

Multiple sequence alignments similarity without sequence similarity

Andrew Torda,
Bioinformatics,
Sommersemester 2011

- Gute Nachrichten
 - "Understanding Bioinformatics" – 5 + 1 Kopien in Stabi

Bis Jetzt

- Man hat eine Sequenz (Protein oder Nukleotid)
- Man will so viel wie möglich finden, um
 - Struktur vorherzusagen
 - Funktion vorherzusagen
- Jetzt Alignments, Evolution & Funktion

Multiple alignments

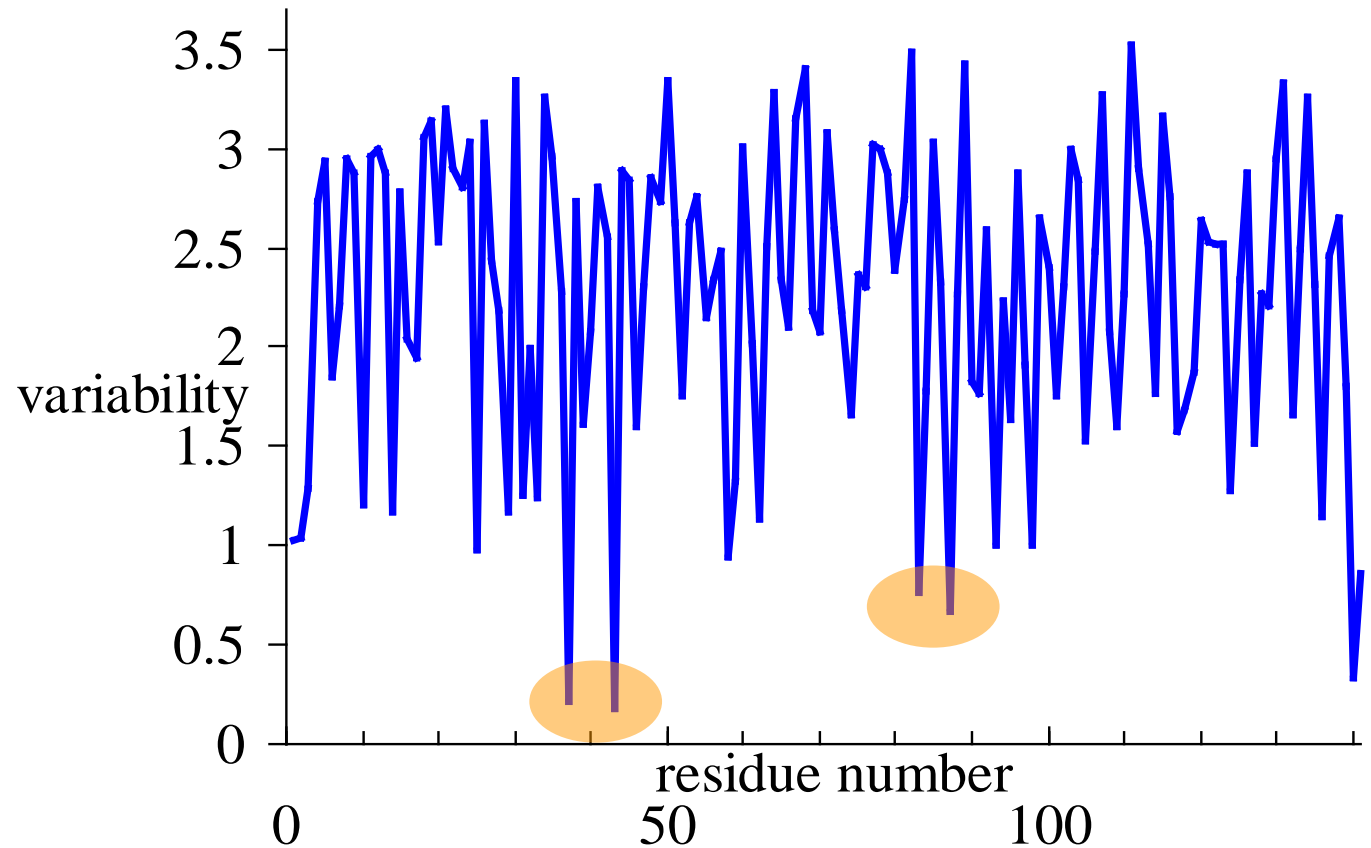
- mostly for proteins
- what does a set of sequences look like ?
- data for a haemoglobin
- summarise this data

```
VLSPADKTNVKAAWGKVGHAHAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGHAHAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGHAHAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGHAHAGEYGAEALEKMFSLFPTTKTYFPHFDLSHGSAQVKGHG
  LSPADKTNVKAAWGKVGHAHAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGHAHAGDYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
VLSPDDKTNVKAAWGKVGHAHAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGHAHAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGHAHAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTHVKAAWGKVGHAHAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGHAHAGEYGAEAWERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGHAHAGEYGAEAWERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGHAHAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGHAHAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGHAHAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
VLSADDKANIKAAWGKIGGHGAEYGAEALERMFCSFPTTKTYFPHFDVSHGSAQVKGHG
MLSPADKTNVKAAWGKVGHAHAGEYGAEAFERMFLSFPTTKTYFPHFDLSHGSAQVKGG
VLSPADKTNVKAAWGKVGHAHAGEYGAEAFERMFLSFPTTKTYFPHFDLSHGSAQVKGQA
VLSAADKSNVKAAWGKVGGNAGAYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGHAHAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKSNVKATWDKIGSHAGEYGGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHG
VLSPADKSNVKAAWGKVGGHAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGHAHAGEYGAEALERMFSLFPTTGTYFPHFDLSHGSAQVKGHG
VLSAADKNNVKACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFDLSHGSAQVQAHG
VLSAADKSNVKAAWGKVGGNAGAYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGHAHAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
VLSANDKSNVKAAWGKVGNHAPYEGAEALERMFSLFPTTKTYFPHFDLSHGSSQVKAHG
VLSPADKSNVKAAWGKVGGHAGDYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHG
```

... ..

