\frac{1}{7} + \frac{1}{7}\ln\frac{1}{7}\right)$$

$$\approx 0.8$$

- example from start of this topic

- look at residues 37, 43, 83 and 87

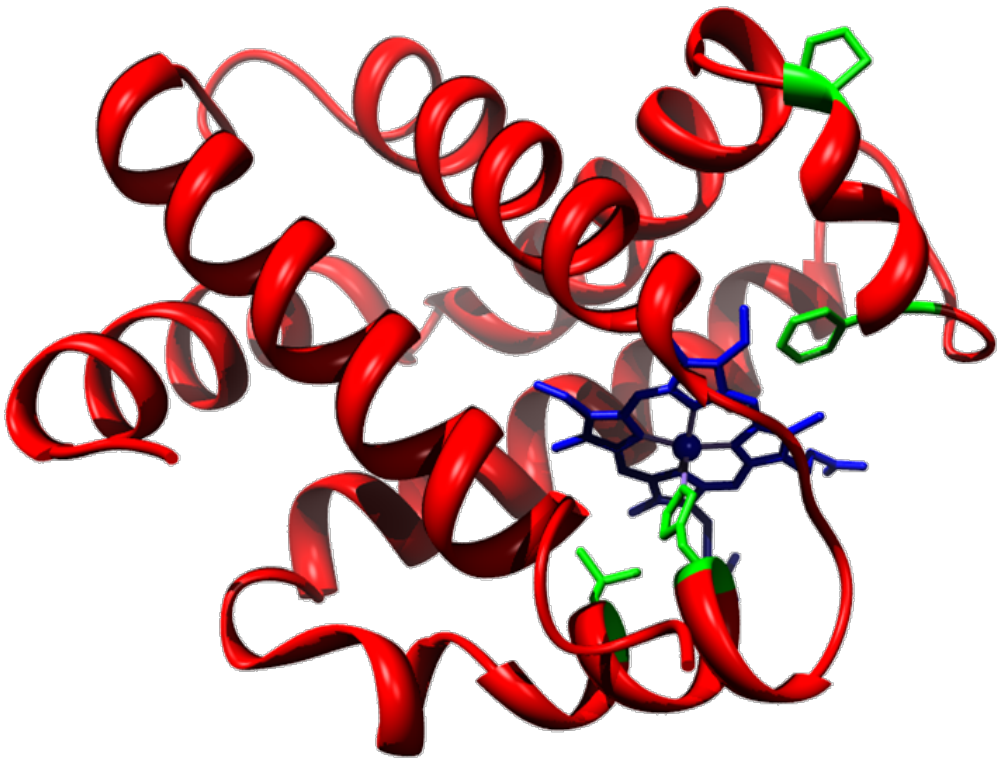


- 4 residues (maybe more) stand out as conserved
  - why ?

conservation – function / structure

- 3 of the sites
  - interact with haem group

- Look at fourth site
  - proline
  - end of a helix



- what is special about proline ?
  - no Hbond donor
- here – if it mutates, maybe haemoglobin does not fold

use of information

- structure not known
  - usually many sequences known
  - which sites can be modified ?
- even if structure known
  - active site residues probably amongst those conserved
- are the non-conserved residues boring ?

