Bücher

- Hütt-Dehnert, Methoden der Bioinformatik (eine Einführung)
  - billig, mehrere Kopien in der Bibliothek
- Selzer, Angewandte Bioinformatik
  - minimal OK, mehrere Kopien in der Bibliothek

- Nicht so viel Hilfe für die zweite Hälfte des Semesters

Prüfungen

- Beispielfrage bald

6 weeks of me

- Done
  - similarities and alignments

- Coming
  - multiple alignments - evolutionary emphasis
  - comparing protein structures - not sequences

Bis jetzt

- Man hat eine Sequenz (Protein oder Nukleotid)
- Man will so viel wie möglich finden um
  - Struktur vorherzusagen
  - Funktion vorherzusagen
- Erinnerung

Erinnerung

- warum braucht man Ähnlichkeiten ?
- Ähnlichkeiten auf dem Sequenz-Niveau
  - wie man sie findet
  - Alignments
- genaue versus schnelle Methoden
- Bewertungsmethoden
- entfernt Homologen
- Signifikanz
- Protein modellierung

- Jetzt multiple Alignments

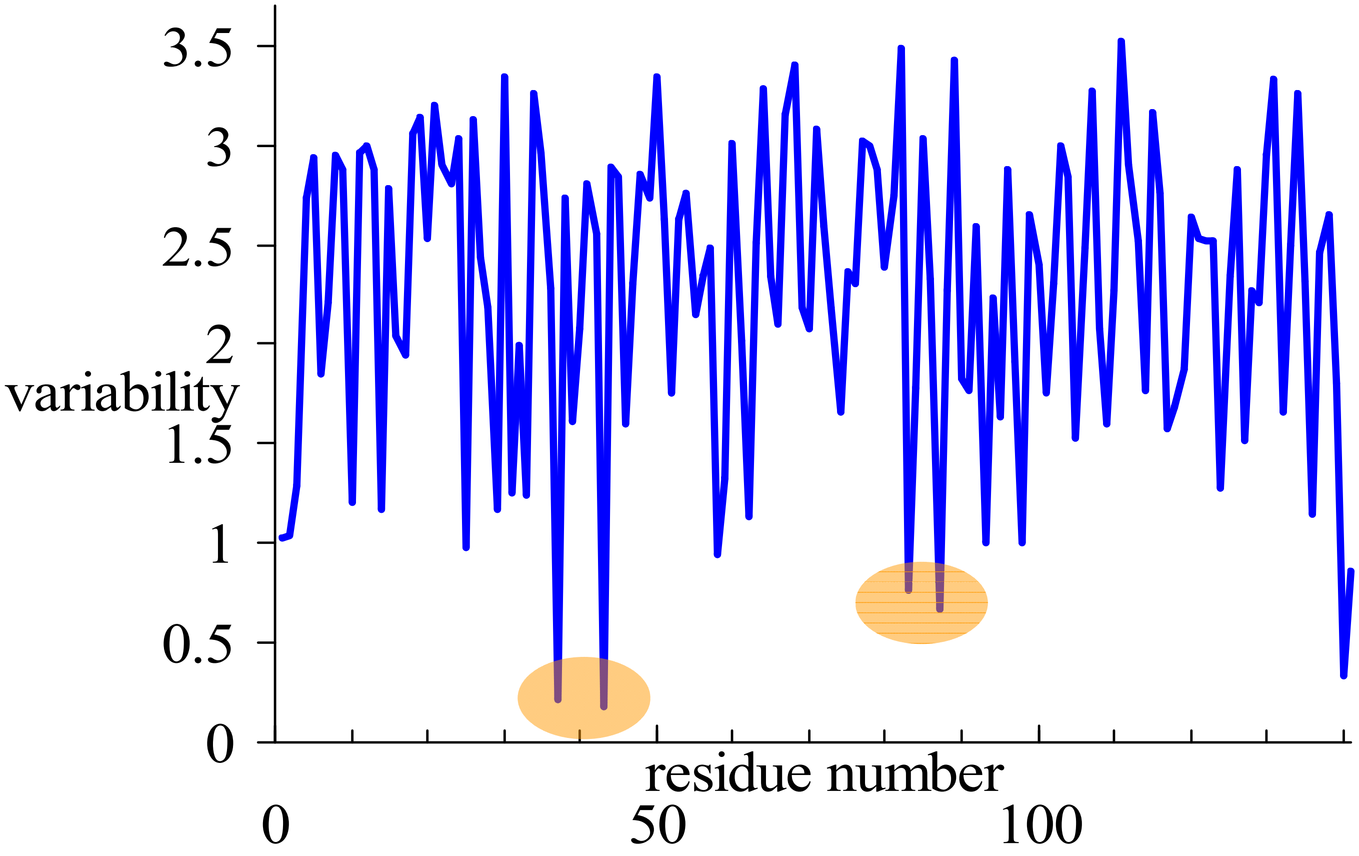
Multiple alignments

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

```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALEKMFLSFPTTKTYFPHFDLSHGSAQVKGHG
 LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
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
    …       …        … …
```

Conservation / variability

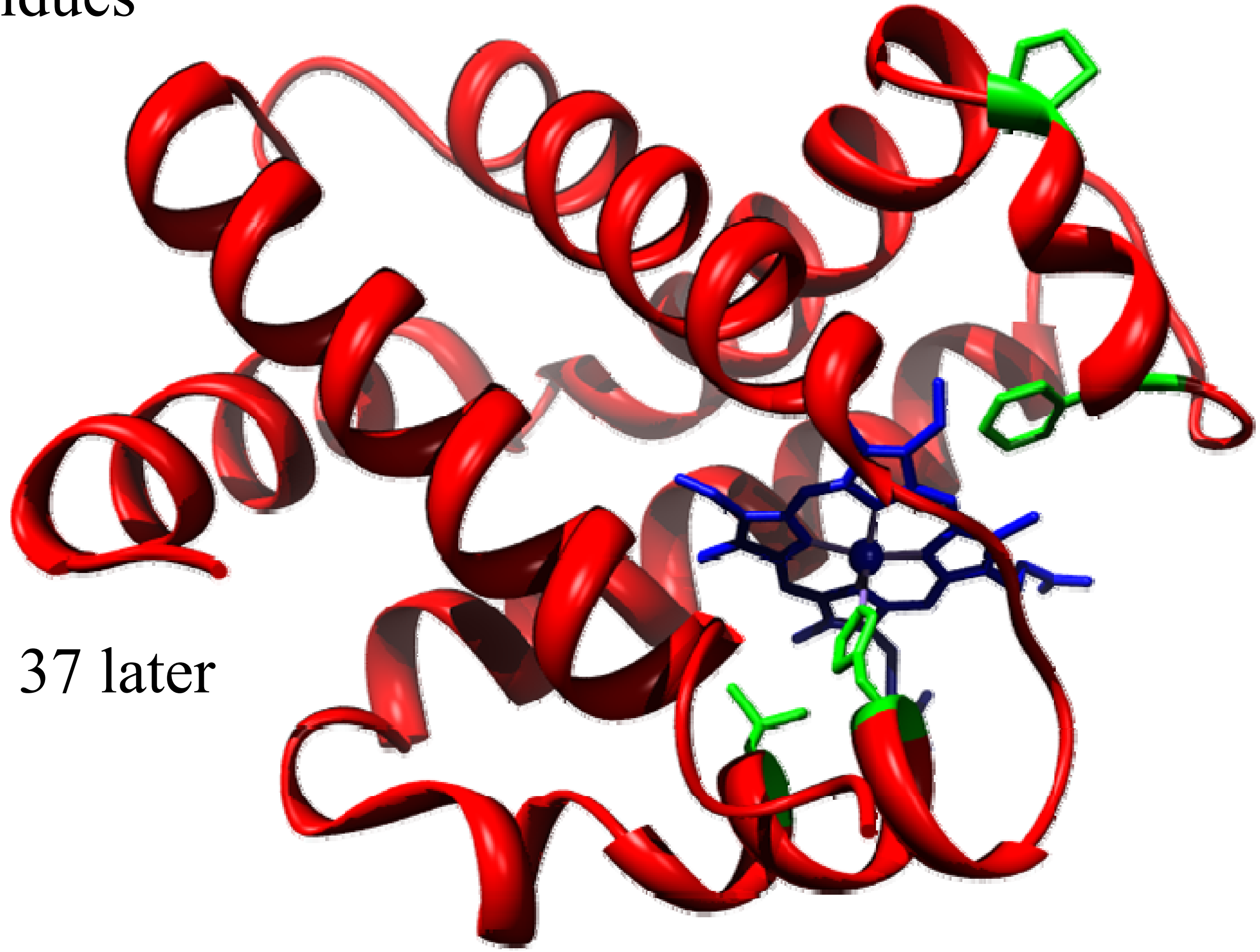
- look at residues 37, 43, 83 and 87



- 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 vary 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
  - …

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

- reasonable starting point
  - begin with pairs of proteins

Scoring schemes

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

```
VLSPADKSNVKAGWGQVGAHAGDYGAEAIERMYLSFPSTKTYFPHTDISHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
```

- In pairwise problem
  - 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 VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
2 VITP-EQSNVKAAWGKVGAHAGEYGAEALEQMFLSYPTTKTYFP-FDLSHGSAQIKGHG
3 MLSPGDKTQVQAGFGRVGAHAG--GAEALDRMFLSFPTTKSFFPYFELTHGSAQVKGHG
4 VLSPAEKTNIKAAWGKVGAHAGEYGAEALEKMF-SYPSTKTYFPHFDISHATAQ-KGHG
5 –VTPGDKTNLQAGW-KIGAHAGEYGAEALDRMFLSFPTTK-YFPHYNLSHGSAQVKGHG
6 VLSPAEKTNVKAAWGRVGAHAGDYGAEALERMFLSFPSTQTYFPHFDLS-GSAQVQAHA
7 VLSPDDKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
```

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

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

Aligning average sequences

```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VITPAEKTNVKAAWGKVGAHAGEYGAEALEQMFLSYPTTKTYFPHFDLSHGSAQIKGHG
```

and