Conservation / variability

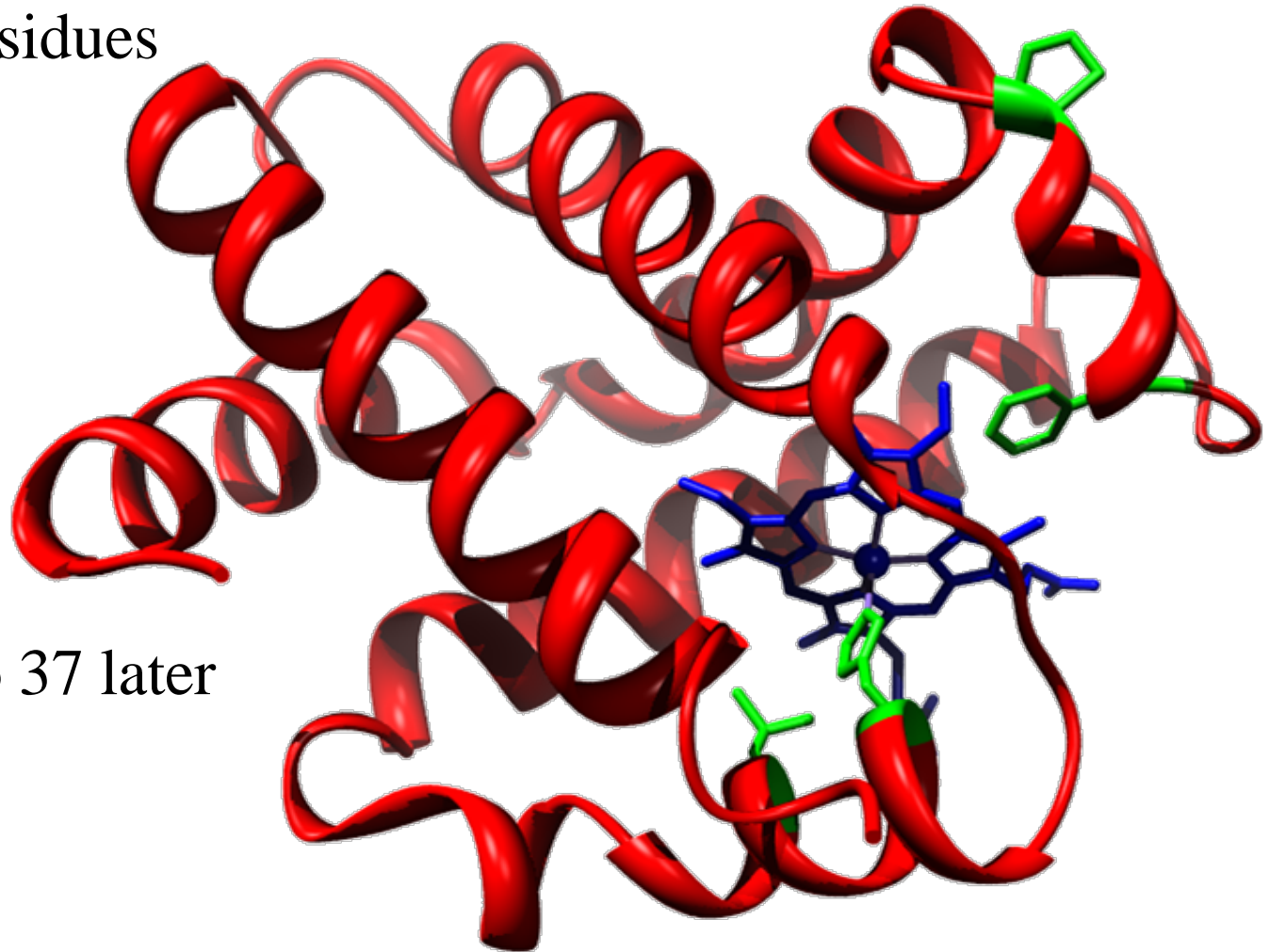
- look at residues 37, 43, 83 and 87



- 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 in multiple sequence alignments

Most proteins found in many organisms

- rarely identical
- where they are conserved will be connected with function
- how much they vary will reflect evolution (phylogeny)


How many homologues might you have ?

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

Many sequences - rigorous alignment

- 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

Scoring schemes

$$S_{a,b} = \sum_{i=1}^{N_{res}} match(s_{a,i}, s_{b,i})$$


- In pairwise problem

VLSPADKSNVKAGWGQVGAHAGDYGAEAI ERM YLSFPSTKTYFPHTDI SHGSAQVKGHG
 MLSPADKTNVKA AWGKVG AHAGEYGA EALERMFLSFPTTKTYFP HFDLSHGSAQVKGHG

- Sum over

where N_{res} is sequence length

- $match(s_{a,i}, s_{b,i})$ is the match/mismatch score of sequence a and b at position i

- invent a distance between two sequences like

$$d_{a,b} = 1 - \frac{S_{a,b}}{100 \times N_{res}} \quad \text{or} \quad d_{a,b} = \frac{1}{S_{a,b}}$$

- distance measure – mainly to see which sequences are most similar to each other

Scoring schemes for a multiple alignment

In the best alignment

- 1 is aligned to 2, 3, ..
- 2 to 3,4, ...

```

1 VLSPADKTNVKAAWGKVGGAHAGEYGAEALERMFLSFPPTTKTYFPHFDLSHGSAQVKGHG
2 VITP-EQSNVKAAWGKVGGAHAGEYGAEALEQMFLSYPTTKTYFFP-FDLSHGSAQIKGHG
3 MLSPGDKTQVQAGFGRVGAHAG--GAEALDRMFLSFPPTTKSFFPYFELTHGSAQVKGHG
4 VLSPAECTNIKAAWGKVGGAHAGEYGAEALEKMF-SYPSTKTYFPHFDISHATAQ-KGHG
5 -VTPGDKTNLQAGW-KIGAHAGEYGAEALDRMFLSFPPTK-YFPHYNLSHGSAQVKGHG
6 VLSPAECTNVKAAWGRVGAHAGDYGAEALERMFLSFPSTQTYFPHFDLS-GSAQVQAHA
7 VLSPDDKTNVKAAWGKVGGAHAGEYGAEALERMFLSFPPTTKTYFPHFDLSHGSAQVKGHG
    
```

Mission: for N_{seq} sequences

- S_{ab} : alignment score sequences a and b
- not quite possible
 - if I move sequences 4 and 5, may make a mess of 5 and 2

$$score = \sum_{b \neq a} \sum_{a=1}^{N_{seq}} S_{a,b}$$

Aligning average sequences

VLSPADKTNVKAAWGKVGHAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VITPAEKTNVKAAWGKVGHAHAGEYGAEALEQMFLSYPTTKTYFPHFDLSHGSAQIKGHG

and

IITPGDKTNVKAAGFKVGAHGGEYGAEALDRMFISFPSTKTYYPHFDSLHASAQVKAHG
VITPAEQTNIGAWGQIGAHAGDYAADALEQMFLSYPTSKTYFPYFDLTHGSAQIKGHG
VITPAEKTQVKAAWGKVGGHAGEYGAEAEIQMFLTYPTTQTYFPHFELSHGTAQIKGHG

- at each position
 - use some kind of average in scoring
 - if a column has $2 \times D$ and $1 \times E$ score
 - score as $\frac{2}{3} D + \frac{1}{3} E$
- later.. call the average of S1 and S2: $av(S1, S2)$

Summarise ingredients

- pairwise scores + distances
- ability to align little groups of sequences

Progressive alignments

- known as guide tree / progressive method
- steps
 - build a distance matrix
 - build a guide tree
 - build up overall alignment in pieces

