Multiple sequence alignments similarity without sequence similarity

Andrew Torda,
Bioinformatics,
Sommersemester 2016

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

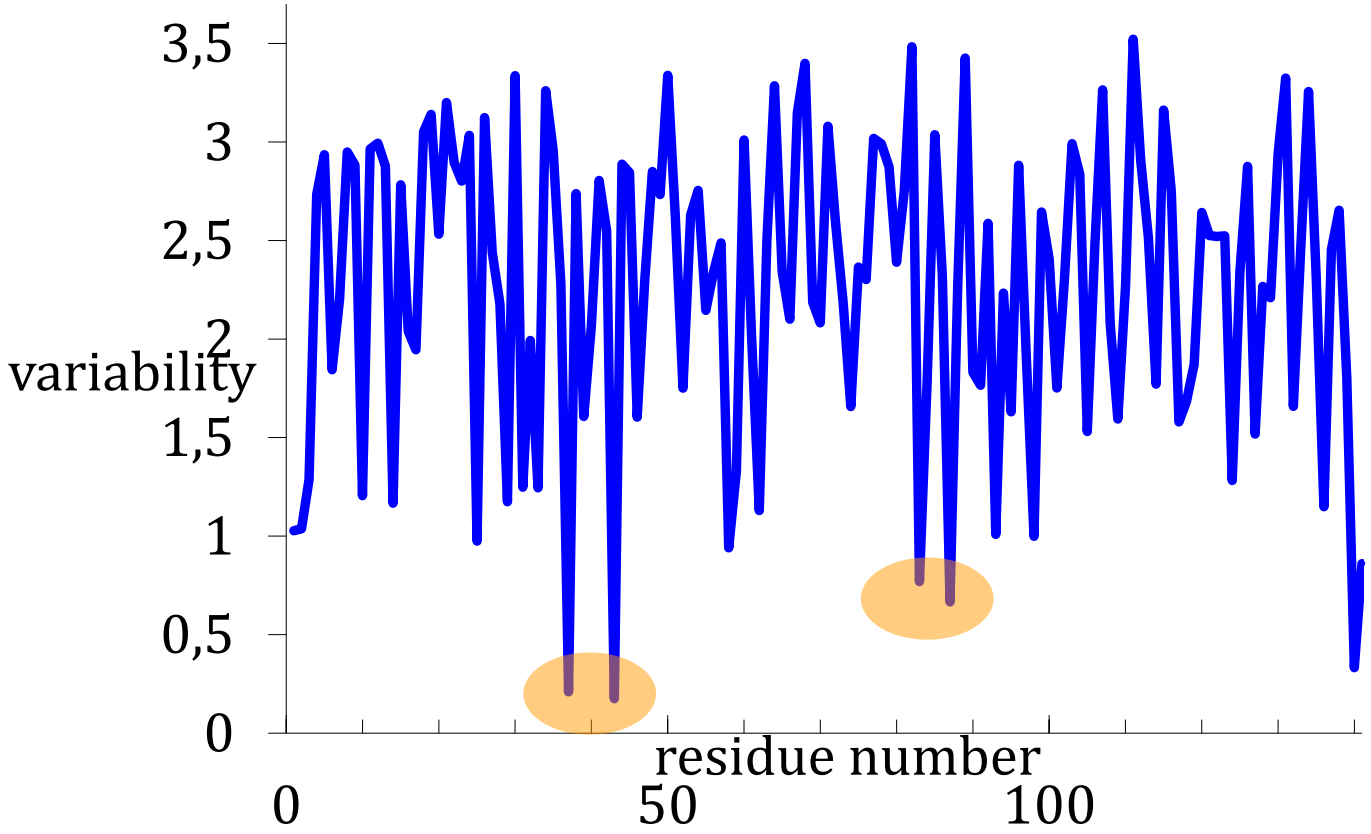
Multiple alignments

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

```
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

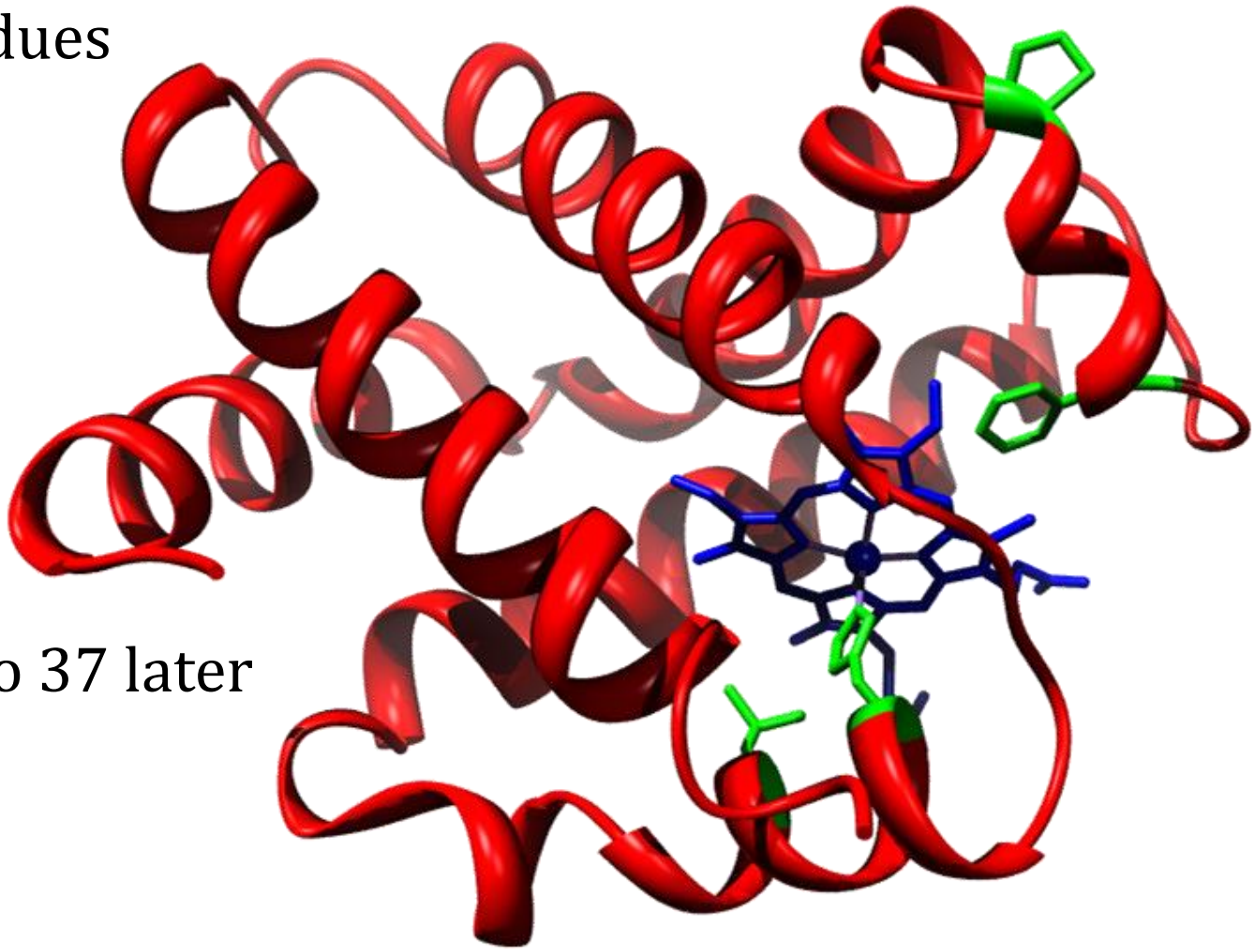