```
IITPGDKTNVKAAFGKVGAHGGEYGAEALDRMFISFPSTKTYYPHFDLSHASAQVKAHG
VITPAEQTNIKGAWGQIGAHAGDYAADALEQMFLSYPTSKTYFPYFDLTHGSAQIKGHG
VITPAEKTQVKAAWGKVGGHAGEYGAEAIEQMFLTYPTTQTYFPHFELSHGTAQIKGHG
```

- at each position
  - use some kind of average in scoring
  - if a column has 2×D and 1×E score
    - score as D (cheating but fast)
    - score as 2/3 D + 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

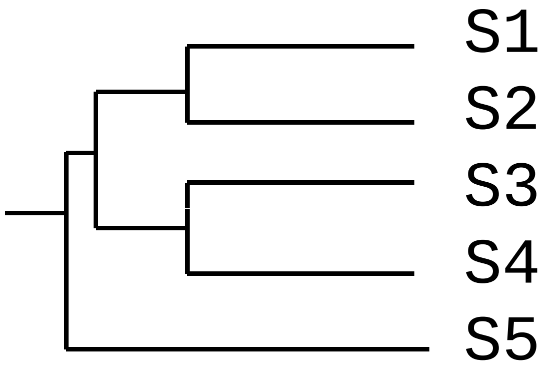
Progressive alignment - tree

S1	ATCTCGAGA
S2	ATCCGAGA
S3	ATGTCGACGA
S4	ATGTCGACAGA
S5	ATTCAACGA

Compute pairwise alignments, calculate the distance matrix

	S1	S2	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

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
```

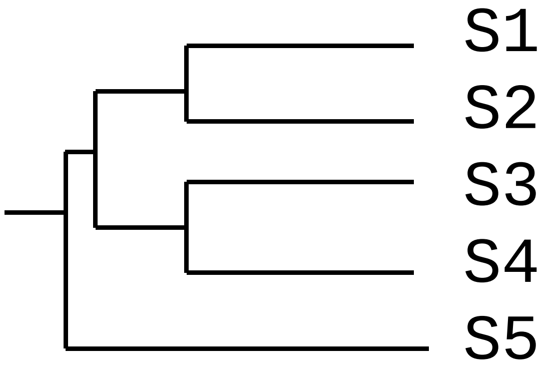
align av(S1,S2,S3,S4) with S5

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

- 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

Problems and variations

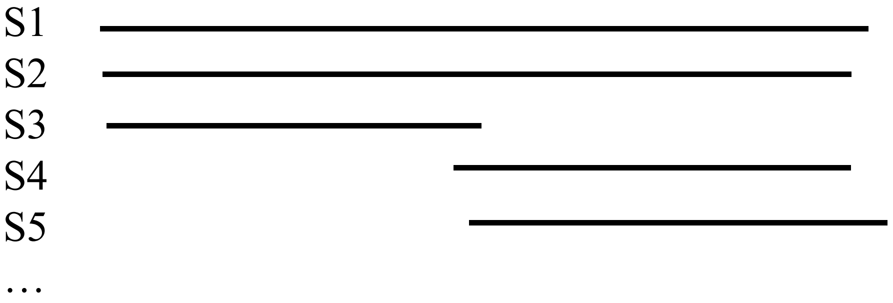


	S1	S2	S3	S4	S5
S1	-				
S2	.11	-			
S3	.20	.30	-		
S4	.27	.36	.09	-	
S5	.30	.33	.23	.27	-

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 ?

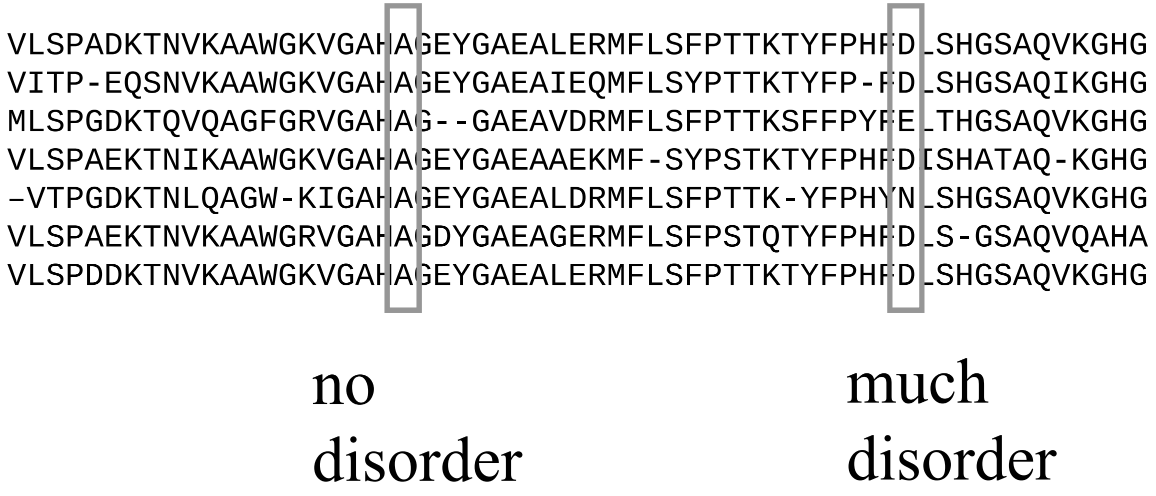
- more later

Measuring conservation / entropy

- Gibbs entropy
  - 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
```

no disorder

much disorder

- Calculate an "entropy" for each column

Entropy

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

- We can forget *k* (Boltzmann – just scaling)
- 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

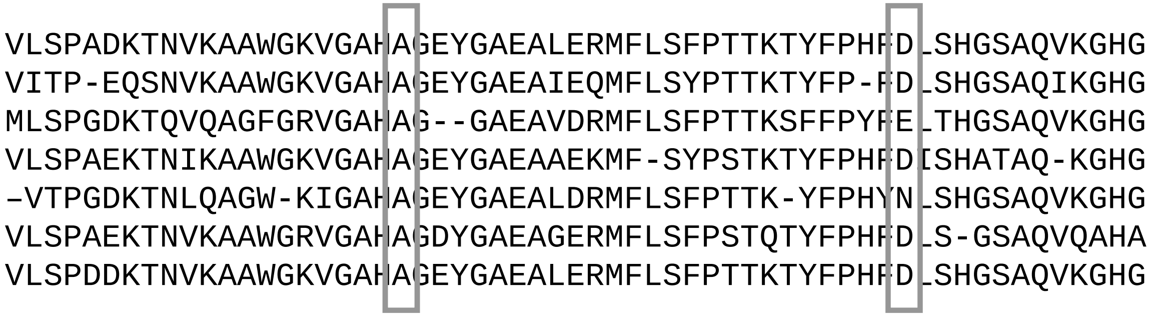
$$S = -\sum_{i=1}^{20} \frac{1}{20} \ln \frac{1}{20} = -20 \cdot \frac{1}{20} \ln \frac{1}{20}$$

$$\approx 3$$

- my toy alignment..

Entropy

- first column is boring
- second
  - $p_{\mathrm{D}} = 5/7$
  - $p_{\mathrm{E}} = 1/7$
  - $p_{\mathrm{N}} = 1/7$



```
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

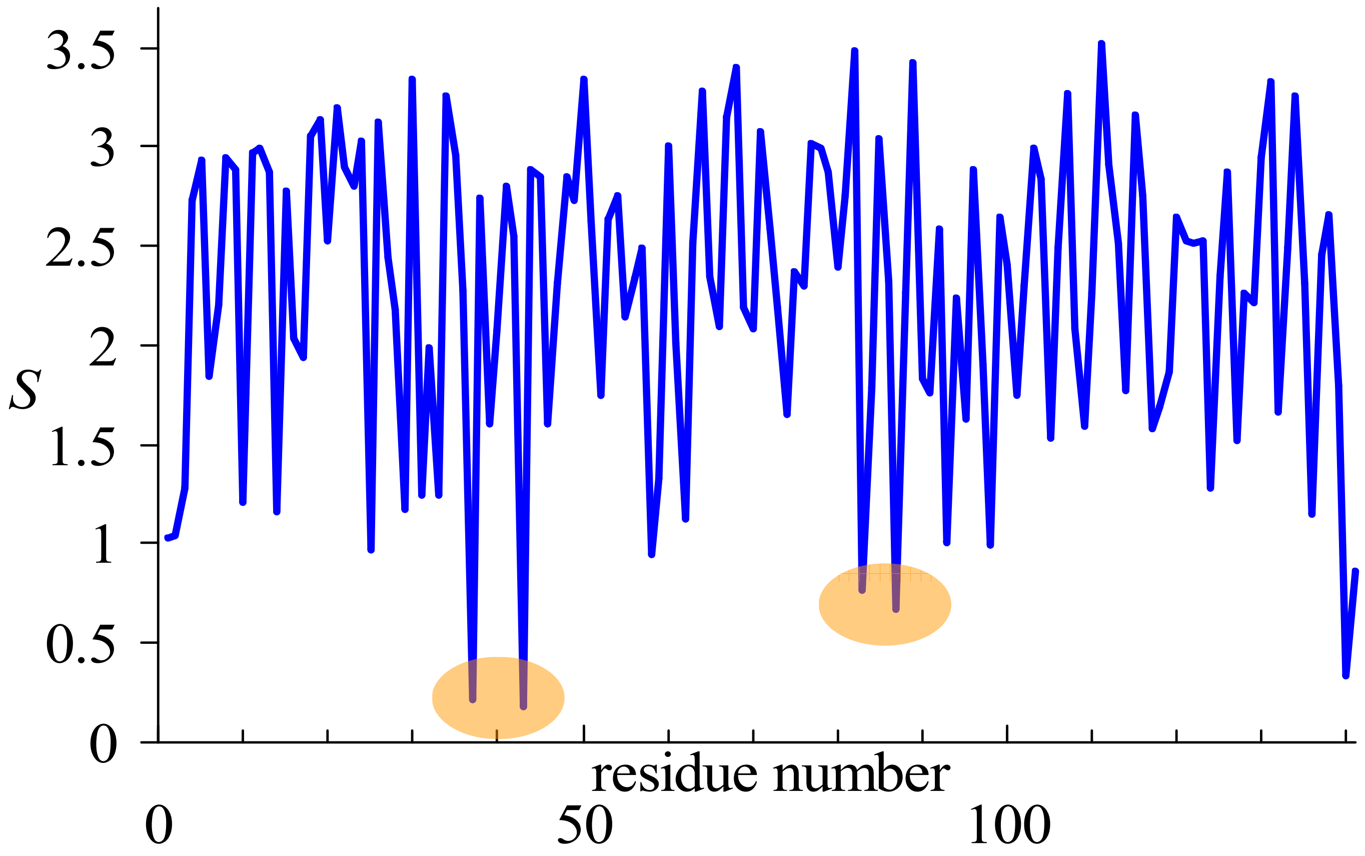
Entropy from DNA

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

$$\begin{aligned} S &= -4\left(\frac{1}{4}\ln\frac{1}{4}\right) \\ &= -\ln\frac{1}{4} \\ &\approx 1.4 \end{aligned}$$