Progressive alignment - tree

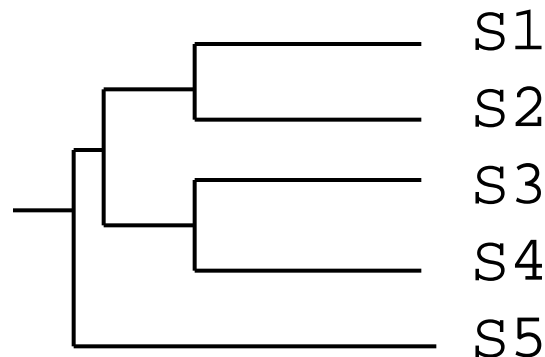
S1 ATCTCGAGA
S2 ATCCGAGA
S3 ATGTCGACGA
S4 ATGTCGACAGA
S5 ATTCAACGA

Compute pairwise alignments,
calculate the distance matrix



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

calculate guide tree



Multiple alignment from guide tree

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

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

align S1 with S2

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

align S3 with S4

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

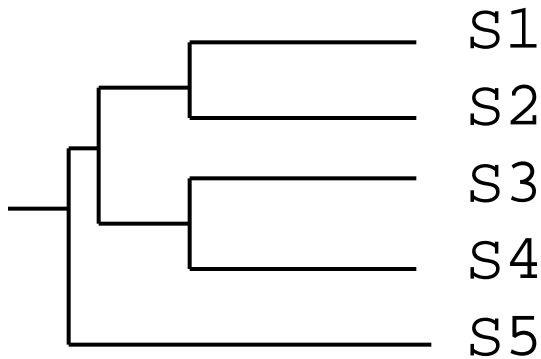
align $av(S1,S2)$ with $av(S3,S4)$

```
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
```

Problems and variations

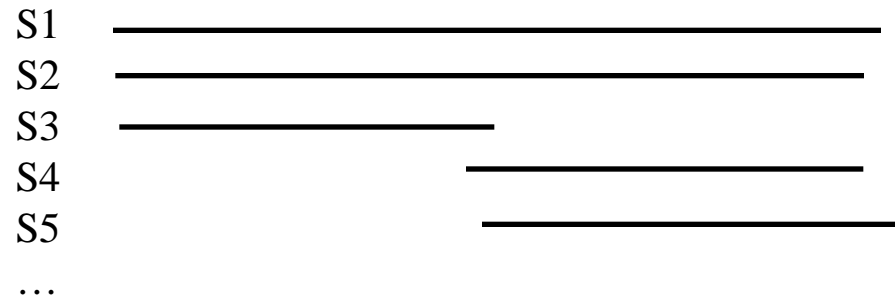


S1	-				
S2	.11	-			
S3	.20	.30	-		
S4	.27	.36	.09	-	
S5	.30	.33	.23	.27	-
	S1	S2	S3	S4	S5

What order should we join ?

- pairs are easy (S1+S2) and (S3+S4)
- which next ?

Real breakdown



- S1 and S2 are multi-domain proteins
 - S3 is not really related to S4 or S5
 - distance matrix elements are rubbish

Given an alignment

How reliable / believable ?

- set of very related proteins (an enzyme from 100 mammals)
 - no problem
- diverse proteins (an enzyme 100 organisms, bacteria to man)
 - maybe lots of little errors
- can break completely (domain example)

Is the tree a "phylogeny" ? A reflection of evolution ?

- more later

Measuring conservation / entropy

- Gibbs entropy

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

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

- Our question

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

```
VLSPADKTNVKAAWGKVGAAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VITP-EQSNVKAAWGKVGAAHAGEYGAEAEIQMFLSYPTTKTYFP-FDLSHGSAQIKGHG
MLSPGDKTQVQAGFGRVGAHAG--GAEAVDRMFLSFPTTKSFFPYFELTHGSAQVKGHG
VLSPAECTNIKAAWGKVGAAHAGEYGAEAAEKMF-SYPSTKTYFPHFDLSHATAQ-KGHG
-VTPGDKTNLQAGW-KIGAAHAGEYGAEALDRMFLSFPTTK-YFPHYNLSHGSAQVKGHG
VLSPAECTNVKAAWGRVGAHAGDYGAEAGERMFLSFSTQTYFPHFDLS-GSAQVQAHA
VLSPDDKTNVKAAWGKVGAAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
```

no
disorder

much
disorder

- Calculate an "entropy" for each column

Entropy

- We can forget k (Boltzmann – just scaling)

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

- We have a protein

- 20 possible states

- What if a residue is always conserved ?

- $S = \ln(1) = 0$ (no entropy)

- What if all residues are equally likely ?

- $p_i = 1/20$

$$S = - \sum_{i=1}^{20} \frac{1}{20} \ln \frac{1}{20} = -20 \cdot \frac{1}{20} \ln \frac{1}{20}$$

$$\approx 3$$

- my toy alignment..

Entropy

- first column is boring
- second

- $p_D = 5/7$
- $p_E = 1/7$
- $p_N = 1/7$

```
VLSPADKTNVKAAWGKVGAFAGEYGAELERMFLSFPTTKTYFPHEDLSHGSAQVKGHG  
VITP-EQSNVKAAWGKVGAFAGEYGAEAIEQMFLSYPTTKTYFP-ELSHGSAQIKGHG  
MLSPGDKTQVQAGFGRVGAHAG--GAEAVDRMFLSFPTTKSFFPYEELTHGSAQVKGHG  
VLSPAECTNIKAAWGKVGAFAGEYGAEEAEKMF-SYPSTKTYFPHEDLSHATAQ-KGHG  
-VTPGDKTNLQAGW-KIGAFAGEYGAELDRMFLSFPTTK-YFPHYNLSHGSAQVKGHG  
VLSPAECTNVKAAWGRVGAHAGDYGAEEGERMFLSFSTQTYFPHEDLS-GSAQVQAHA  
VLSPDDKTNVKAAWGKVGAFAGEYGAELERMFLSFPTTKTYFPHEDLSHGSAQVKGHG
```

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

$$\approx 0.8$$

- example from start of this topic

Entropy from DNA

- exactly as for proteins
- will numbers be larger or smaller ?