- how do we get these and what does it mean ?
- what does it mean for this protein ?

Conserved residues

Proximity to haem group

- green residues



- more on pro 37 later

Beliefs in multiple sequence alignments

Similar proteins found in many organisms

- where they are conserved - connected with function
- variation reflects evolution (phylogeny)

How many homologues might you have ?

- many
  - some DNA replication proteins – almost every form of life
  - profilin – cell mobility – bacteria, mammals, plants
  - ..
- few
  - exotic viral proteins
  - messengers exclusively in human biochemistry
  - …

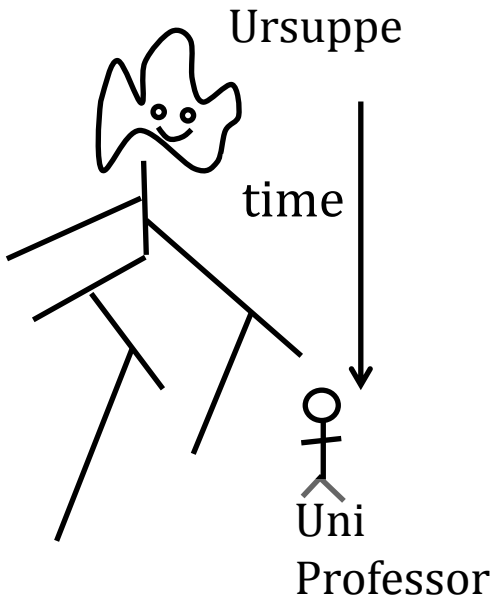
Trees / Phylogeny

Multiple sequence alignments are fun

- conservation, function…

What next ? Phylogeny - making trees