Haemoglobin conservation

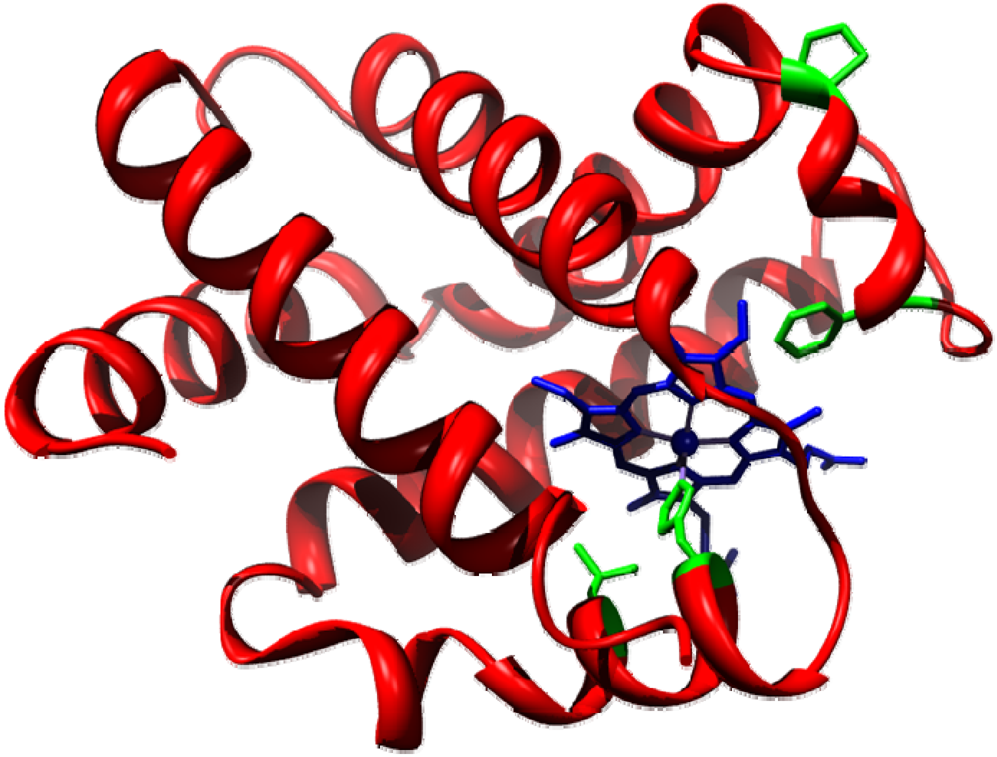
- look at residues 37, 43, 83 and 87



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

Conserved residues in haemoglobin

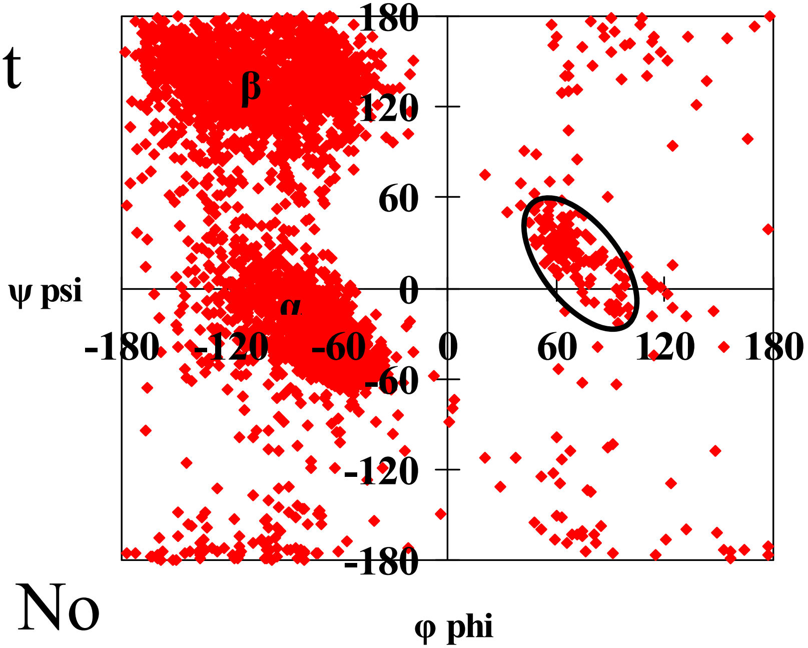
- 3 of the sites are easy to explain
  - interact with haem group
- Look at fourth site
  - proline
  - end of a helix



- 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

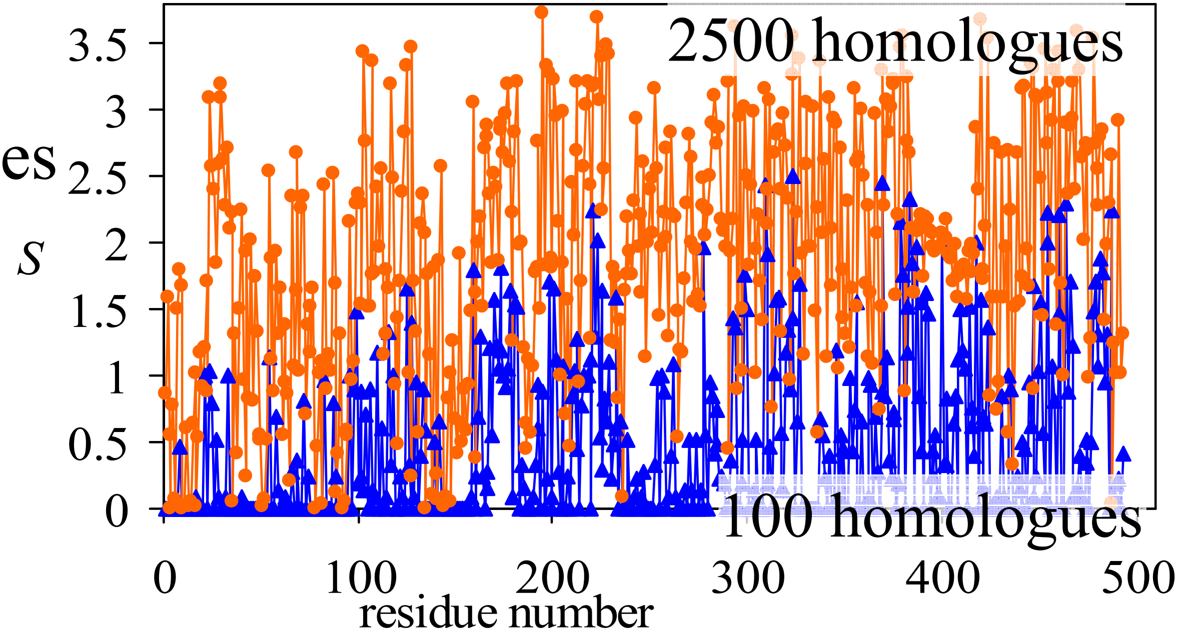
If you have the choice, where would you evolve to ?

1. very fragile
2. likely to survive mutations

Conservation – how meaningful ?

- example sequence (1ab4, DNA gyrase)
- find 100 close homologues (mostly > 80% similarity) – calculate conservation
- find 2500 close homologues (mostly > 50 % similarity) calculate conservation

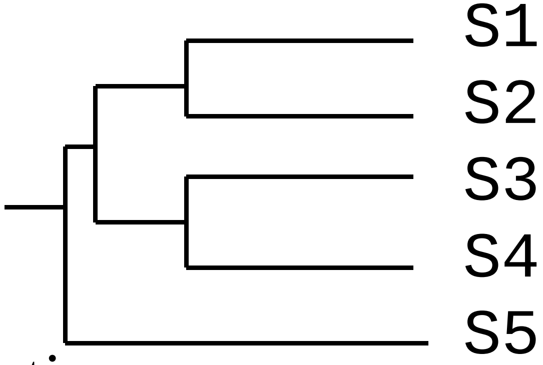
- fewer sequences
  - lots of conserved sites
- you can get the answer you want



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

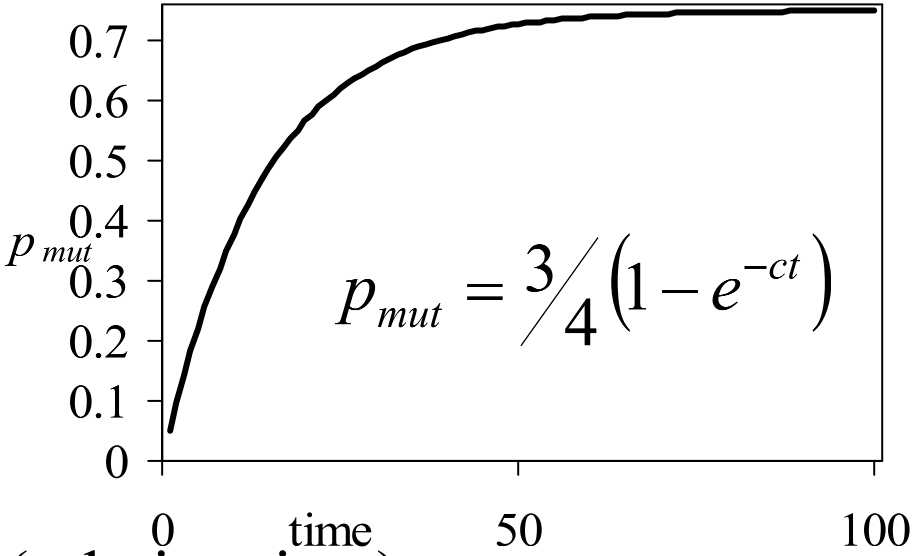
Evolutionary time

- compare two DNA sequences see
  - 1 mutation (represents time *t*)
  - 2 mutations (time 2*t*)
  - 3 mutations (time 3*t*)…
  - No !
- After some evolution
  - `A` → `C` → `G` two events (although looks like `A`→`G`)
  - `A` → `C` → `G` → `C` → `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

Mutation probability

- time units are rather arbitrary
- how would I estimate time ?