do not trust conservation

Imagine: two possible systems for some important enzyme

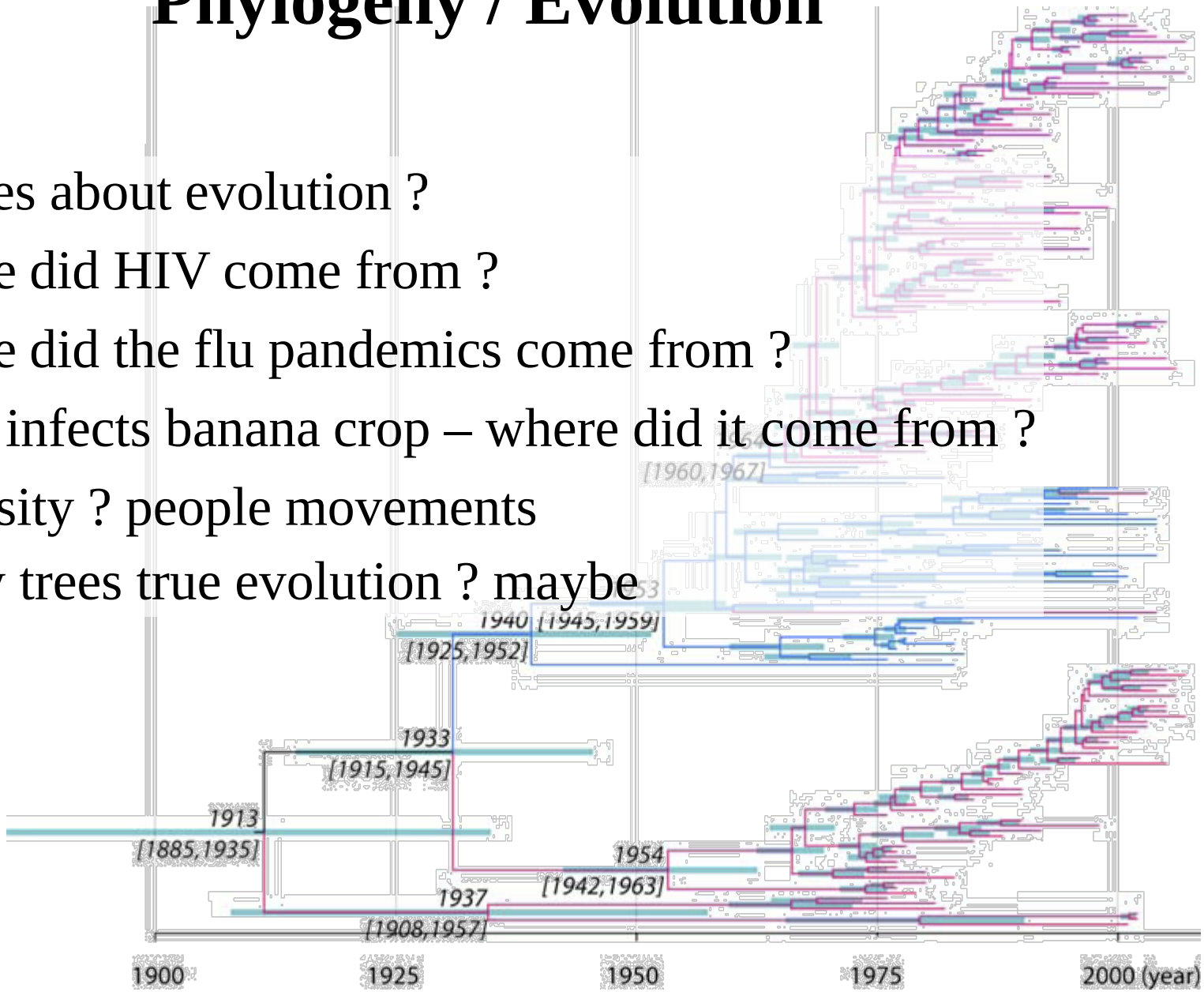
1. active site fits to essential biochemistry
   - any mutation – you lose
   - you see active site residues as conserved in a conservation plot
2. maybe enzyme is not absolutely perfect
   - some mutations kill you
   - some mutations OK
   - site does not appear perfectly conserved

If you have the choice, where would you evolve to ?