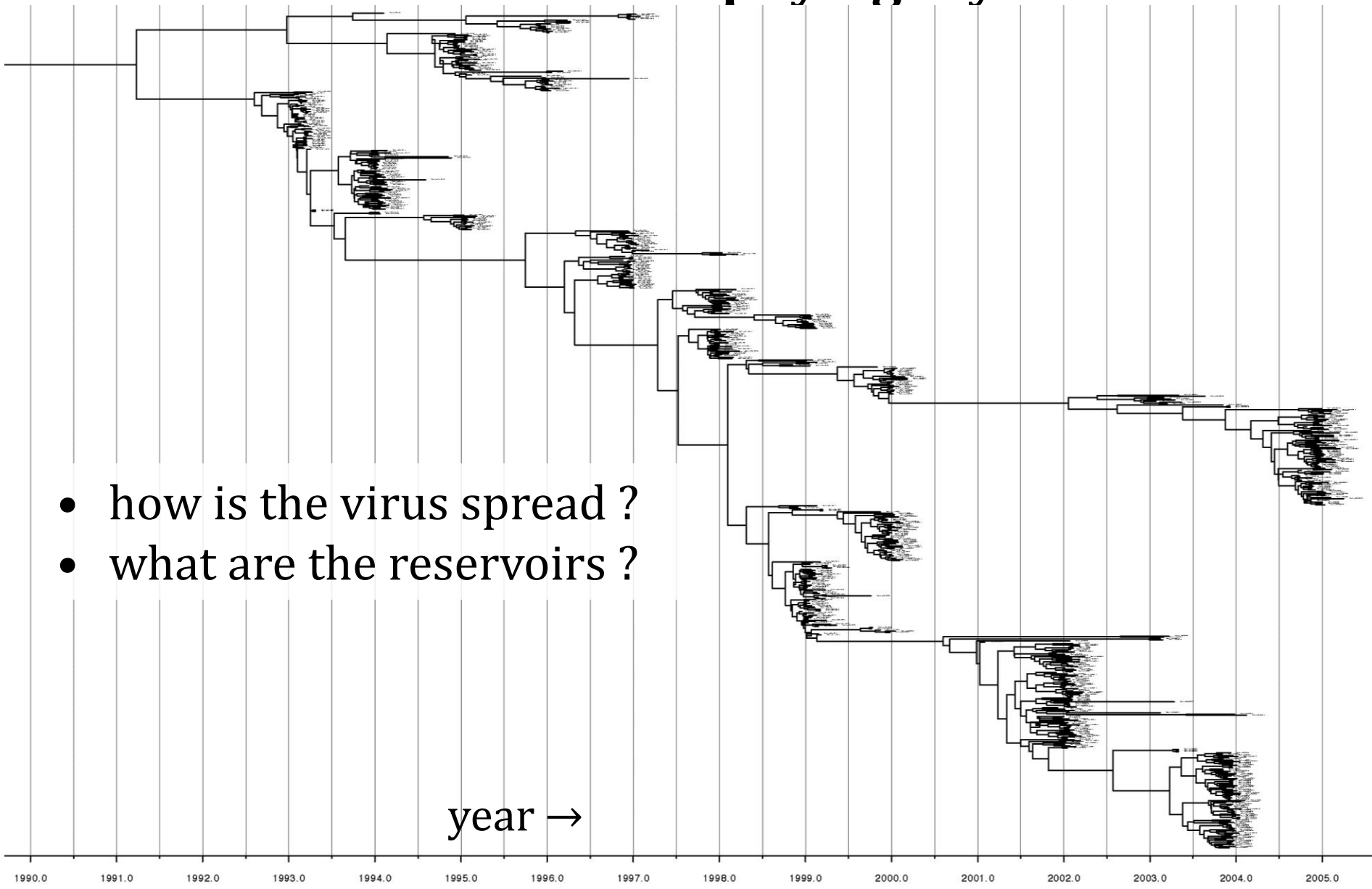
- Need multiple sequence alignments to make trees



Do you just want the tree of life ?

- who killed the bananas ?
- where does influenza come from ?
- lassa, swine flu, ebola
- who killed the ladies ?

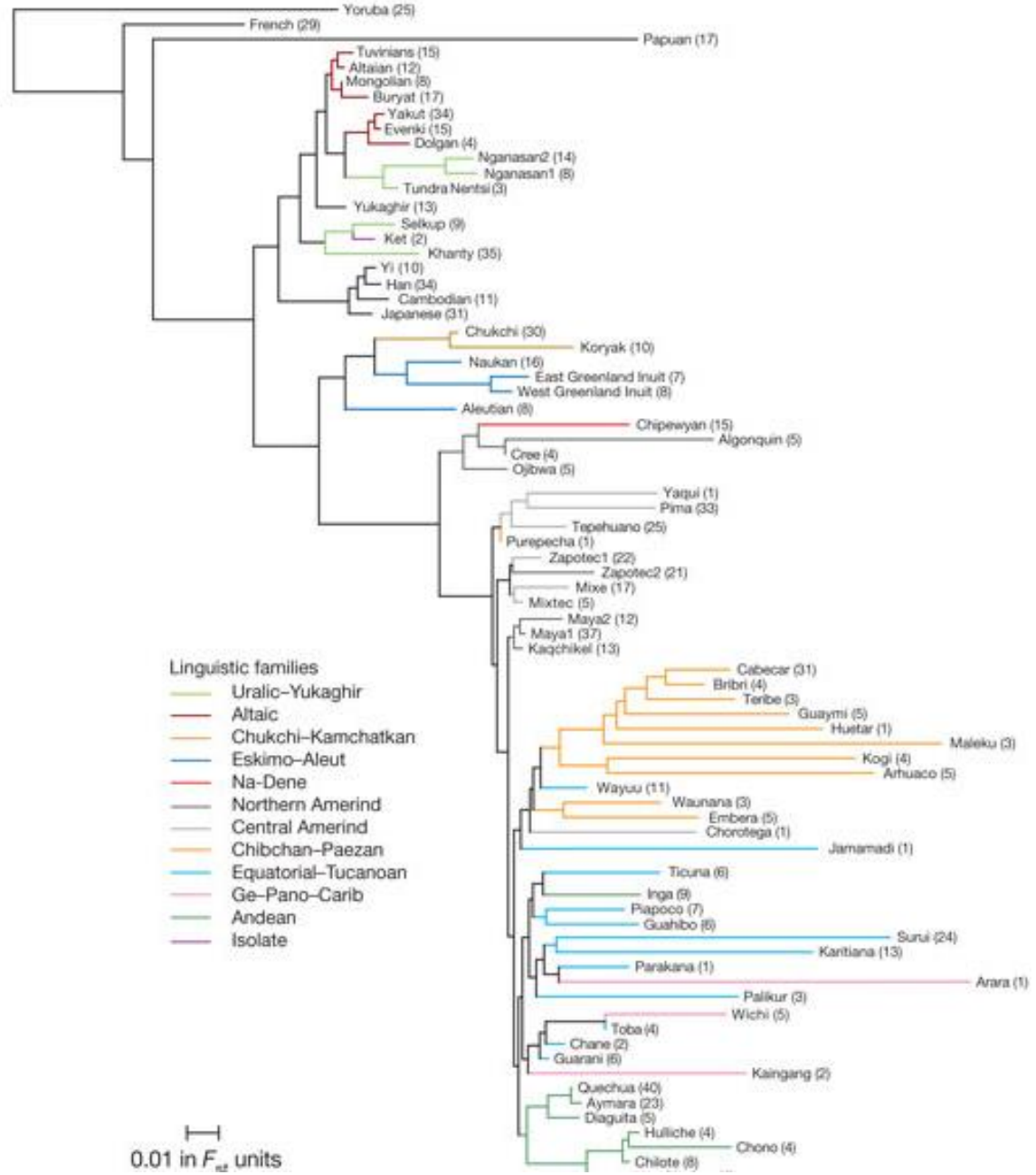
Influenza virus phylogeny



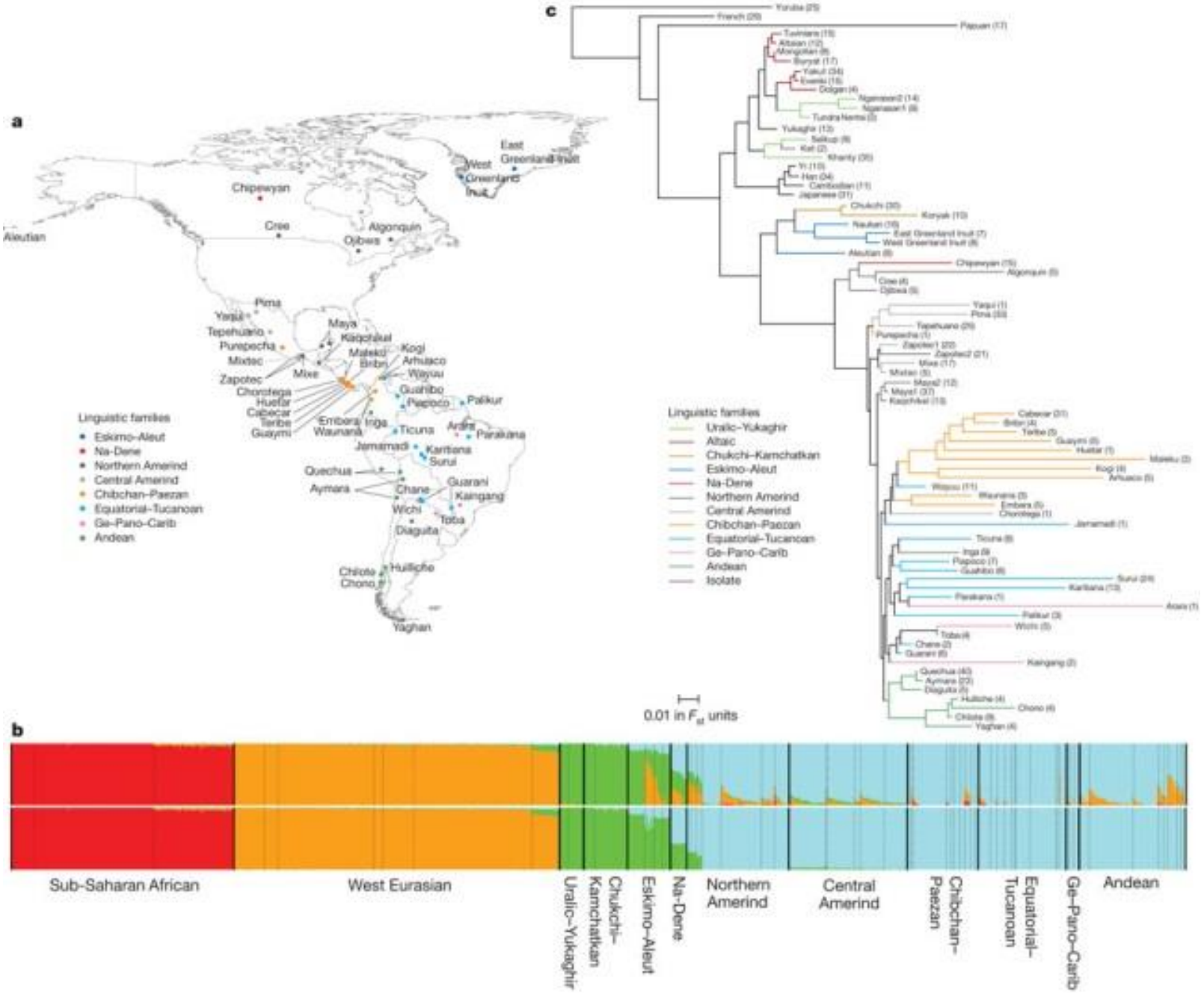
- how is the virus spread ?
- what are the reservoirs ?