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



$$p_{mut} = \tfrac{3}{4}\left(1 - e^{-ct}\right)$$

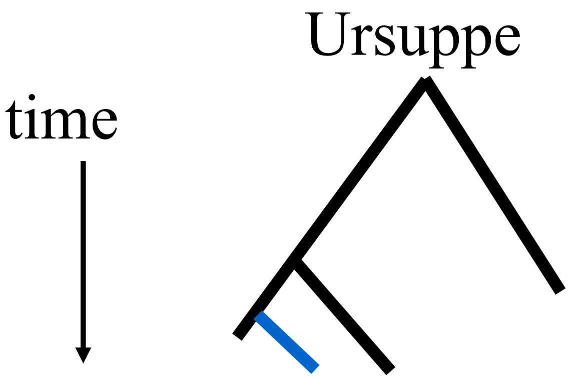
- p_{mut} ? count n_{mut} / n_{res}
- scaling of *t* not so important (relative time)

- for short times, p_{mut} changes fast
  - for small *t*, distances will be more reliable
    - as will be alignments

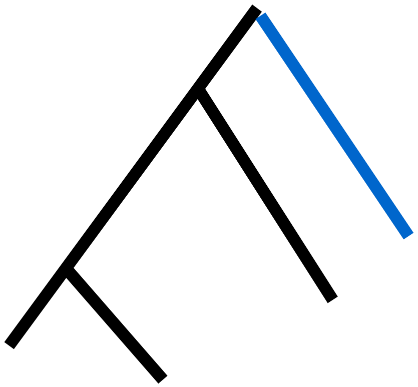
- is this enough for phylogeny ?
  - what about reliability ?

Problems in phylogeny

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



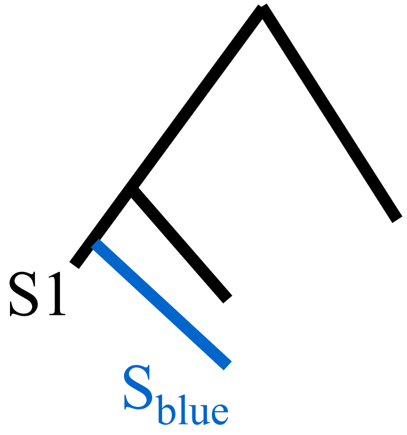
but blue species (protein) mutates quickly



- blue appears to have branched off earlier
- less drastic..

Problems in trees

- blue evolves a bit faster
- when we make average sequences
    - av(S1, S_{blue}) and sub-tree seems further from other sequences
    - all nearby nodes will be distorted



Problems estimating time

- mutation rates vary wildly
  - changing environments – pH, temperature,..
- can the distances ever be accurate ?

- imagine time t is such that p_{mut}=0.25
  - we have random events
  - sometimes you see 23% mutation, sometime 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 ?

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

- if the data is robust /reliable
  - not much
- if the tree is very fragile /sensitive
  - tree will change

- better
  - repeat 10^2 to 10^3 times
    - delete 5 to 10 % of columns
    - copy random columns so as to have original size
    - recalculate tree

Monster example

- generate lots of trees
- for each subtree
  - see how often it is is present

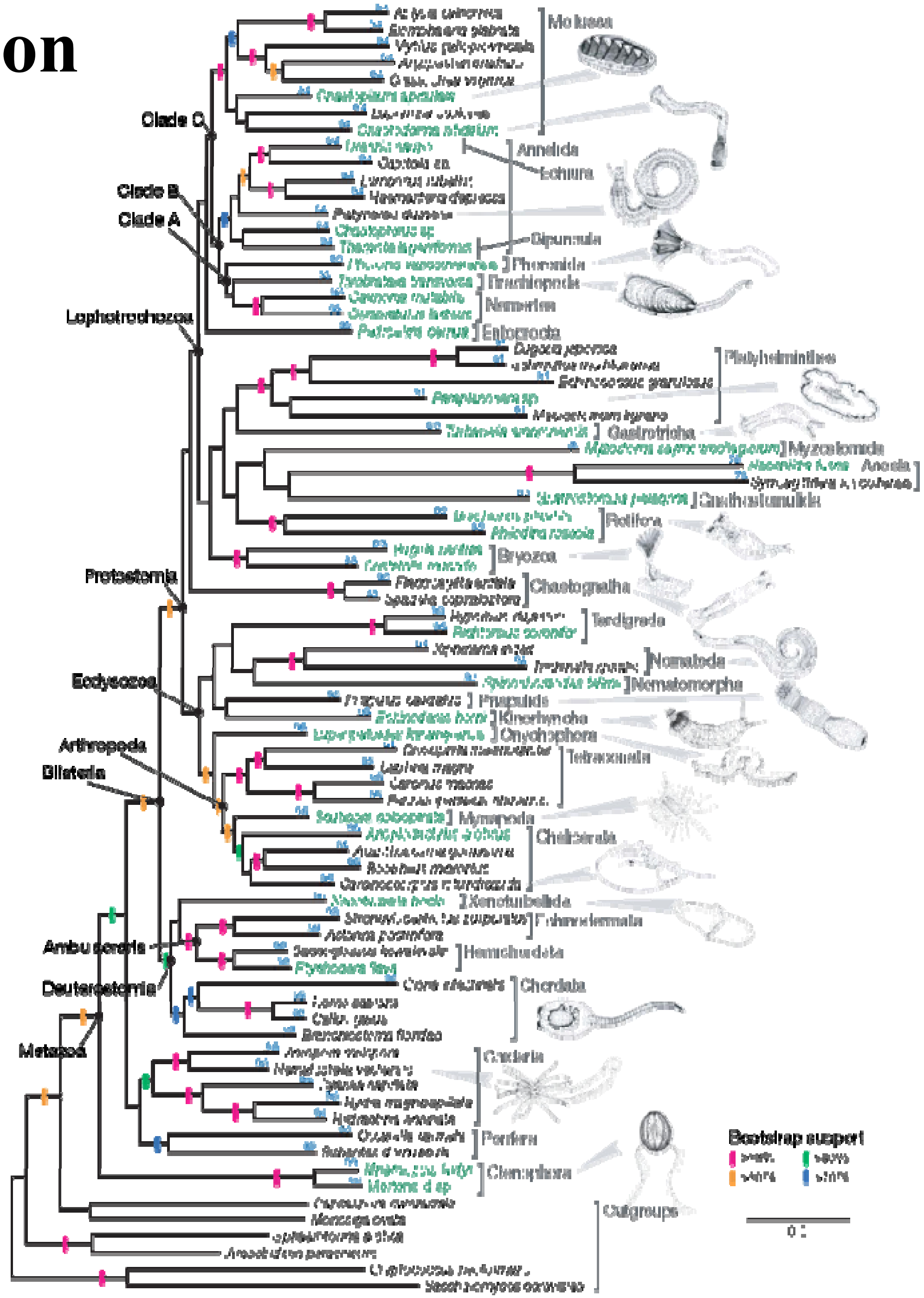
- example from cover of nature

Monster calculation

- we are usually placed near Hühne
- we are not so reliably placed with little worms

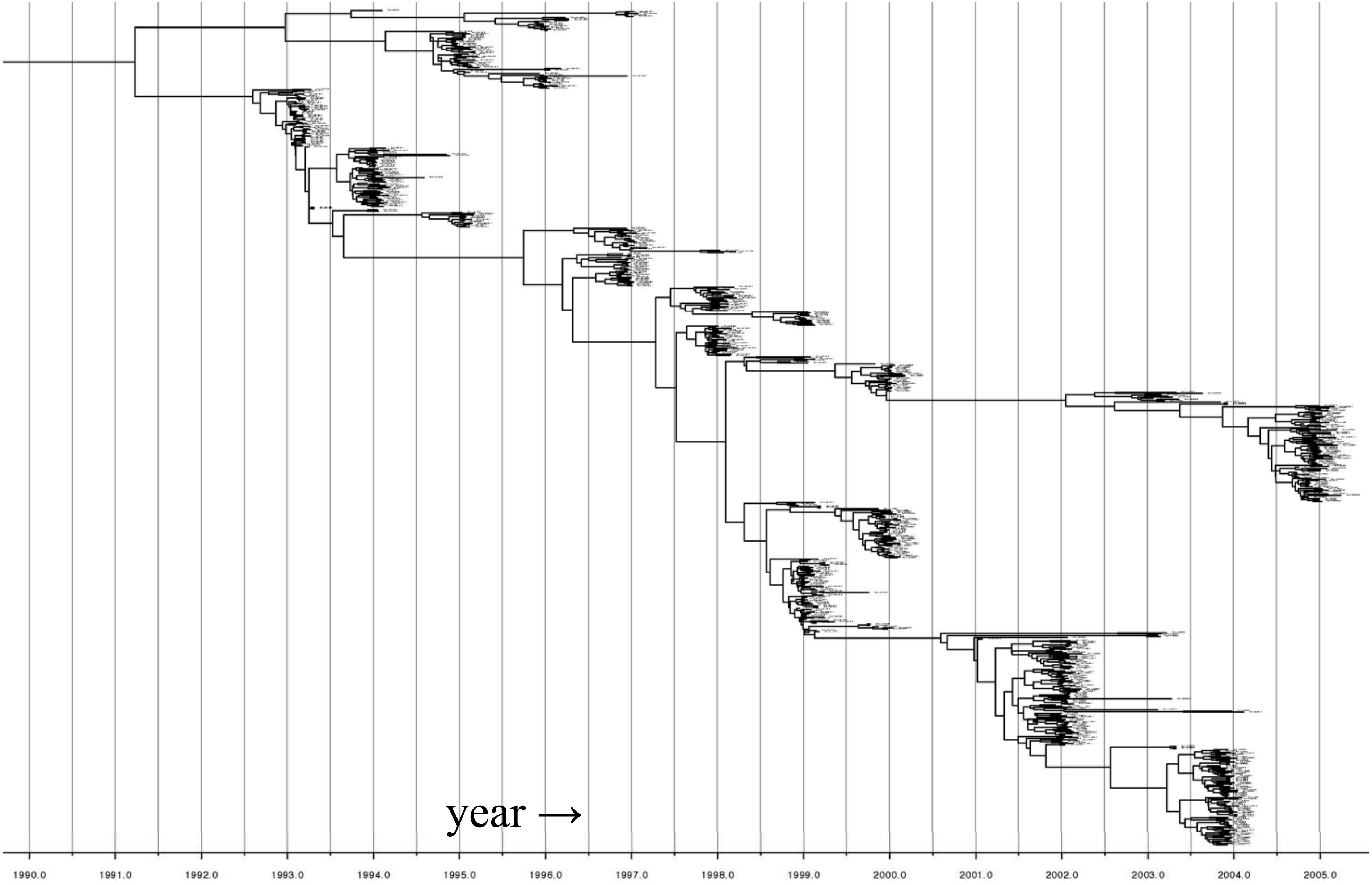
- how long does this take?
  - months on 120 processors

- a more applied example..