- max possible entropy

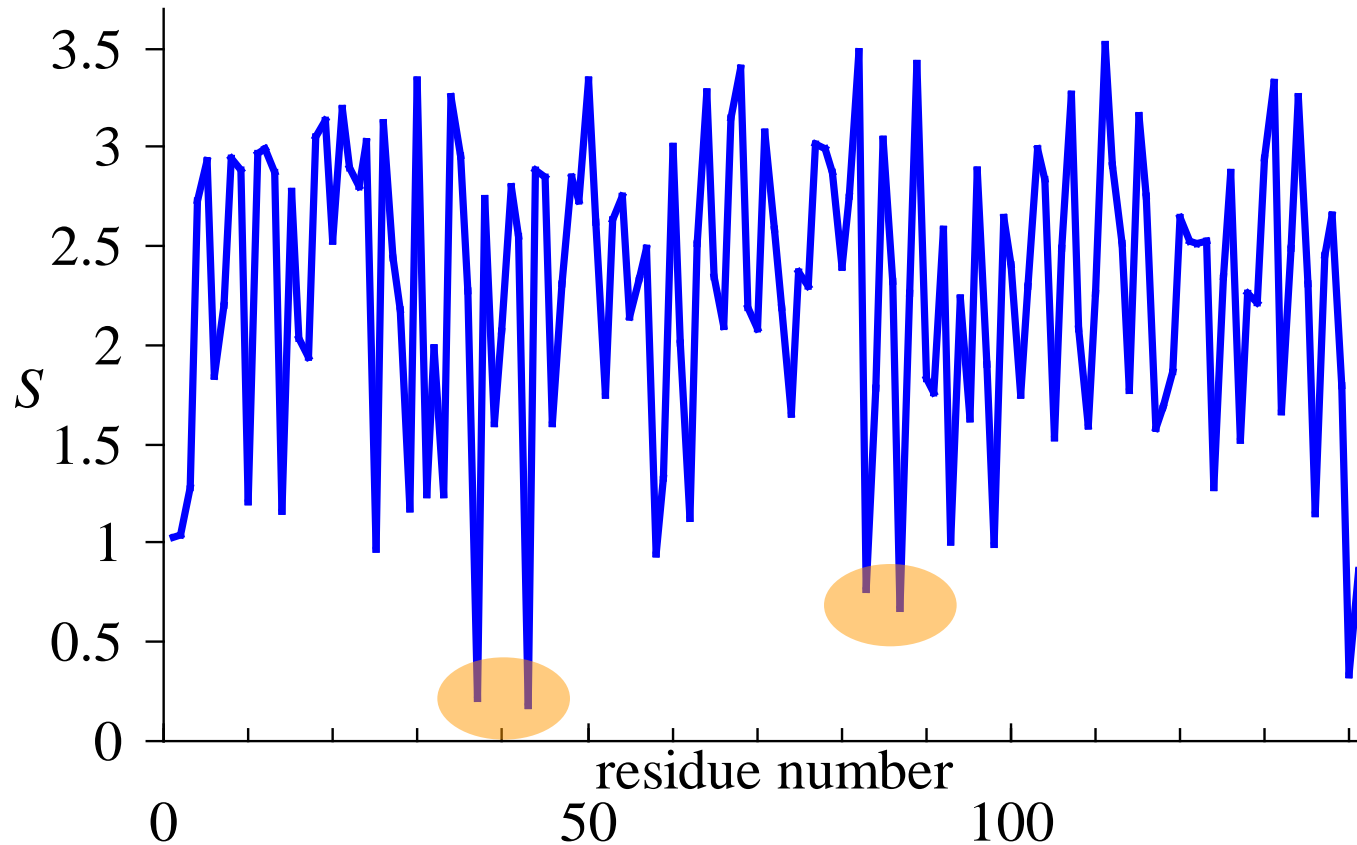
$$S = -4 \left(\frac{1}{4} \ln \frac{1}{4} \right)$$

$$= -\ln \frac{1}{4}$$

$$\approx 1.4$$

Haemoglobin conservation

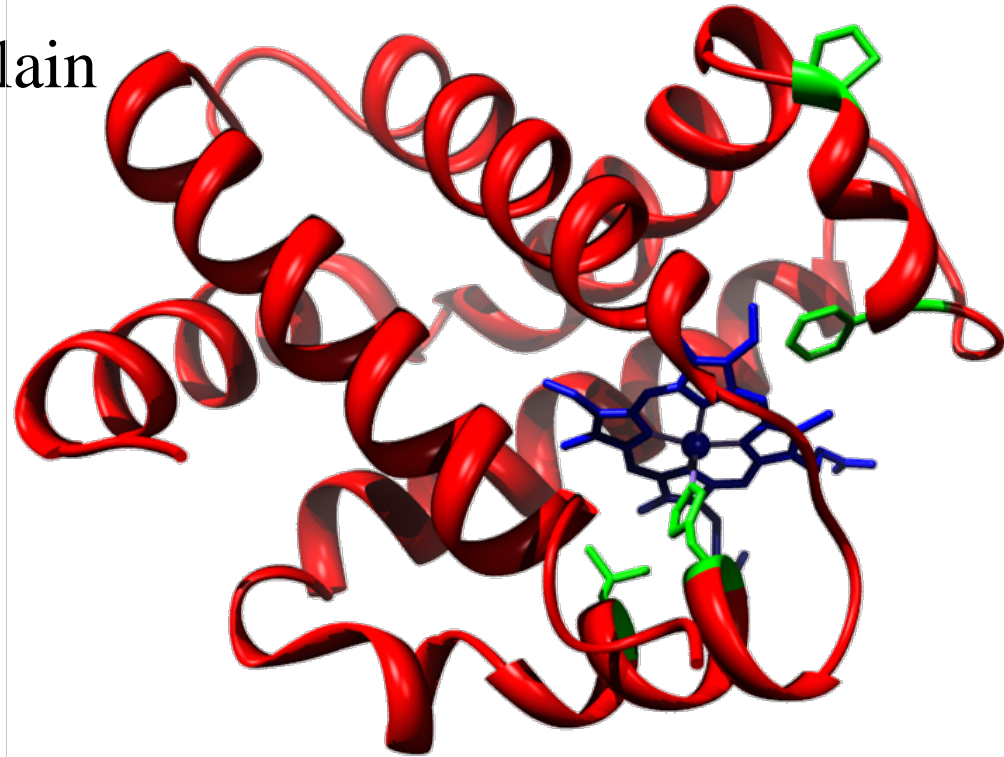
- look at residues 37, 43, 83 and 87



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

Conserved residues in haemoglobin

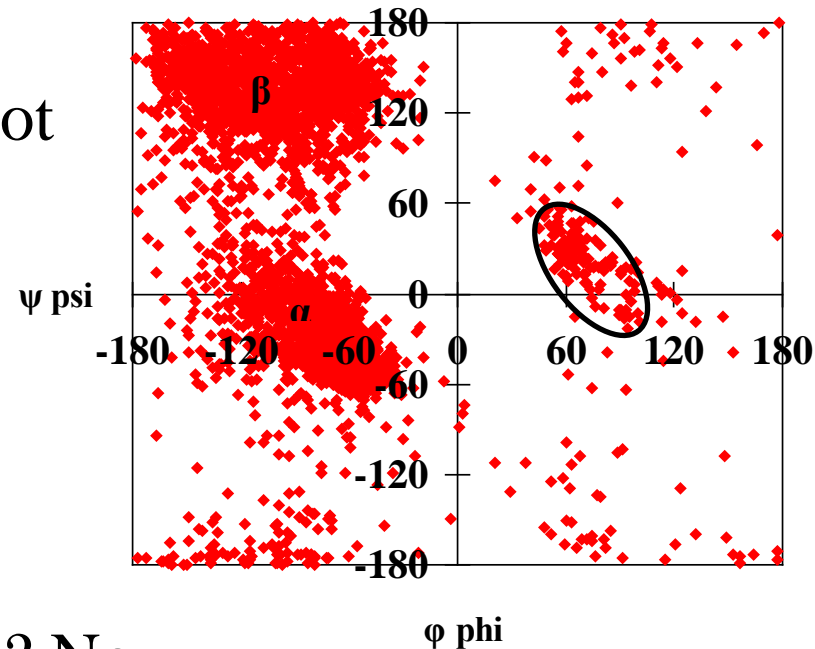
- 3 of the sites are easy to explain
 - interact with haem group
- Look at fourth site
 - proline
 - end of a helix