1. very fragile
2. likely to survive mutations

Phylogeny / Evolution

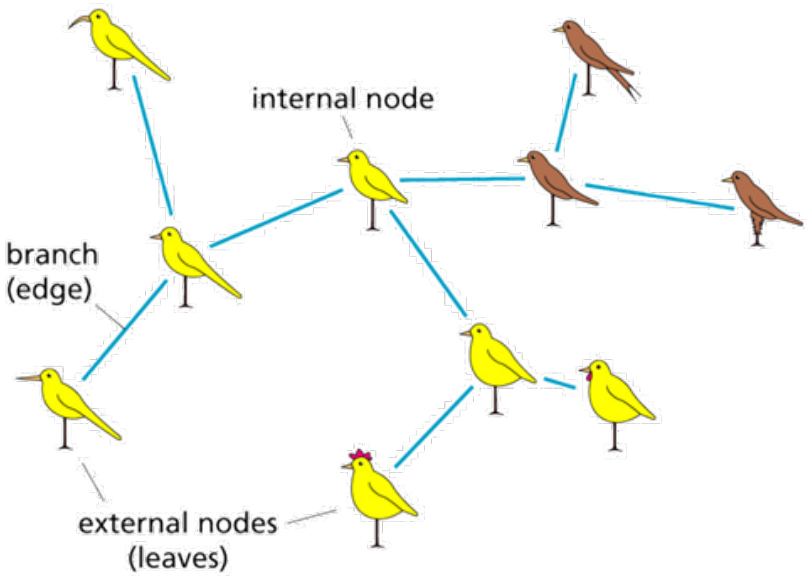
- who cares about evolution ?
  - where did HIV come from ?
  - where did the flu pandemics come from ?
  - virus infects banana crop – where did it come from ?
  - curiosity ? people movements
- were my trees true evolution ? maybe