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

Influenza virus phylogeny



Rambaut, A., .. Holmes, C. The genomic.. influenza A virus, Nature 452, 1-6, 2008

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

Protein structures and comparisons

Ultimate aim

- how to find out the most about a protein
- what you can get from sequence and structure information

On the way..

- remote similarities between proteins
- sequence versus structural similarity
- Detour
  - protein coordinates – representation, accuracy
- measures for similarity of coordinates

- Later
  - classifications of proteins

Sequence and structure similarity

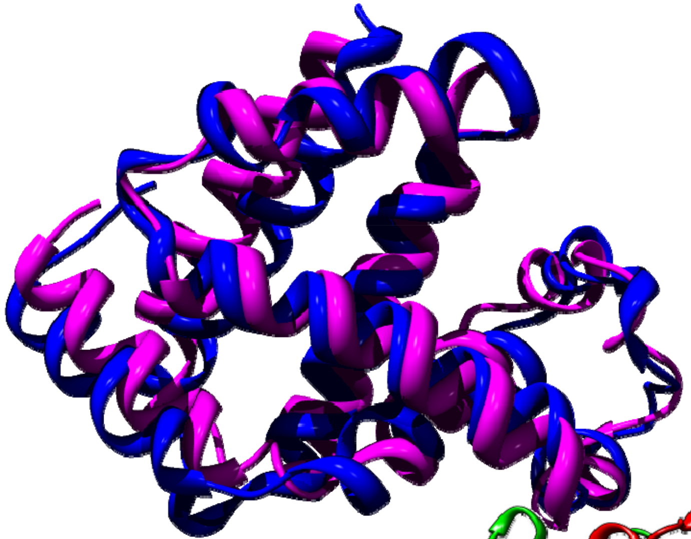
Claim from before

- if two sequences are similar – they are related – structures are similar

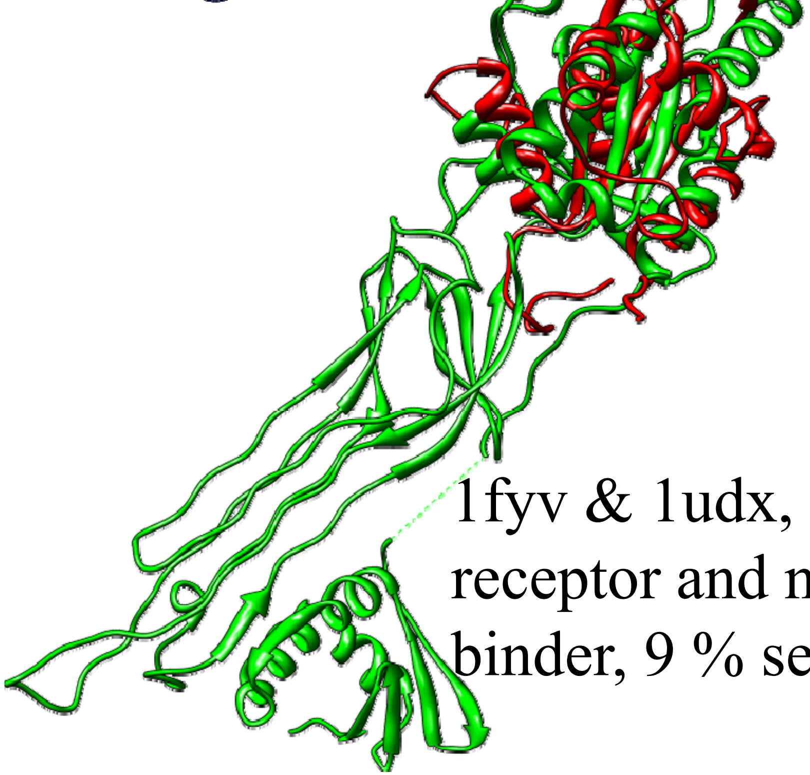
Question

- if two sequences are different - are their structures different ?

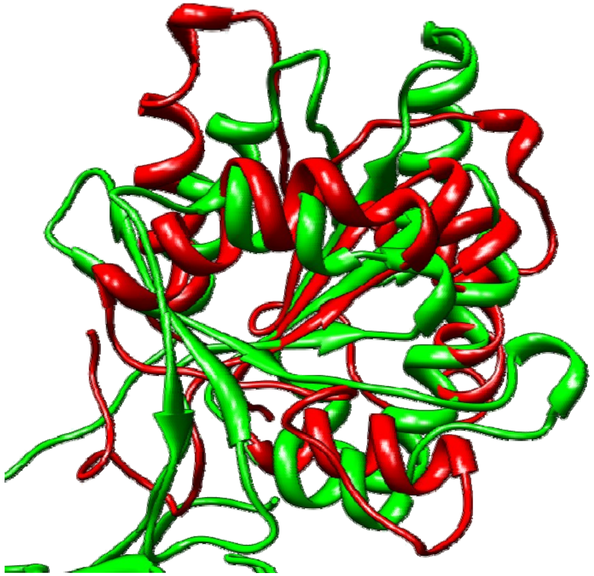
Remote similarities



1cbl & 1eca (haemoglobin & erythrocruorin)
14 % sequence id

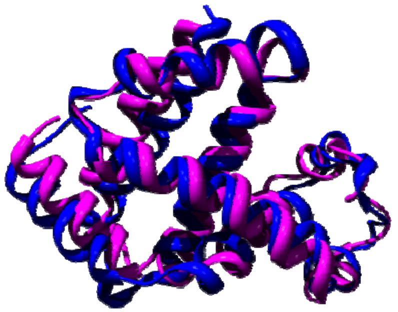


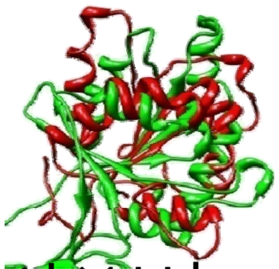
1fyv & 1udx, TLR receptor and nucleotide binder, 9 % sequence id



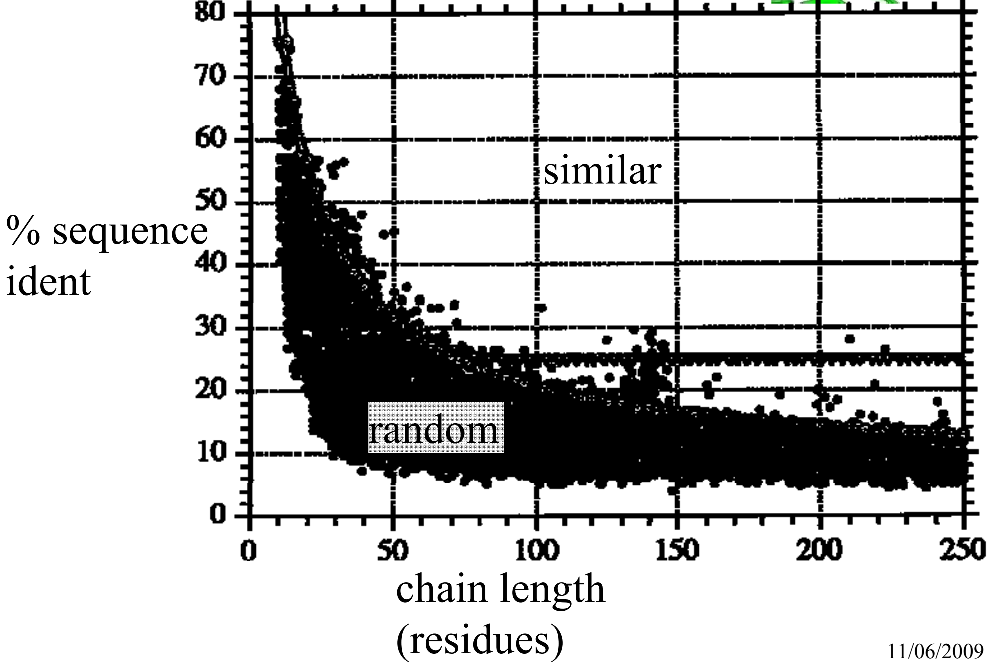
No sequence similarity – similar structures

- Are these rare ?
  - easy to find 100s of examples
- does this agree with previous claims ?
  - dot in diagram – two structures seem different





- if sequences are similar
  - structures will be similar
- if sequences are different
  - one does not know



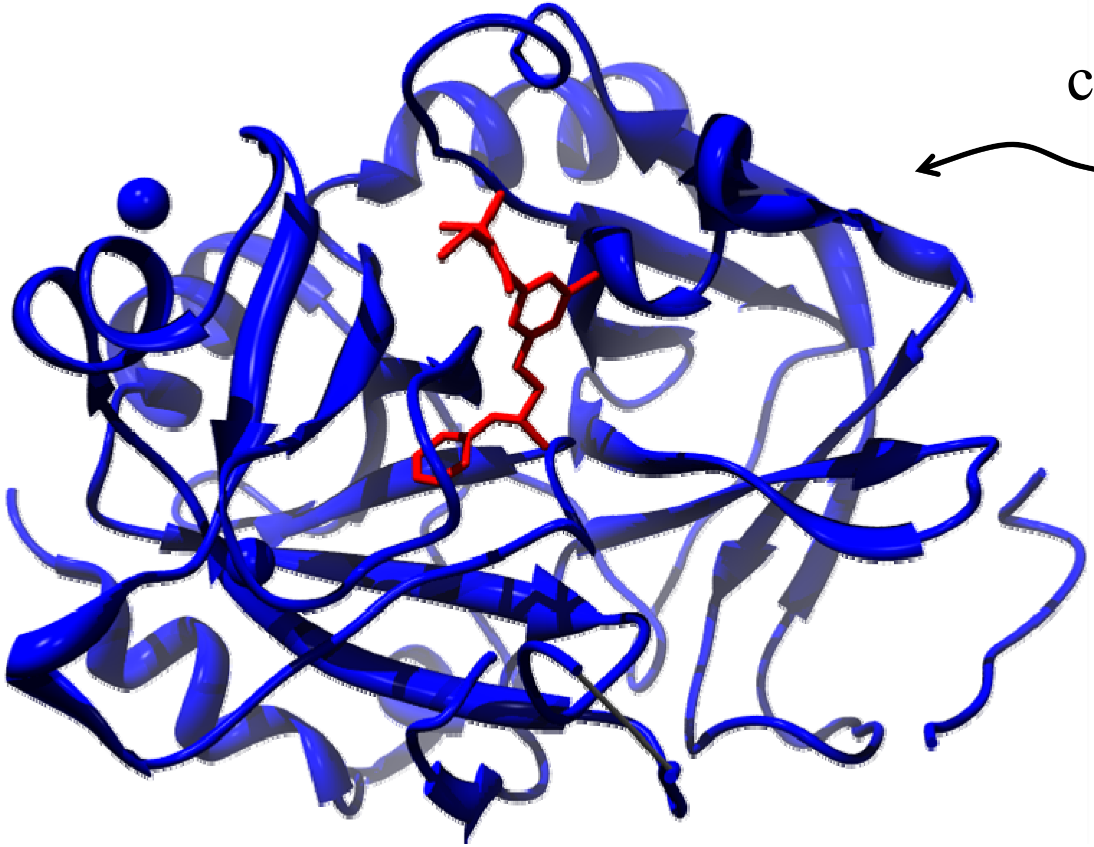
Rost, B.Prot Eng, 12,85–94, 1999

Structure versus sequence similarity

- Clear statement
  - sequence changes faster than structure
- Reason ? Unclear
  - possibility..