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

Conservation for structure

- some residues have very special structural roles
 - proline – not an H-bond donor
 - often end of a helix
 - glycine – can visit part of φ ψ plot
 - found in some turns
- are all gly residues so important ?
 - NO – they occur in many places
sometimes in turns
- are all pro residues very conserved ? No



Conservation for function

- in a serine protease
 - always a "catalytic serine"
 - can it mutate ? Not often
- in haemoglobin – residues necessary for binding haem
 - can they mutate ? rarely
 - changes properties of haemoglobin (bad news)
- dogma
 - residues in active site will be more conserved than other sites

Important summary

- conservation may reflect
 - important function
 - structural role
- mutagenesis / chemistry
 - what residue may I change to allow binding to a solid substrate ? (for biosensor/immobilized enzyme ?)
 - I want to try error prone PCR to select for new enzyme activity – which sites might I start with (active site) ?
- drug design example
 - target is an essential protein (basic metabolism, DNA synthesis, protein synthesis..)
 - is there some set of sequence features common to pathogen, different to mammalian protein ?

Evolution – do not trust conservation

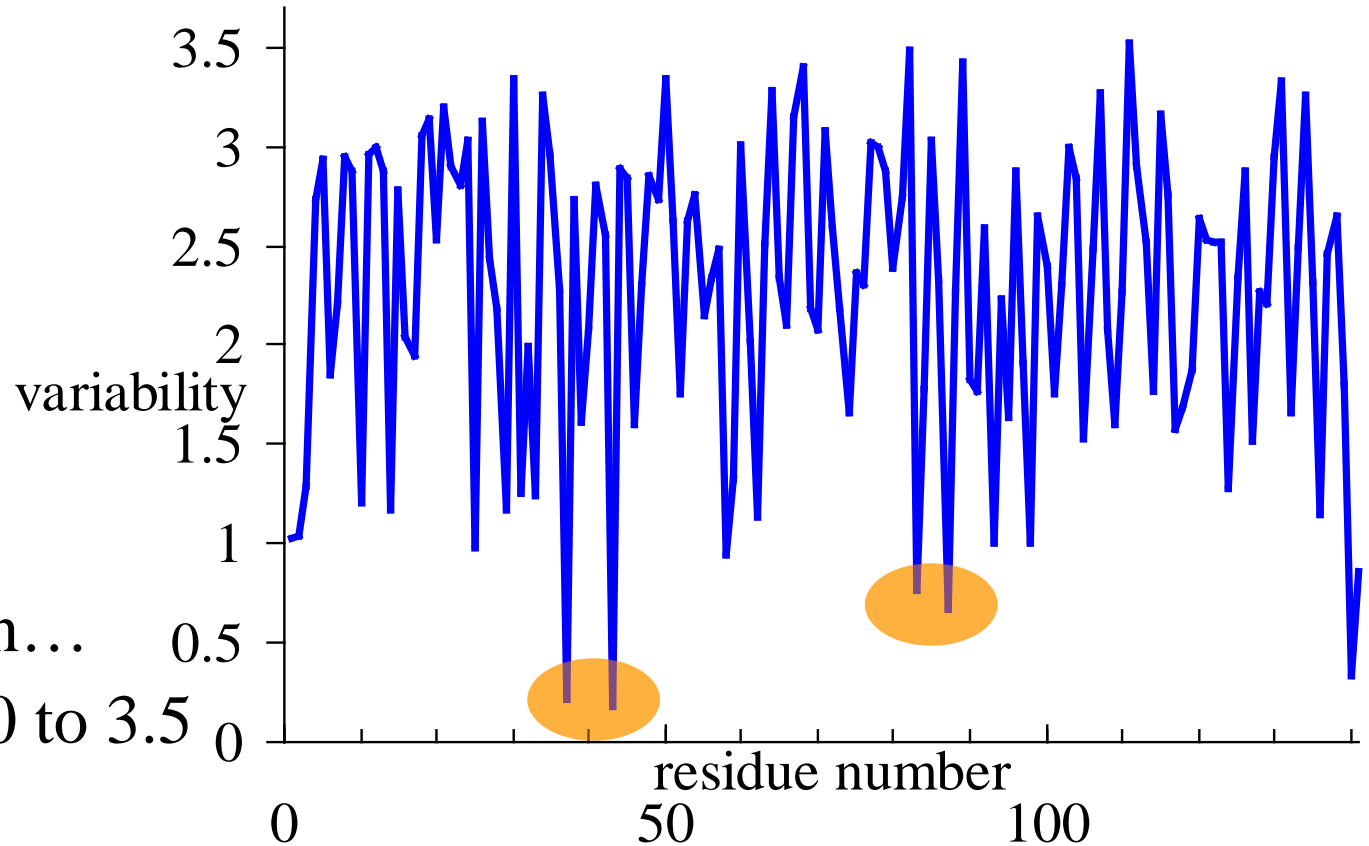
Imagine: two possible systems for some important enzyme

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

Conservation – how meaningful ?