Chen R. and Holmes, E.C., J Mol Evol (2010) 70,98–105

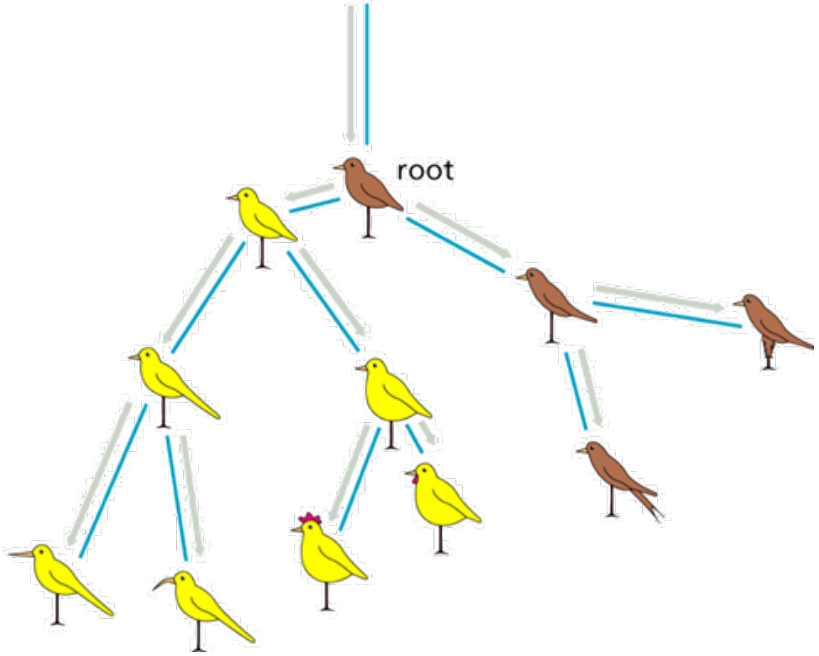
different types of trees

unrooted



easier
no direction
of time

rooted

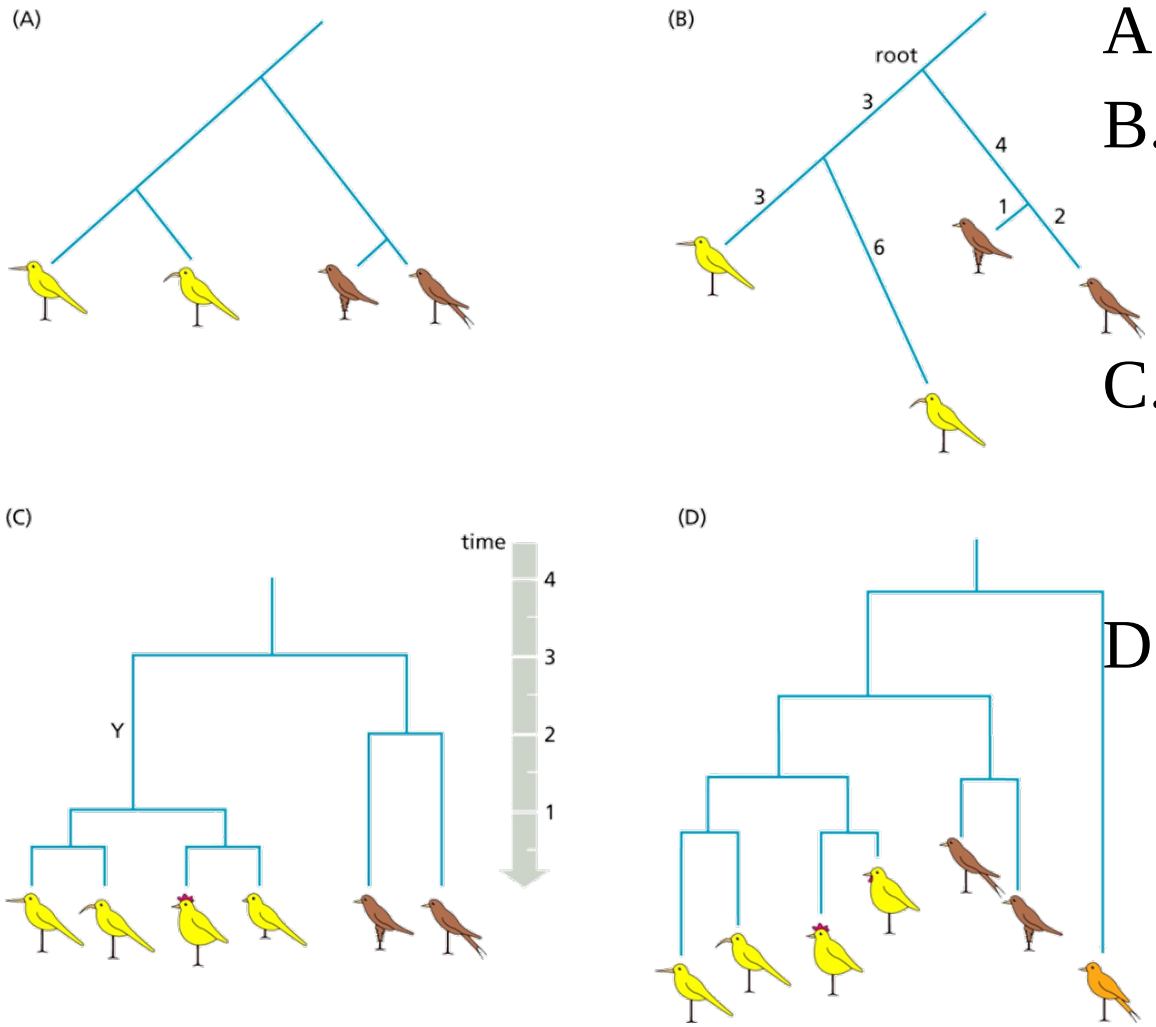


clear evolutionary
path

- only the leaves are real birds

from Zvelebil, M, Baum, J.O. 2008, Understanding Bioinformatics, Garland Science

how ambitious are you ?



A. "clades"
B. branch lengths = evolutionary divergence
C. ultrametric – assume constant rate of mutation
D. add an outlier to get a root

from Zvelebil, M, Baum, J.O. 2008, Understanding Bioinformatics, Garland Science

Phylogeny method 1 - clustering

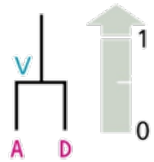
- very simple – nearest neighbour cluster method

```
get distance between each sequence
each sequence in own cluster
while more than one cluster
  join nearest two clusters
```

- textbook picture

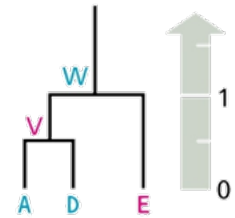
(A)

d_{ij}	A	B	C	D	E	F
A	–	6	8	1	2	6
B		–	8	6	6	4
C			–	8	8	8
D				–	2	6
E					–	6



(B)