- protein function depends on having groups in orientation in space

Why can sequence change



change here

residue changes ? OK

structure changes ? Bad

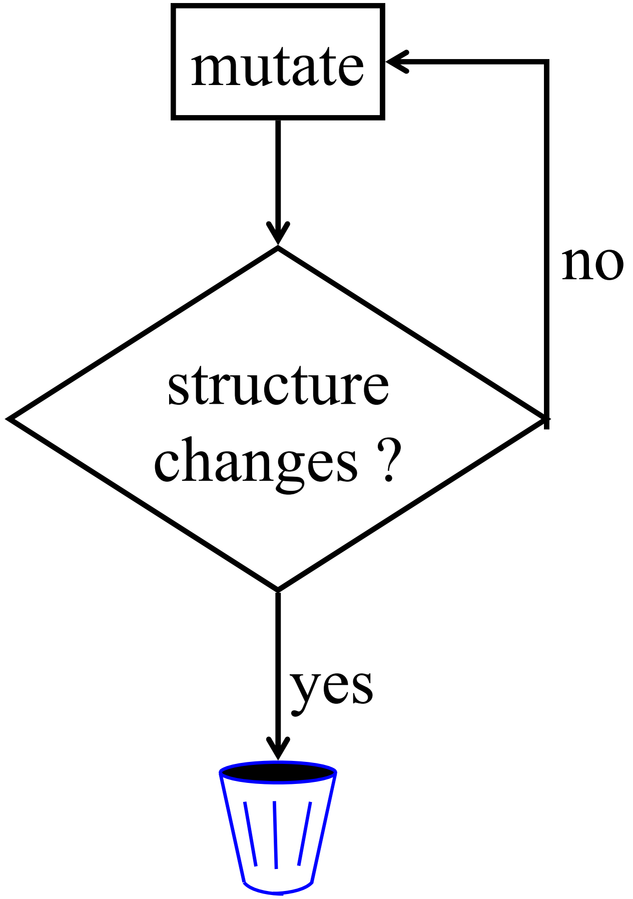
- a view of molecular evolution…

2j9m, 2cdk + aminopyridine

Simple view of molecular evolution

mutate continuously

- mutations which are not lethal
  - may be passed on (fixed)
- if structure changes
  - protein probably will not function
  - not passed on



Result

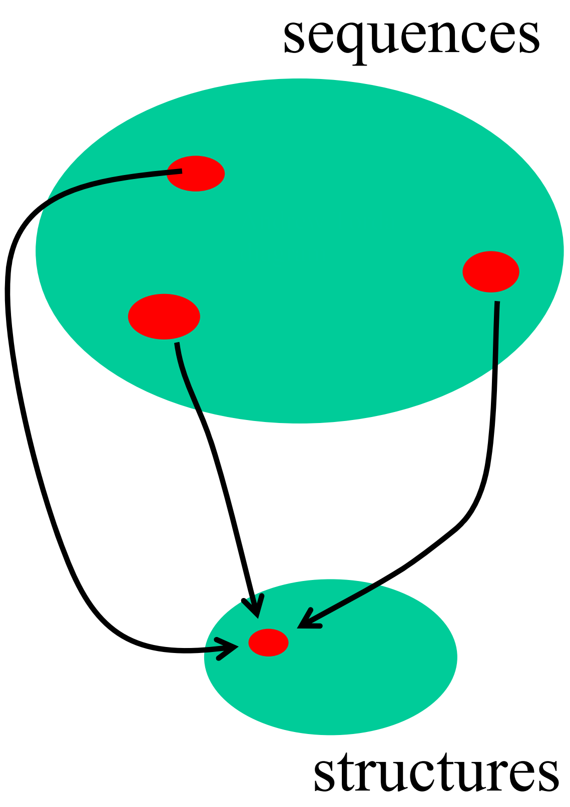
- evolution will find many sequences
  - compatible with structure
  - compatible with function
- how else would we see this ?

Sequence vs structure evolution

Sayings..

- Sequence and structure space
  - sequence space is larger
    - many different sequences map to similar structure
- sequence evolves faster than structure

- Truths…



Practical Consequences

Sequences of proteins are nearly always known

- similar sequence
  - usually similar structure, similar function
- sequences not (obviously) related
  - maybe similar structure
  - maybe similar function
- What if structures are known ?

Sequence and structure similarity

			structures	
			similar	different
sequence	similar	frequency	always	never
		function similar	yes	
	different	frequency	often	normal
		function similar	sometimes	no

- summarise from a different point of view

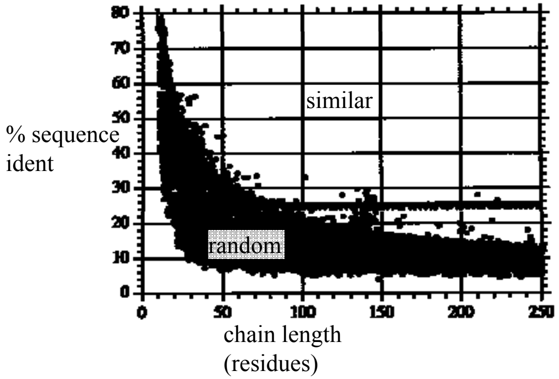
Sequence vs structure similarity

When comparing proteins

- more information is always better (sequence, structure,function)

Similar sequences

- structure and function will be similar
  - remember threshold graphs from earlier



Similar structures, different sequences

- evolutionary relationship implied but
  - bigger evolutionary distance
- not enough to be confident about function

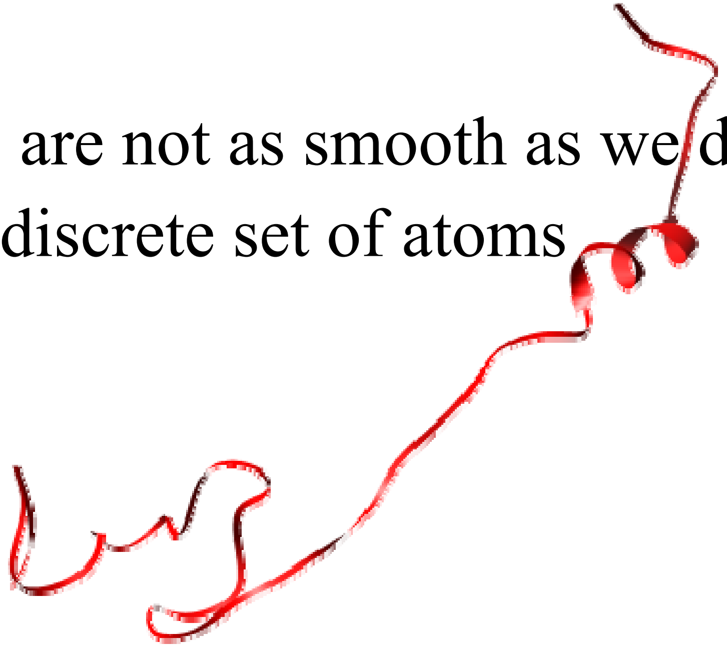
- what do we mean by similar structures ?

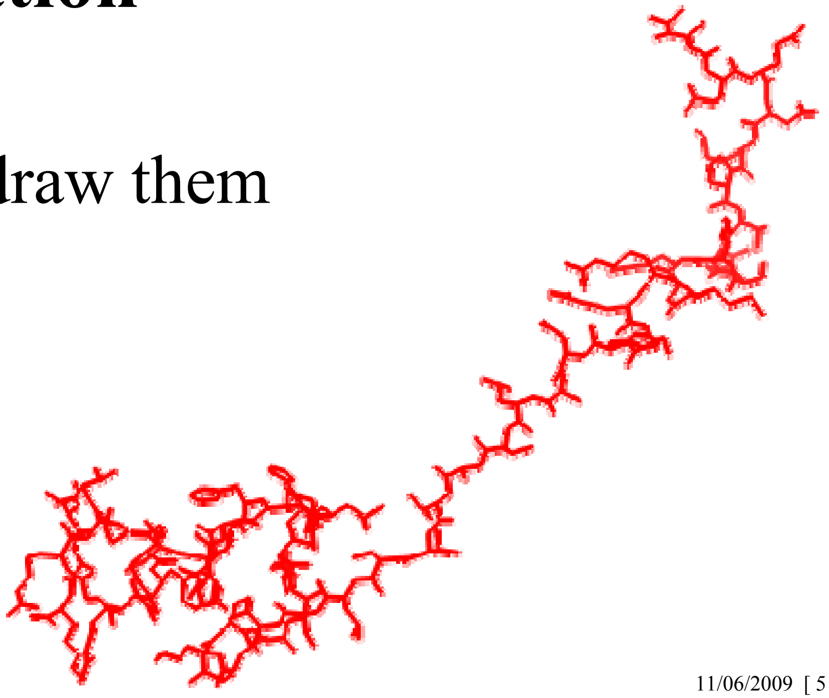
Comparing proteins

- Representation of proteins
- comparison
- classification (later)

Representation

- Proteins are not as smooth as we draw them
  - very discrete set of atoms





2vhm chain 0

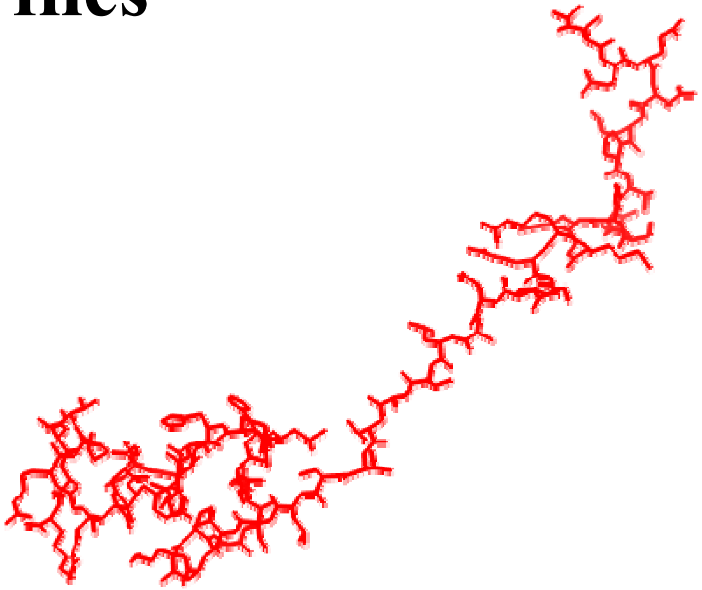
Protein coordinate files

Detour - Protein data bank (www.rcsb.org)

- only significant database of protein coordinates
- deposition of coordinates – often requirement of publication
- $\approx 60 \times 10^3$ structures
  - huge redundancy (> 500 T4 lysozyme)
- biases : 1. soluble, globular proteins 2. interesting proteins
- X-ray crystallography ≈ 85 %
- NMR ≈ 14 % (more in smaller proteins)

- File formats – standardisation - boring but important
  - all programs agree on a format – exchange of information
  - two PDB formats
    - one common – flat files..