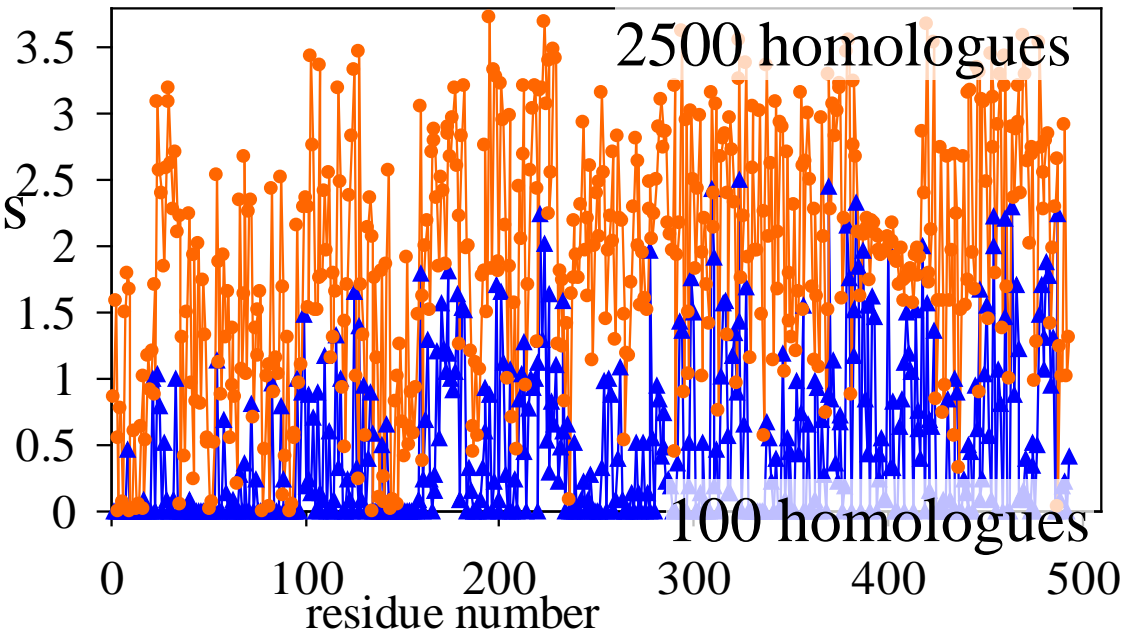


- Earlier Folien...
- values from 0 to 3.5
- what if I used more homologues ?

Conservation – how meaningful ?

- example sequence (1ab4, DNA gyrase)
- find 100 close homologues (mostly > 80% similarity)
– calculate conservation
- find 2 500 close homologues (mostly > 50 % similarity)
calculate conservation

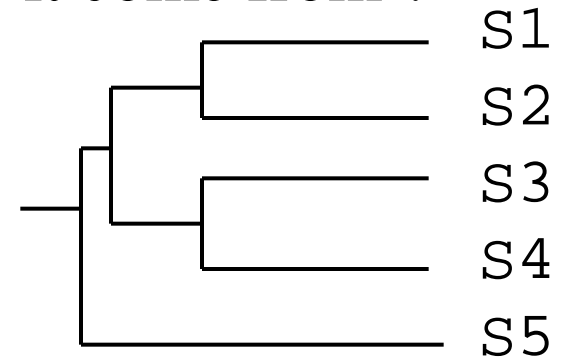
- fewer sequences
 - lots of conserved sites
- you can get the answer you want



Phylogeny / Evolution

Purely academic ? For fun ? Not always

- possibly useful in explaining disease propagation
 - where did HIV come from ?
 - where did the flu pandemics come from ?
 - virus infects banana crop – where did it come from ?
- previously we had a "guide tree"
 - did (S1,S2) and (S3,S4) share an ancestor but not S5 ?
 - not so good
- branch lengths do not reflect evolutionary time
- there may be other similar trees which could be evolutionary paths



Evolutionary time

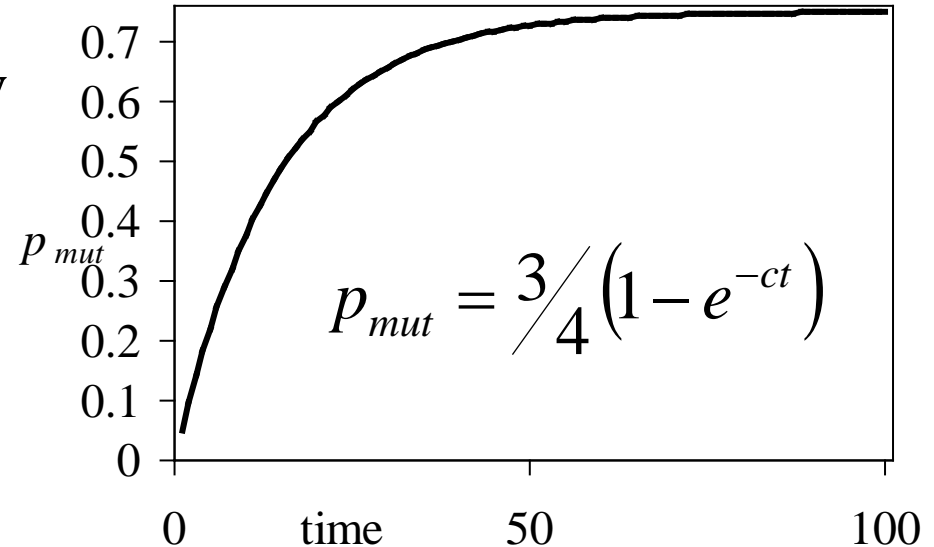
- compare two DNA sequences see
 - 1 mutation (represents time t)
 - 2 mutations (time $2t$)
 - 3 mutations (time $3t$)...
 - No !
- After some evolution
 - $A \rightarrow C \rightarrow G$ two events (although looks like $A \rightarrow G$)
 - $A \rightarrow C \rightarrow G \rightarrow C \rightarrow A$ looks like zero mutations
- If I have infinite time
 - all bases / residues equally likely
 - $p_{mut} = 3/4 = 0.75$ (DNA) or $p_{mut} = 19/20$ (protein)

Mutation probability

- time units are rather arbitrary
- how would I estimate time ?
(for DNA)

$$t \propto -\ln\left(1 - \frac{4}{3} p_{mut}\right)$$

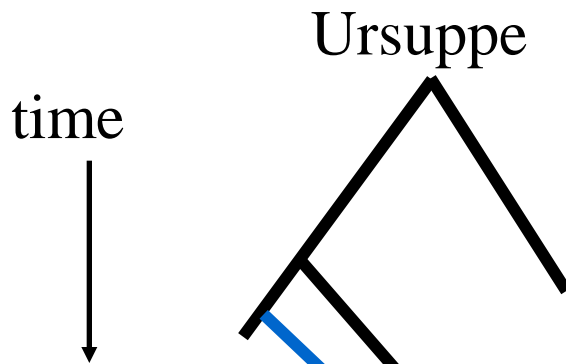
- p_{mut} ? count n_{mut} / n_{res}
- scaling of t not so important (relative time)



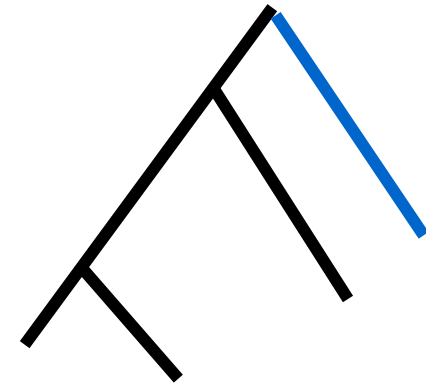
- for short times, p_{mut} changes fast
 - for small t , distances will be more reliable
 - as will be alignments
- is this enough for phylogeny ?
 - what about reliability ?