d_{ij}	B	C	E	F	V
B	–	8	6	4	6
C		–	8	8	8
E			–	6	2
F				–	6

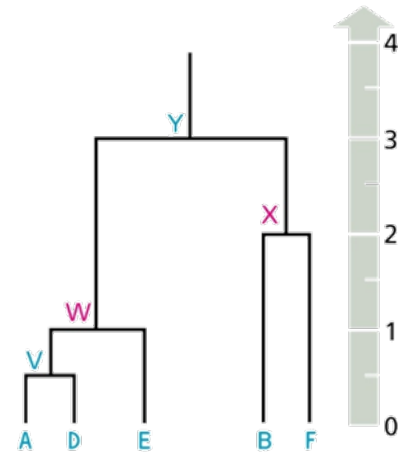


(C)

d_{ij}	B	C	F	W
B	–	8	4	6
C		–	8	8
F			–	6

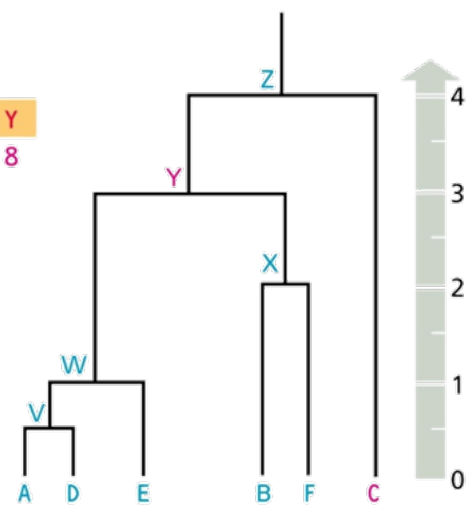
(D)

d_{ij}	C	W	X
C	–	8	8
W		–	6



(E)

d_{ij}	C	Y
C	–	8



- debate..
  - what is distance to a cluster ?
  - simple averaging

from Zvelebil, M, Baum, J.O. 2008, Understanding Bioinformatics, Garland Science]

Reliability

- trees too simple, but nice properties
  - fast – easy to estimate reliability
  - example – jackknifing / bootstrapping
- assume we have a multiple sequence alignment

```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VITP-EQSNVKAAWGKVGAHAGEYGAEAIEQMFLSYPTTKTYFP-FDLSHGSAQIKGHG
MLSPGDKTQVQAGFGRVGAHAG--GAEAVDRMFLSFPTTKSFFPYFELTHGSAQVKGHG
VLSPAEKTNIKAAWGKVGAHAGEYGAEAAEKMF-SYPSTKTYFPHFDISHATAQ-KGHG
–VTPGDKTNLQAGW-KIGAHAGEYGAEALDRMFLSFPTTK-YFPHYNLSHGSAQVKGHG
VLSPAEKTNVKAAWGRVGAHAGDYGAEAGERMFLSFPSTQTYFPHFDLS-GSAQVQAHA
VLSPDDKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
```

- work on half the columns
  - should get same answer as original

Reliability

```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VITP-EQSNVKAAWGKVGAHAGEYGAEAIEQMFLSYPTTKTYFP-FDLSHGSAQIKGHG
MLSPGDKTQVQAGFGRVGAHAG--GAEAVDRMFLSFPTTKSFFPYFELTHGSAQVKGHG
VLSPAEKTNIKAAWGKVGAHAGEYGAEAAEKMF-SYPSTKTYFPHFDISHATAQ-KGHG
–VTPGDKTNLQAGW-KIGAHAGEYGAEALDRMFLSFPTTK-YFPHYNLSHGSAQVKGHG
VLSPAEKTNVKAAWGRVGAHAGDYGAEAGERMFLSFPSTQTYFPHFDLS-GSAQVQAHA
VLSPDDKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
```