Protein coordinate files



What would you expect ?

- Define the chain direction
  - N to C terminus
- within each residue
  - order of atoms
    - backbone
    - sidechain going away from backbone
- unit Å
- usually no Hydrogens

PDB File

```
ATOM      1  N   ARG A   1      26.465  27.452  -2.490  1.00 25.18           N
ATOM      2  CA  ARG A   1      25.497  26.862  -1.573  1.00 17.63           C
ATOM      3  C   ARG A   1      26.193  26.179  -0.437  1.00 17.26           C
ATOM      4  O   ARG A   1      27.270  25.549  -0.624  1.00 21.07           O
ATOM      5  CB  ARG A   1      24.583  25.804  -2.239  1.00 23.27           C
ATOM      6  CG  ARG A   1      25.091  24.375  -2.409  1.00 13.42           C
ATOM      7  CD  ARG A   1      24.019  23.428  -2.996  1.00 17.32           C
ATOM      8  NE  ARG A   1      23.591  24.028  -4.287  1.00 17.90           N
ATOM      9  CZ  ARG A   1      24.299  23.972  -5.389  1.00 19.71           C
ATOM     10  NH1 ARG A   1      25.432  23.261  -5.440  1.00 24.10           N
ATOM     11  NH2 ARG A   1      23.721  24.373  -6.467  1.00 14.01           N
ATOM     12  N   PRO A   2      25.667  26.396   0.708  1.00 10.92           N
…
ATOM     38  N   CYS A   5      23.095  22.004   2.522  1.00  7.84           N
ATOM     39  CA  CYS A   5      22.106  21.863   1.467  1.00  9.61           C
ATOM     40  C   CYS A   5      22.192  20.518   0.830  1.00 10.97           C
ATOM     41  O   CYS A   5      21.230  20.068   0.167  1.00  9.33           O
ATOM     42  CB  CYS A   5      22.358  22.904   0.371  1.00 10.97           C
ATOM     43  SG  CYS A   5      22.145  24.592   0.888  1.00 12.56           S
```

x *y* *z*

- Note coordinates
  - three decimal places – often 5 significant digits

PDB File

```
ATOM       1  N   ARG A   1      26.465  27.452  -2.490  1.00 25.18           N
ATOM       2  CA  ARG A   1      25.497  26.862  -1.573  1.00 17.63           C
ATOM       3  C   ARG A   1      26.193  26.179  -0.437  1.00 17.26           C
ATOM       4  O   ARG A   1      27.270  25.549  -0.624  1.00 21.07           O
ATOM       5  CB  ARG A   1      24.583  25.804  -2.239  1.00 23.27           C
ATOM       6  CG  ARG A   1      25.091  24.375  -2.409  1.00 13.42           C
ATOM       7  CD  ARG A   1      24.019  23.428  -2.996  1.00 17.32           C
ATOM       8  NE  ARG A   1      23.591  24.028  -4.287  1.00 17.90           N
ATOM       9  CZ  ARG A   1      24.299  23.972  -5.389  1.00 19.71           C
ATOM      10  NH1 ARG A   1      25.432  23.261  -5.440  1.00 24.10           N
ATOM      11  NH2 ARG A   1      23.721  24.373  -6.467  1.00 14.01           N
ATOM      12  N   PRO A   2      25.667  26.396   0.708  1.00 10.92           N
…
ATOM      38  N   CYS A   5      23.095  22.004   2.522  1.00  7.84           N
ATOM      39  CA  CYS A   5      22.106  21.863   1.467  1.00  9.61           C
ATOM      40  C   CYS A   5      22.192  20.518   0.830  1.00 10.97           C
ATOM      41  O   CYS A   5      21.230  20.068   0.167  1.00  9.33           O
ATOM      42  CB  CYS A   5      22.358  22.904   0.371  1.00 10.97           C
ATOM      43  SG  CYS A   5      22.145  24.592   0.888  1.00 12.56           S
```

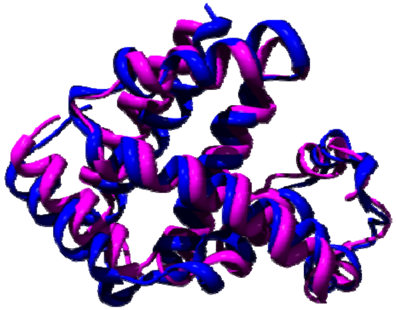
residue

mobility

- Given some coordinates – how to compare them ?

Comparing coordinates

- These are very similar



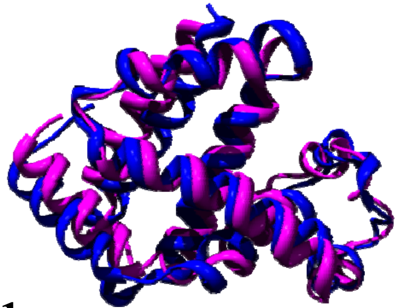
- These are clearly related, less similar

- We want to put numbers on this property

First some notation

- We have spoken of x, y, z coordinates. Easier..
  - vector $\vec{r}$ or for atom i, $\vec{r}_i$
  - for two proteins let us have position i in protein a and b
  - $\vec{r}_i^{\,a}$ and $\vec{r}_i^{\,b}$

Comparing two proteins



- take one atom (C^{α}) from residue *i*
- what do I know from the picture ?
- if my two proteins are similar $\vec{r}_i^{\,a} - \vec{r}_i^{\,b}$ will be a short vector
- for each residue *i*
- define $\left|\vec{r}_i^{\,a} - \vec{r}_i^{\,b}\right|$ distance between $\vec{r}_i^{\,a}$ and $\vec{r}_i^{\,b}$

- I want a single number that tells me
  - usually
  - how close is a residue in *a* to the corresponding residue in *b*
  - think of the set of distances $\left|\vec{r}_i^{\,a} - \vec{r}_i^{\,b}\right|$
  - how spread out is this population of distances ?
    - like a standard deviation (standard Abweichung)

Root mean square (rms)

- normal formula for standard deviation $\sigma_x = \left(\frac{1}{N} \sum_{i=1}^{N} \left(x_i - \bar{x} \right)^2 \right)^{1/2}$

- something similar for coordinates

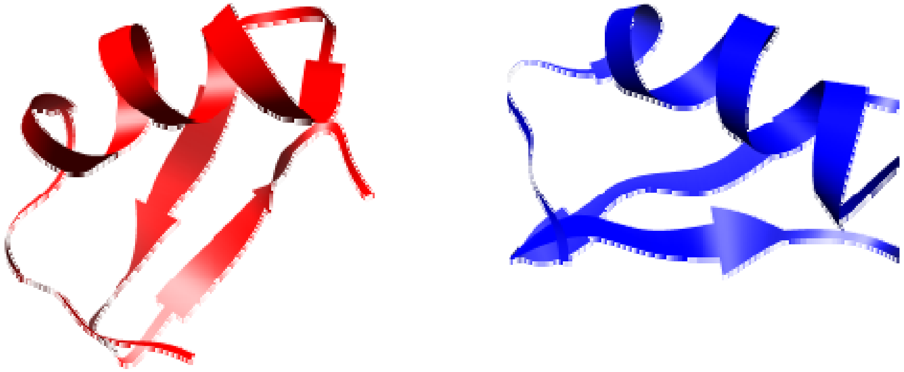
$$r_{rmsd} = \left(\frac{1}{N_{res}} \sum_{i=1}^{N_{res}} \left| \vec{r}_i^{\,a} - \vec{r}_i^{\,b} \right|^2 \right)^{1/2}$$

- where proteins *a* and *b* have N_{res} residues
- *rmsd* is “root mean square difference”