Problems in phylogeny

- not all sites mutate equally quickly
- not all species mutate equally quickly



but blue
species
(protein)
mutates
quickly



- blue appears to have branched off earlier

Problems estimating time

1. mutation rates vary wildly
 - changing environments – pH, temperature,..
 2. imagine time t is such that $p_{mut} = 0.25$
 - we have random events
 - sometimes you see 23% mutation, sometimes 28%
- time estimates will never be accurate
 - maybe we cannot find the correct tree
 - can we roughly estimate reliability ?

Reliability

- think of first alignment
- what would happen if you deleted a column ?
- if the data is robust /reliable
 - not much
- if the tree is very fragile /sensitive
 - tree will change
- better...

```
VLSPADKTNVKAAWGKVGAAHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VITP-EQSNVKAAWGKVGAAHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPGDKTQVQAGFGRVGAHAG--GAEAVDRMFLSFPTTKSFFPYFELTHGSAQVKGHG
VLSPAECTNIKAAWGKVGAAHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
-VTPGDKTNLQAGW-KIGAAHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPAECTNVKAAWGRVGAHAGDYGAELERMFLSFPTTKTYFPHFDLSHGSAQVQAHA
VLSPDDKTNVKAAWGKVGAAHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
```

Reliability

- repeat 10^2 to 10^3 times
 - delete 5 to 10 % of columns
 - copy random columns so as to have original size
 - recalculate tree
- how often did you see each branch

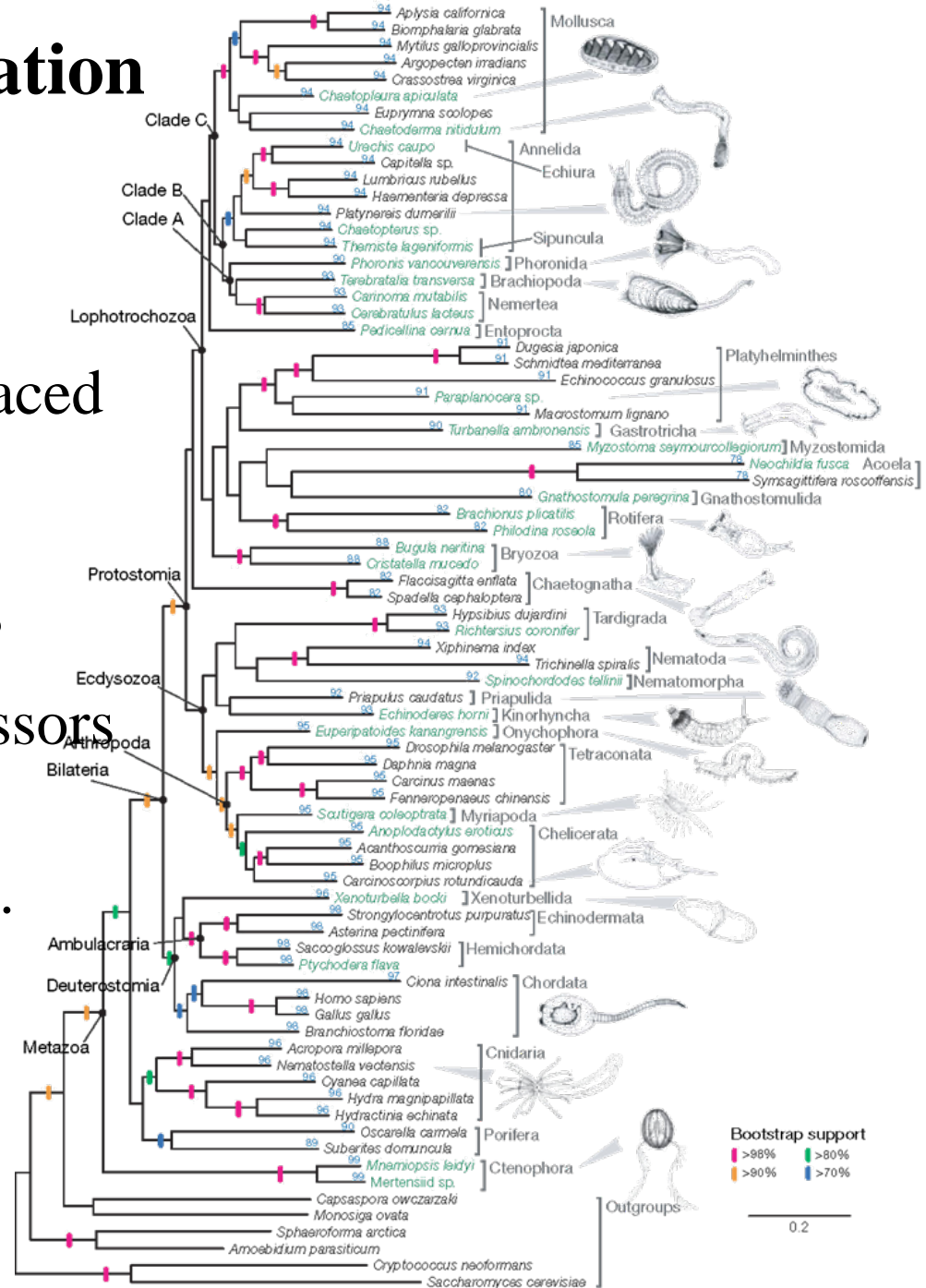


Monster example

- generate lots of trees
- for each subtree
 - see how often it is present
- example from cover of nature