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

52 Native American groups

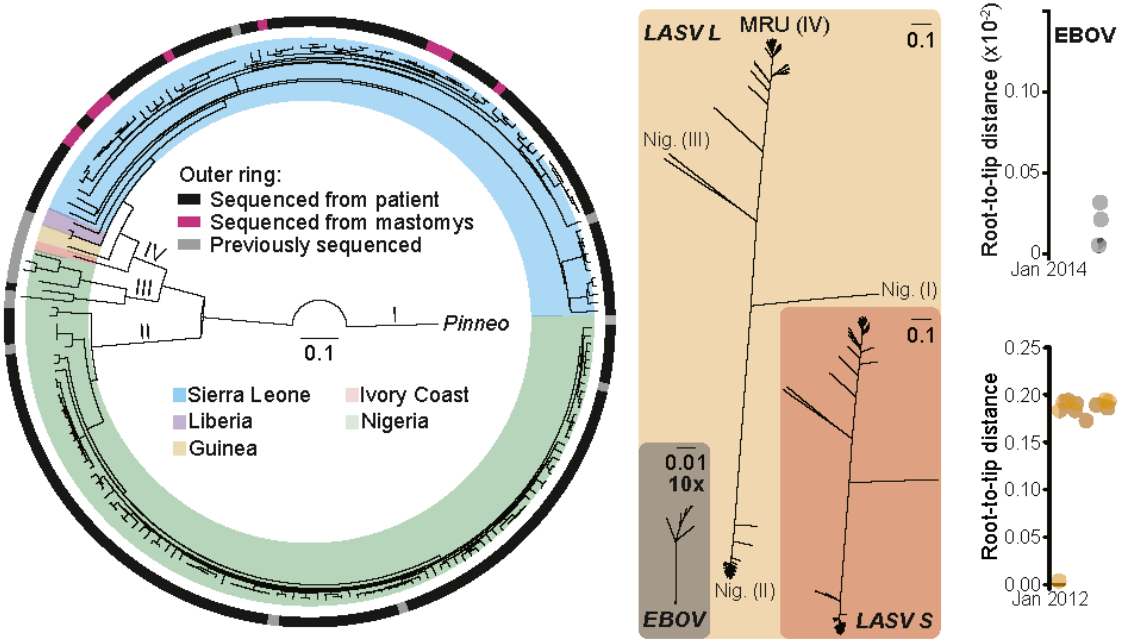


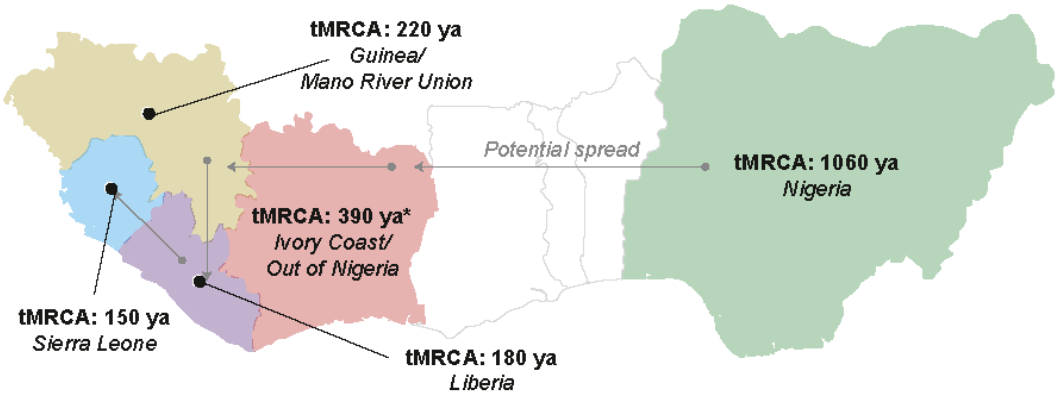
Reich, D., …Ruiz-Linares, A., Nature, 488, 370 (2012), Reconstructing Native American .. History