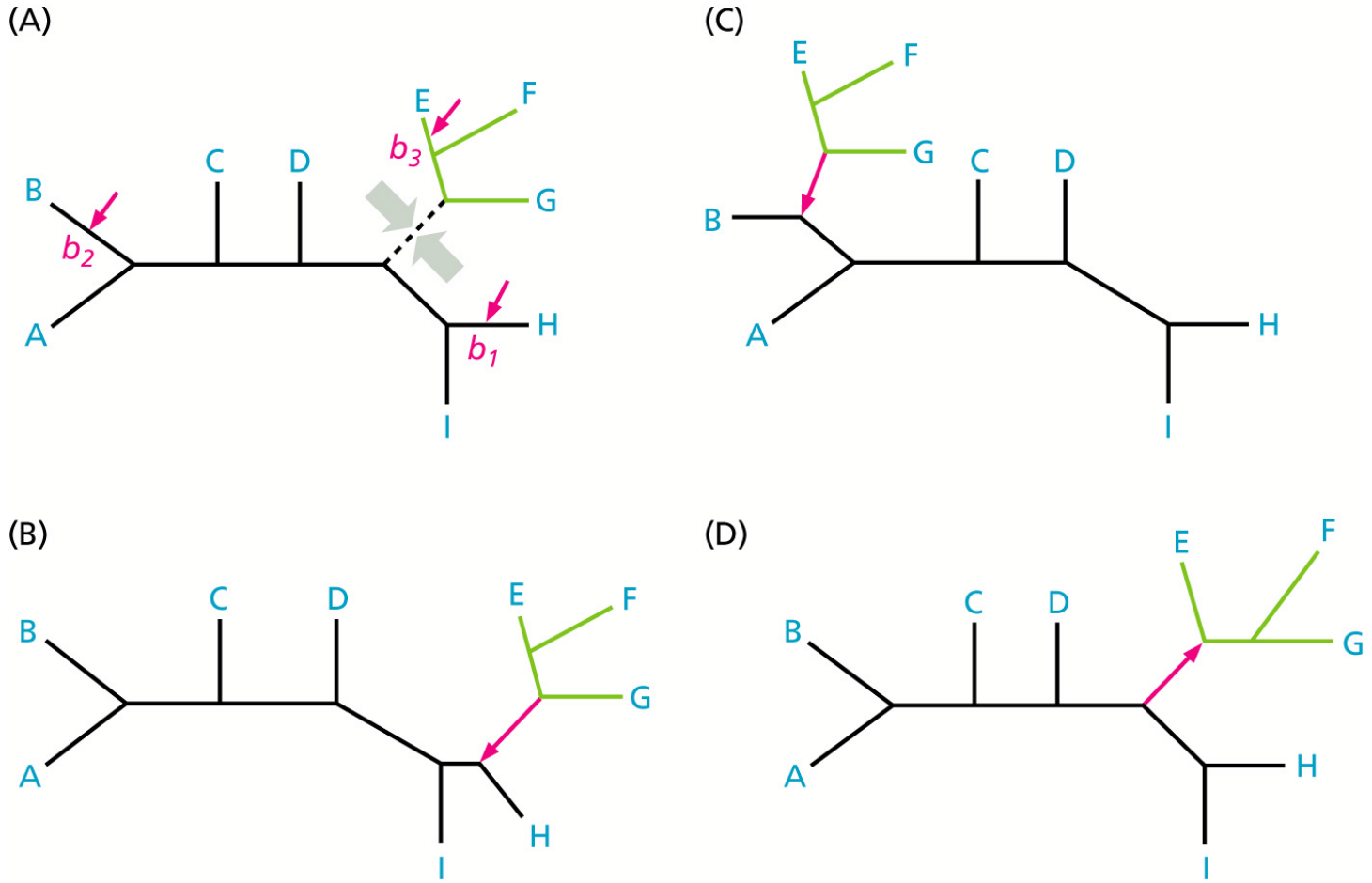
- example
  - *m* times
    - pick *n* (5%) columns randomly delete
    - from remaining columns pick *n*
    - put in place of missing columns
  - in each of *m* trees
    - look for frequency of branch
  - simple estimate of reliability / robustness

more complicated methods

- what is the best tree ?
  - some model of evolution
  - the one with fewest mutations at internal nodes