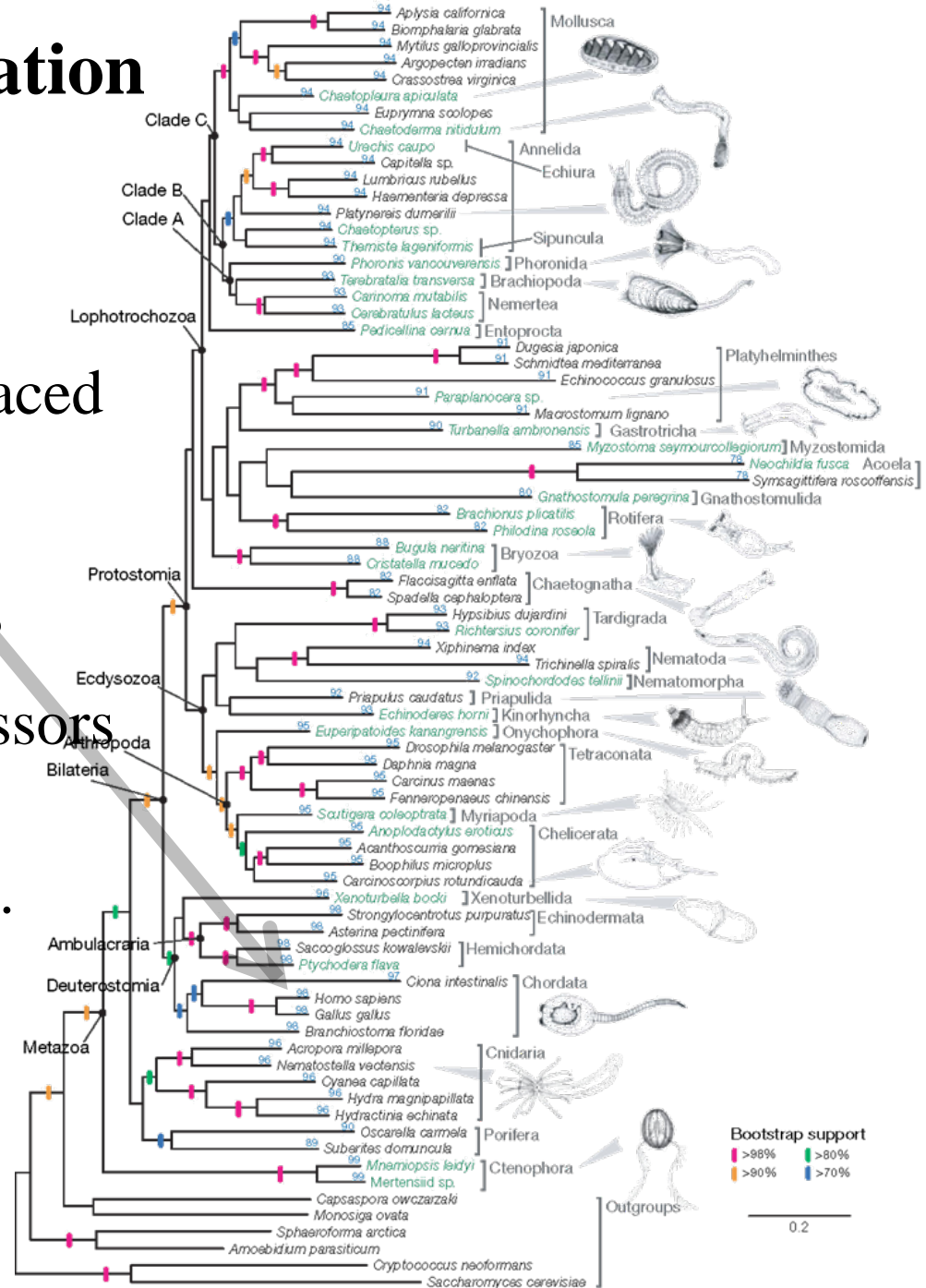
Monster calculation

- we are usually placed near Hühne
- we are not so reliably placed with little worms
- how long does this take?
 - months on 120 processors
- a more applied example..

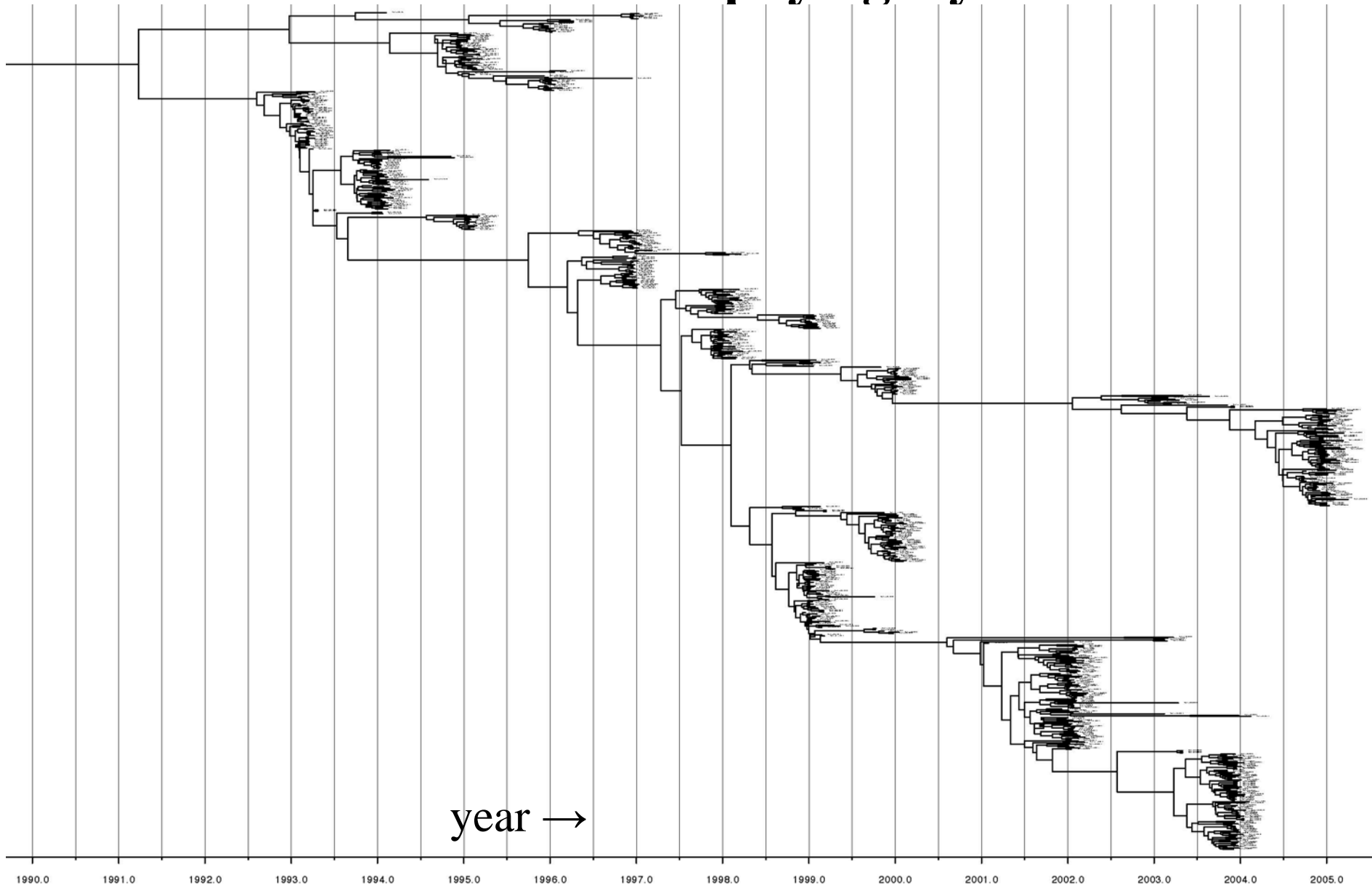


Monster calculation

- we are usually placed near Hühne
- we are not so reliably placed with little worms
- how long does this take?
 - months on 120 processors
- a more applied example..



Influenza virus phylogeny



Summary

- multiple sequence alignment – conservation
 - find important residues (function or structure)
 - can quantify conservation
- relations between most similar proteins are most reliable
- best tree is never found
 - too difficult algorithmically
 - lots of errors – evolution is a random process
- rough idea of reliability
- quick tree – possible for hundreds of sequences
- more complicated methods – only practical for smaller numbers of sequences