Reich, D., …Ruiz-Linares, A., Nature, 488, 370 (2012), Reconstructing Native American .. History

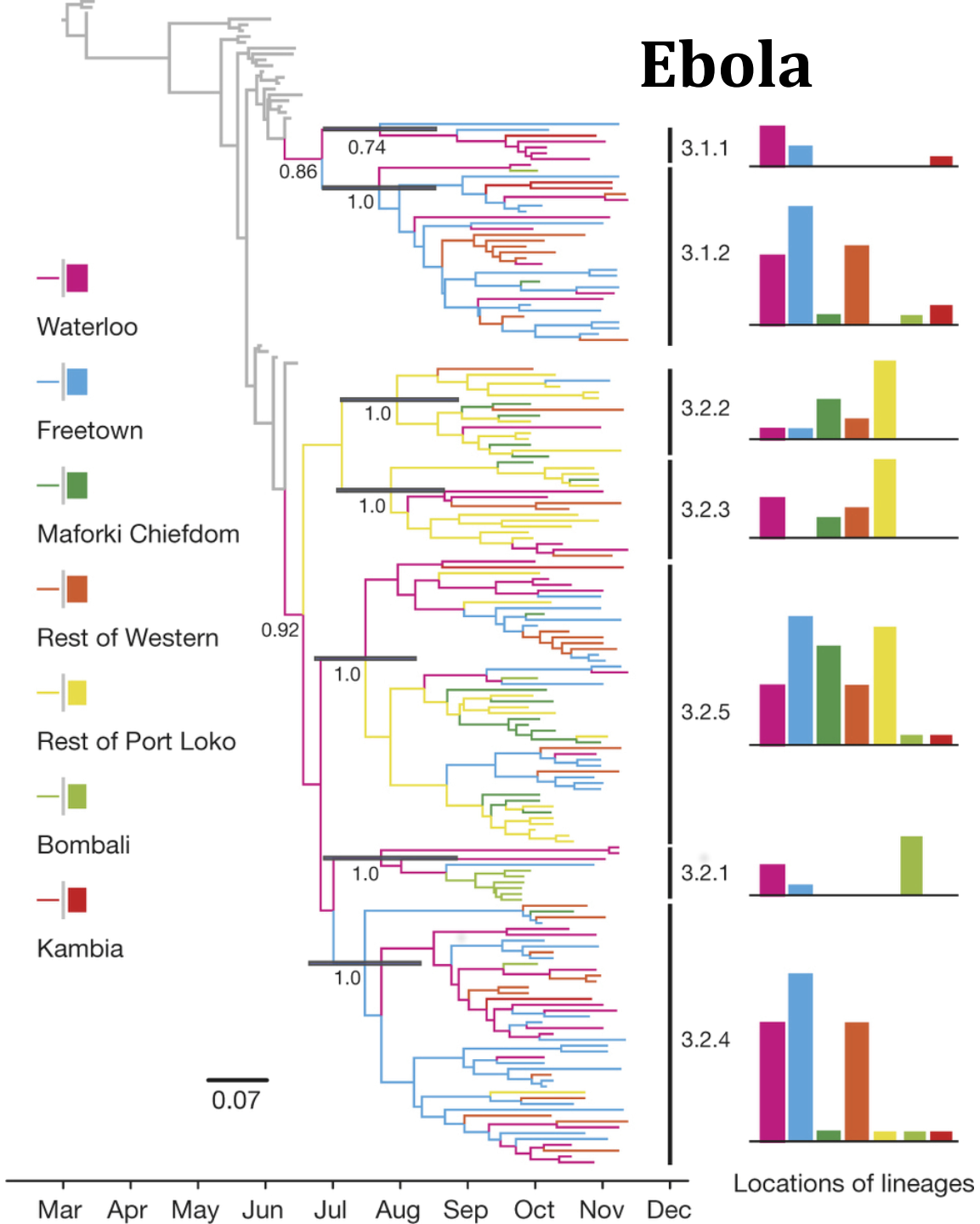
lassa virus





Andersen, KG... Sabeti, PC, Cell, 162, 738-750 (2015), "Clinical sequencing .. lassa virus"

Ebola

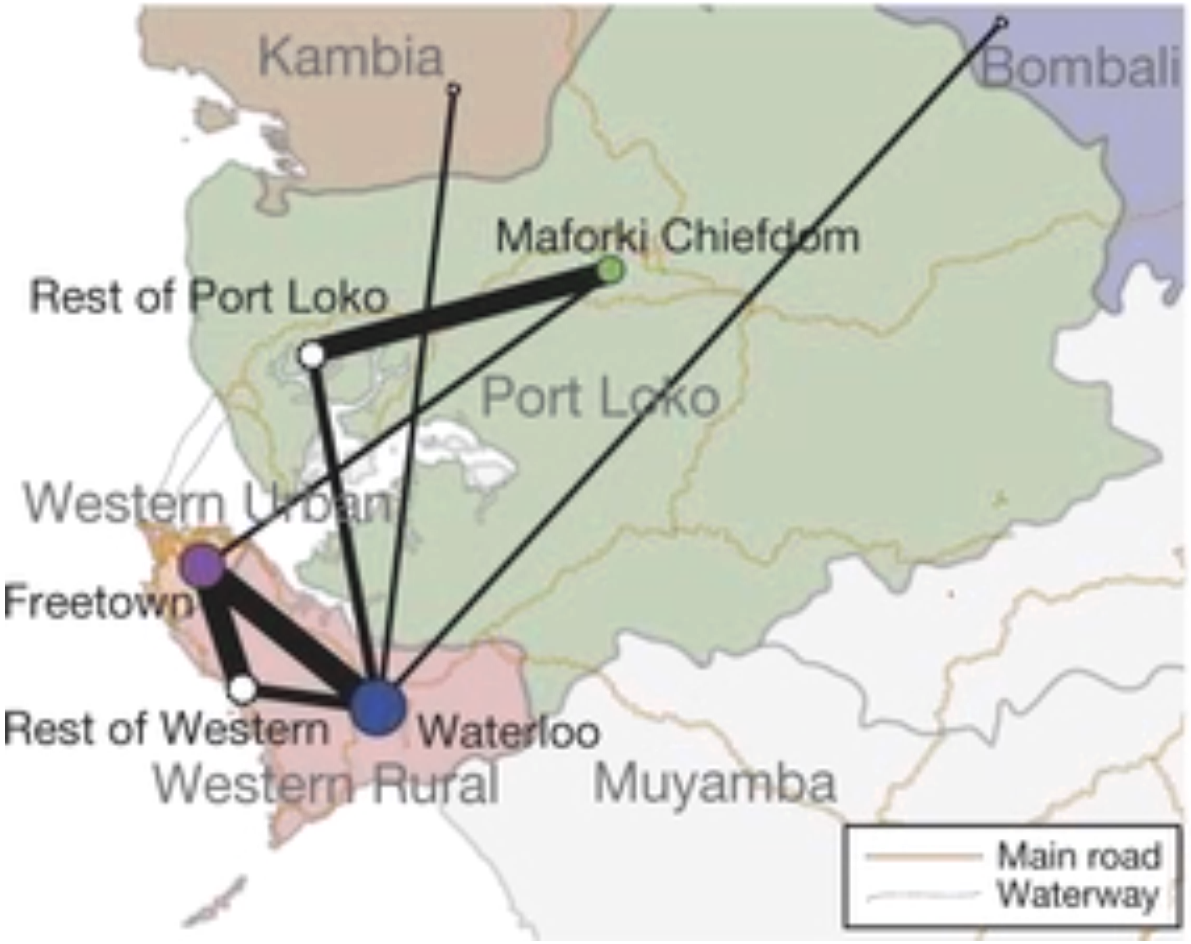


How much of each type was found in each region ?