- complications

Before calculating rmsd

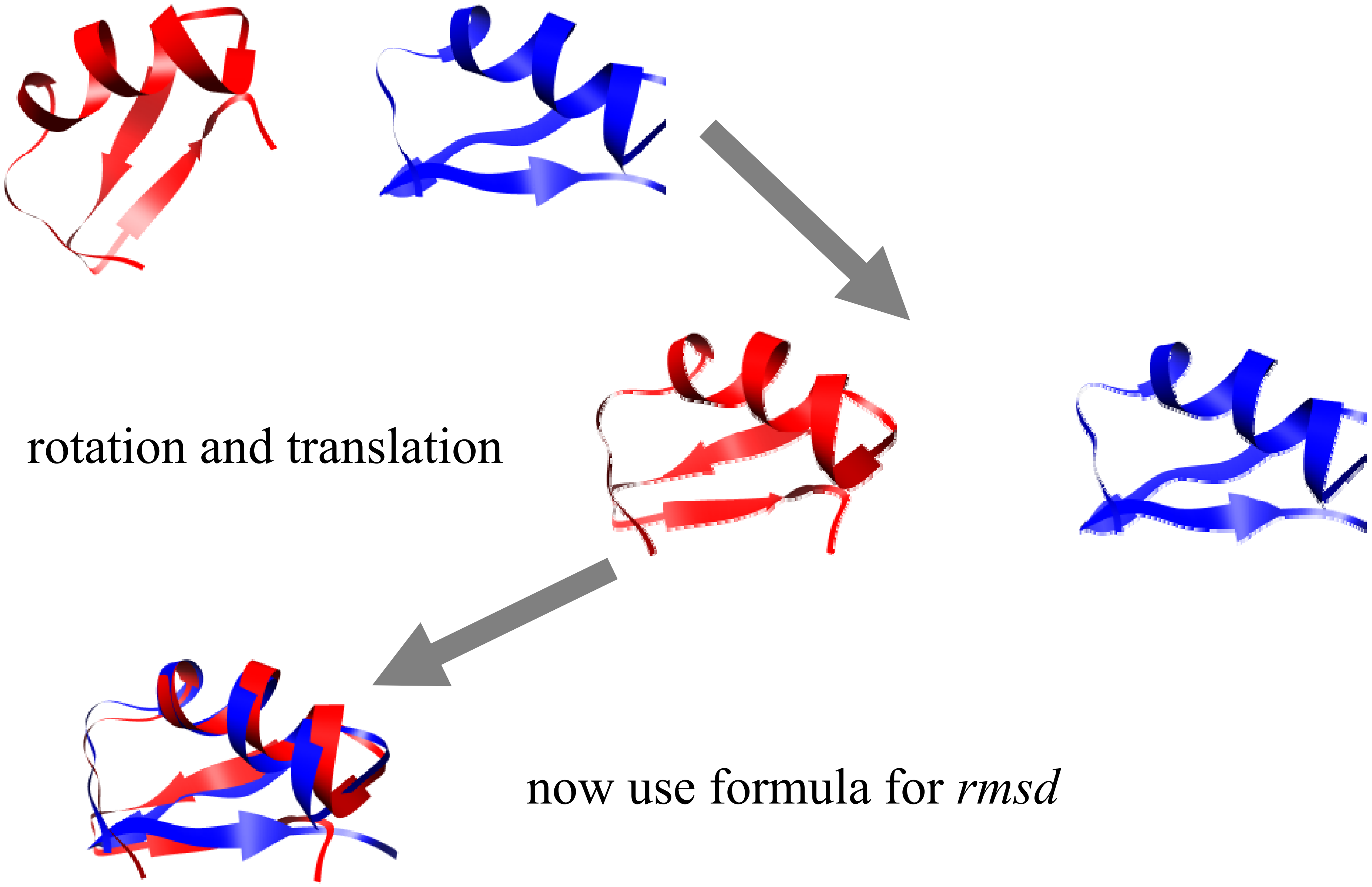
- two very similar proteins
  - coordinates are in different orientations
  - not on top of each other



- what are the orientations of files in PDB ?
  - totally arbitrary

- first some other steps

Superposition of coordinates



rotation and translation

now use formula for *rmsd*

First problems with *rmsd*

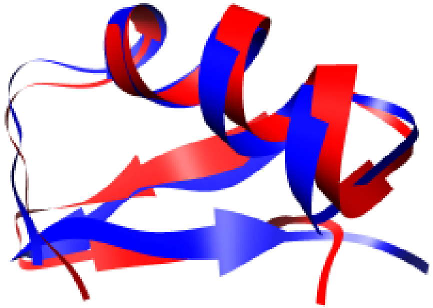
- Before calculating *rmsd*
  - coordinates must be “superimposed” (translation + rotation)

- if you and I use slightly different superpositions
  - our *rmsd* values (similarity) will be different

Meaning of *rmsd*

- units Å
- *rmsd* is size dependent
  - 5 Å in a small protein (50 residues) will not look similar
  - 5 Å in a big protein (250 residues) will look similar

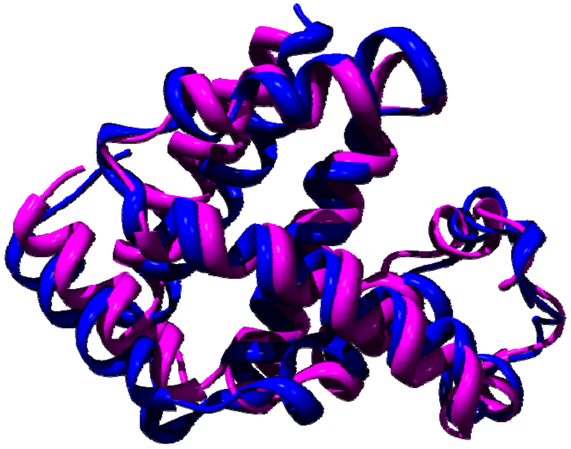
Difficulty with *rmsd*



- these two proteins have the same number of residues

$$r_{rmsd} = \left(\frac{1}{N_{res}} \sum_{i=1}^{N_{res}} \left| \vec{r}_i^{\,a} - \vec{r}_i^{\,b} \right|^2 \right)^{1/2}$$

- if *i* = 1, 2, 3, .. we use residue 1, 2, 3 in both proteins

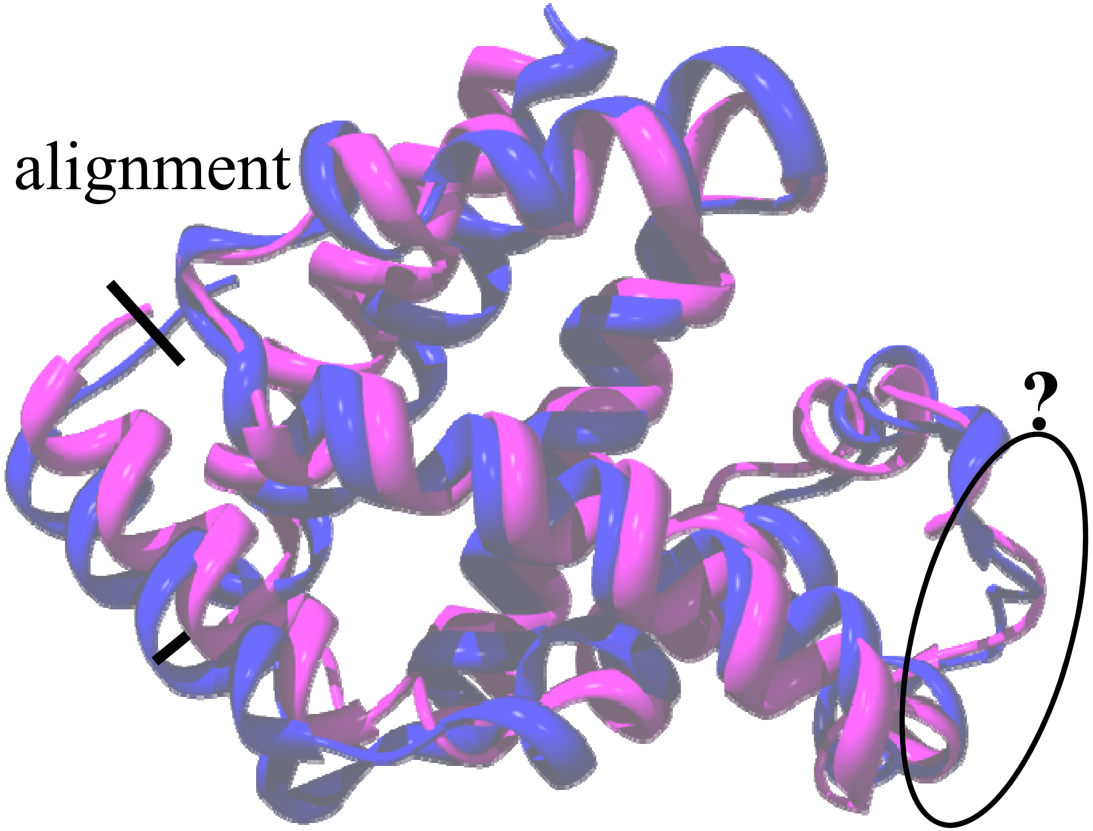


- these two proteins have slightly different numbers of residues

- we cannot compare residue 1 to 1, 2 to 2..

Proteins of different sizes – first version

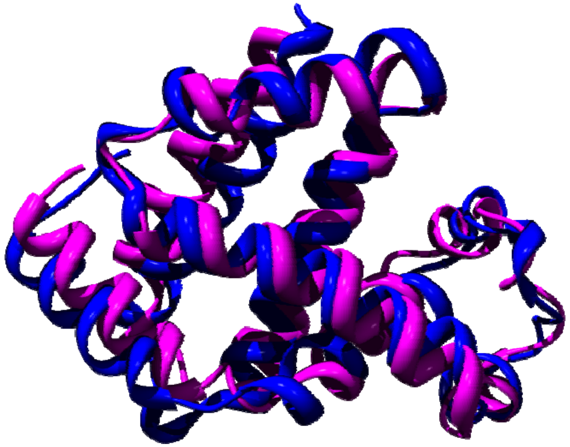
- Problem - for each residue *i* in protein *a* we need matching residue in protein *b*
- One approach
- first build a sequence alignment



Selecting residues for alignment

- take the sequence of each protein, calculate alignment

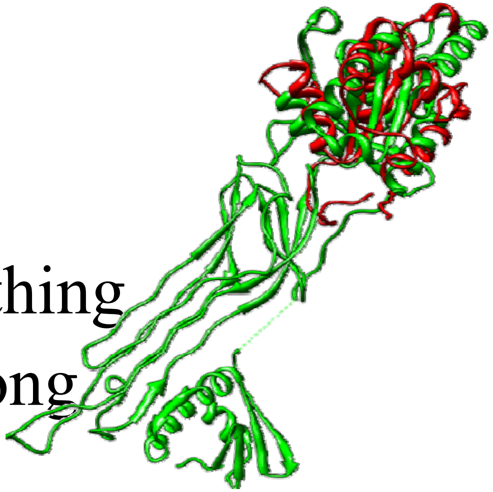
```
ACDEFG-IK-MNP..
A-DEGGHIKLMNP..
```



use these residues

AC**DEFG**-**IK**-**MNP**..
A-**DEGG**H**IK**L**MNP**..

- will find corresponding residues
- will allow for missing / inserted residues
- used in some programs – chimera
- problem … sequence similarity may be near nothing
    - a sequence based alignment may be very wrong



Selecting residues for alignment - better

- We need corresponding residues
  - some kind of alignment
- can one do an alignment based on structures ?

- Answer : yes but..
  - no guaranteed correct solution
  - many different methods

Summary of comparing two structures

- we want a single measure of similarity (like *rmsd*)
- this requires we have a set of corresponding residues in the two proteins
- if there is good sequence similarity – use it
- naïve methods will not give the best superposition
- structure-based alignments can be calculated
  - require approximations
  - often slow
  - can not guarantee the best answer

Summary of everything

- Similarities
  - Sequence level – finding them
    - Multiple sequence alignments leads to evolution
  - Structure
    - Harder to find – more valuable for remote relations