- general approach
  - start with a simple fast tree
  - calculate cost
    - total number of mutations
  - try to move branches / nodes to reduce cost

example moves within tree

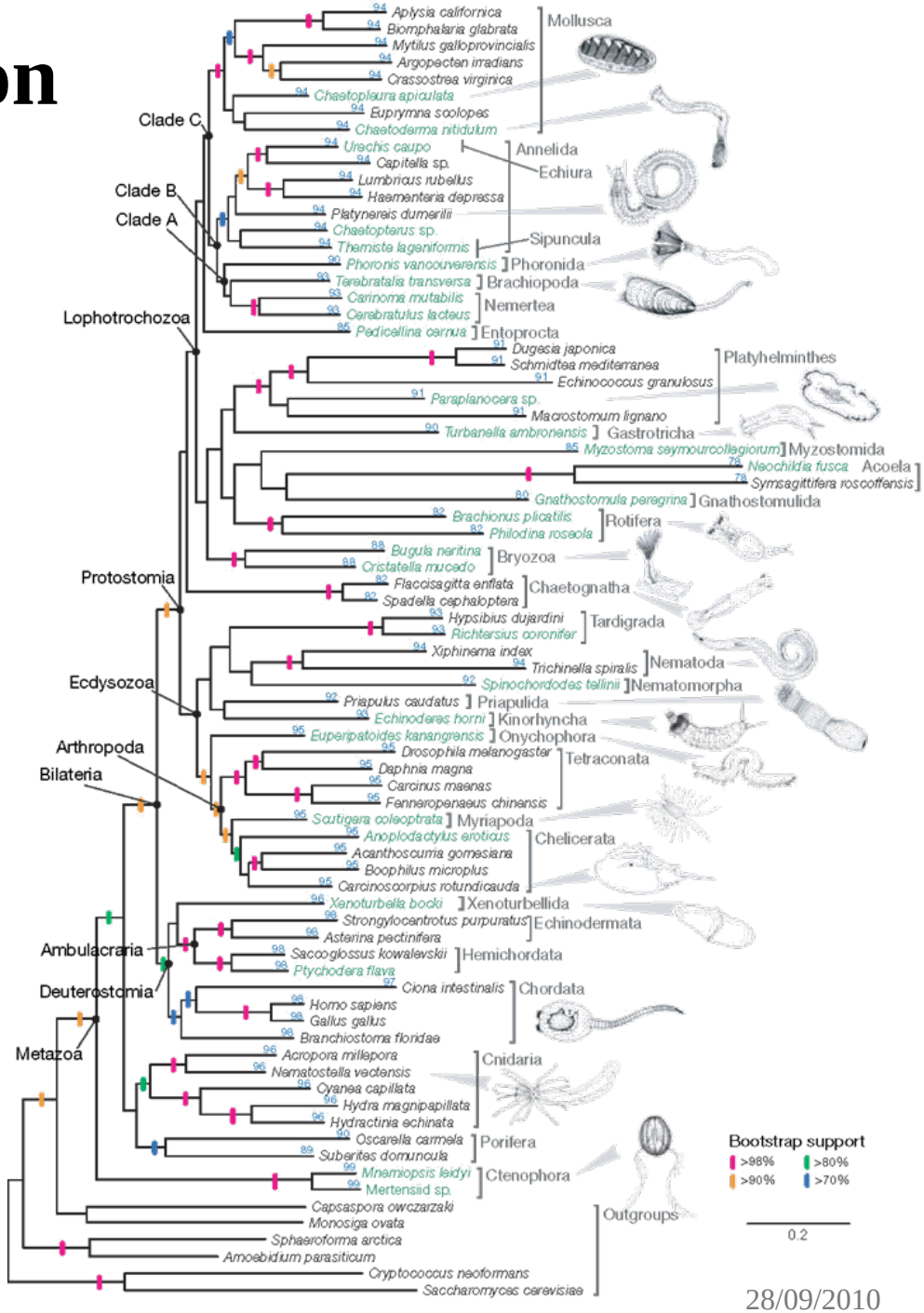


Searching for trees

- how hard ?
  - huge search space
  - local minima
  - cost functions can be complicated
- good points
  - find many trees with similar scores
  - in what % of trees are certain branches reproduced ?
    - reliability score
- huge literature example