How did the virus spread ?

Tong, Y-G., Shi, W.F. *et* al, Nature, 524, 93-96, 2015

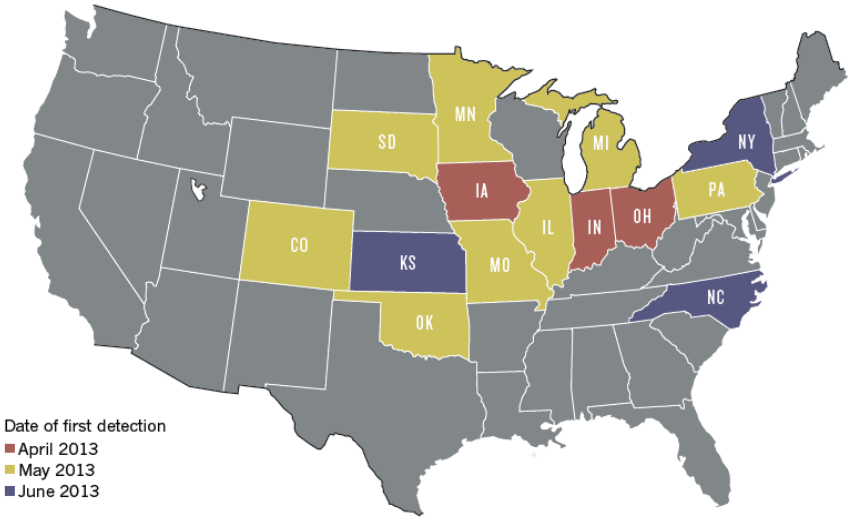
How did the virus spread ?



Thickness of lines – closeness of sequences

PIG VIRUS ON THE WING

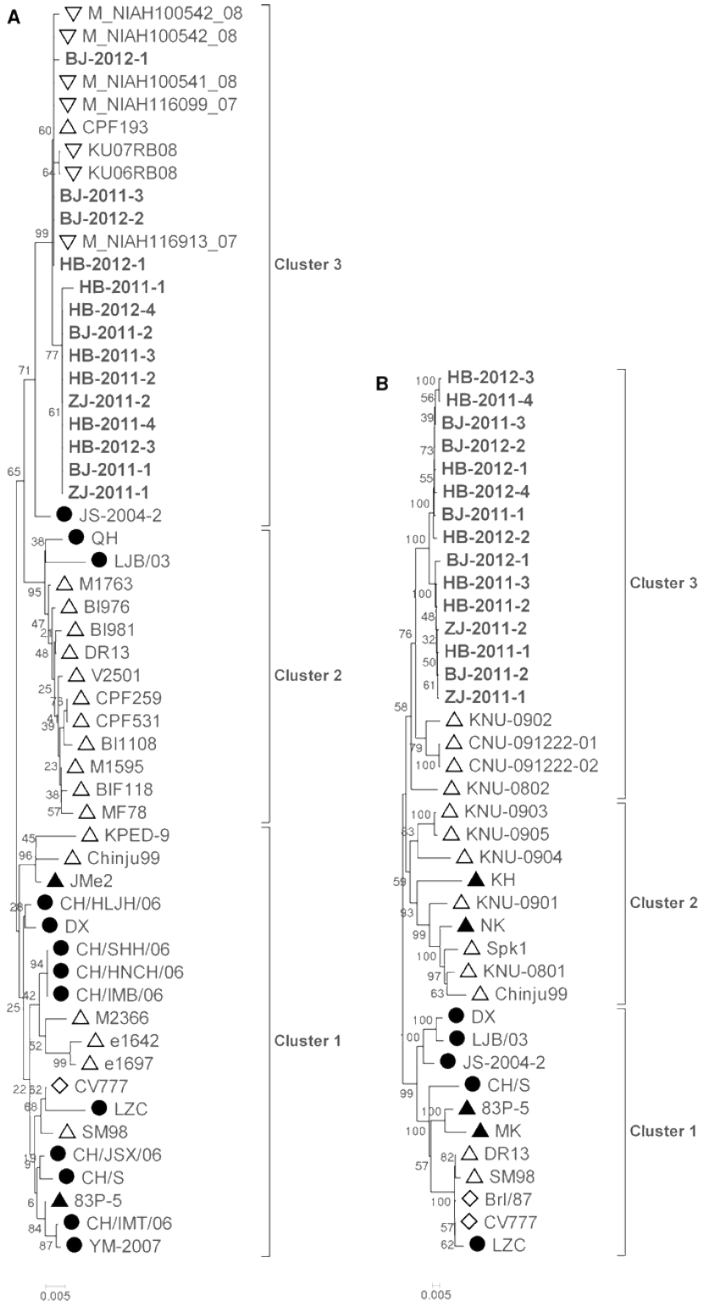
Porcine epidemic diarrhoea virus, a type of coronavirus that can kill piglets, has been detected in 14 US states.