Monster calculation

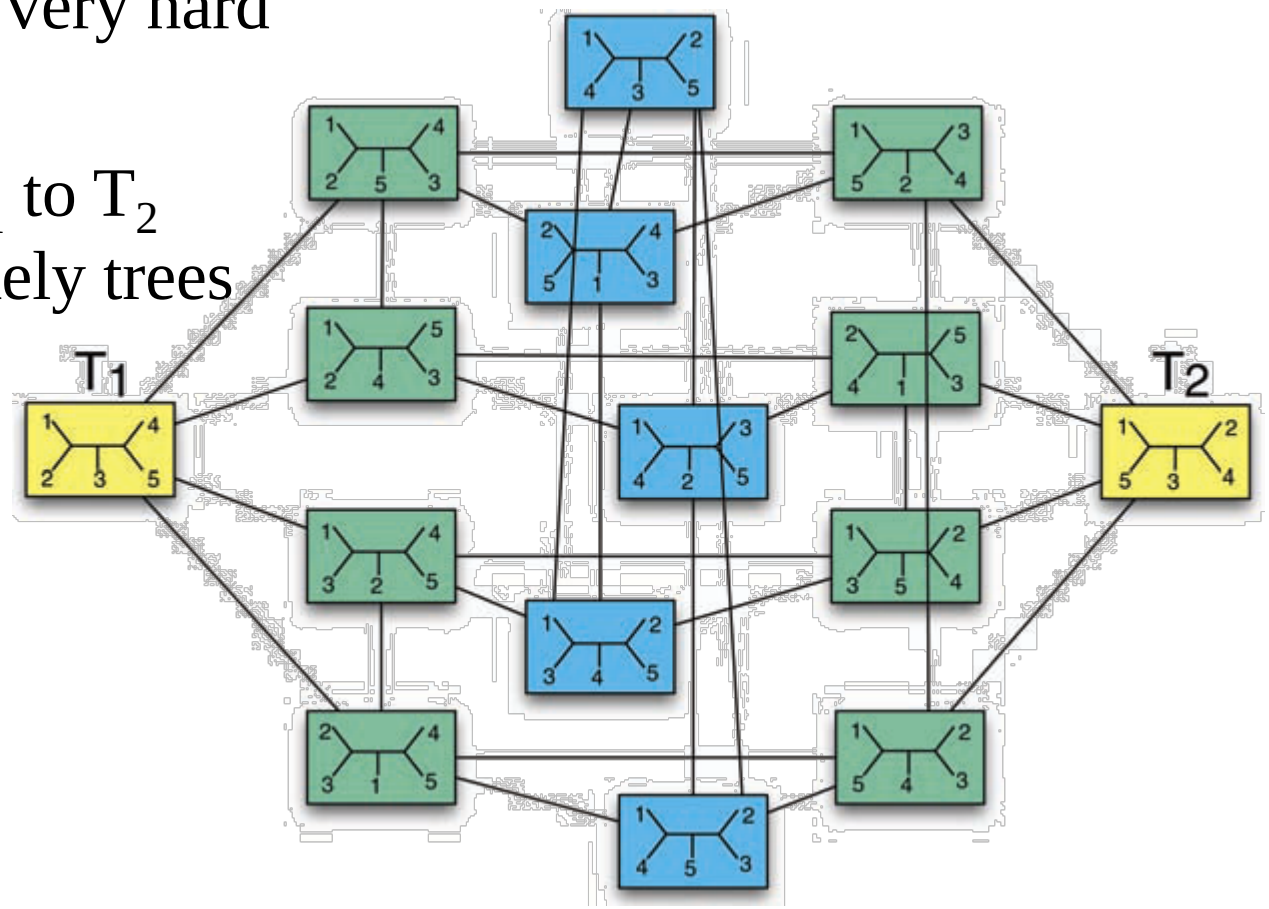
- we are usually placed near chickens
- we are not so reliably placed with little worms
- how long does this take?
  - months on 120 processors



Dunn, CW et al, Nature, 402, 745-750 (2008)

More fun aspects

- how much should you believe huge calculations ?
- sampling can be very hard
- all paths from T_1 to T_2 go through unlikely trees



Mossel & Vigoda, Science, 309, 2207-2209 (2005)

Lots of assumptions / limitations

- common rates of evolution
  - bacteria reproduce faster than people
  - some proteins hardly mutate (DNA copying)
  - within one protein – rates vary
- rare events
  - duplications, fragmenting
  - transfer of genes (drug resistance)