ANIMAL DISEASE

Deadly pig virus slips through US borders

Researchers race to track spread of coronavirus.



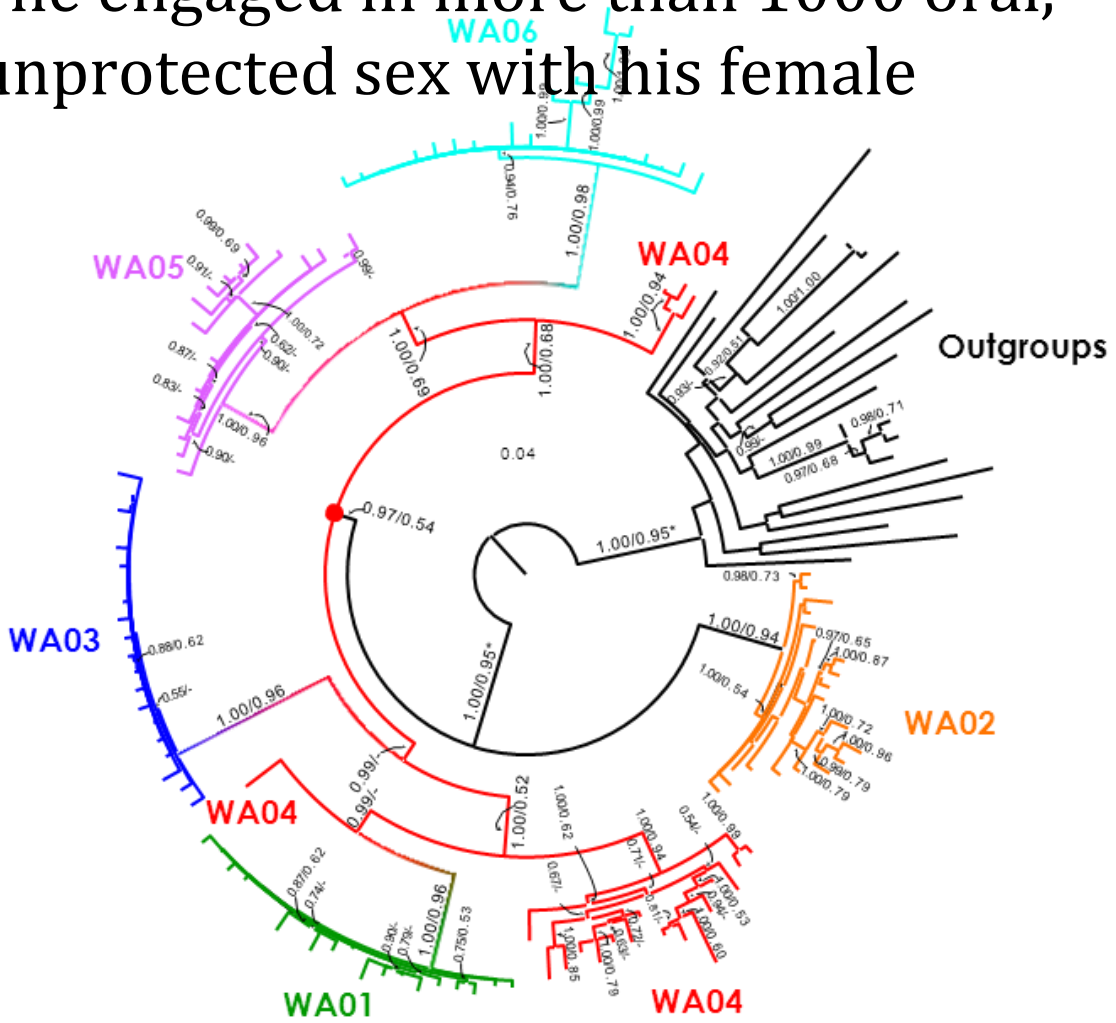
Mole, B., Nature, 499, 388 (2013)

Gao, Y. Kou, Q., Ge, X, Zhou, L., Guo, Yang, H., Arch Virol, 158, 711-715 (2012)

1000 acts of sex

"the defendant intended to inflict “great bodily harm”… Between 1999 and 2004, he engaged in more than 1000 oral, vaginal, and anal acts of unprotected sex with his female partners"

Colour changes where WA04 infected ladies



The plan

- optimise and alignment and tree simultaneously

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

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

```
VLSPADKSNVKAGWGQVGAHAGDYGAEAIERMYLSFPSTKTYFPHTDISHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
```

In pairwise problem

- Sum over match()
  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

$$d_{a,b} = \frac{1}{S_{a,b}}$$

- distance measure..
  which sequences are most dissimilar 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
```

- then I move 5 and 2 & 5 and 3 – messes up 2 and 3

Mission: for N_{seq} sequences

- $S_{a,b}$: alignment score sequences a and b

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

- not quite possible
  - this method is just an approximation

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 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

Guide tree / progressive / neighbour joining method

Steps

- build a distance matrix
- build a guide tree
- build up overall alignment in pieces

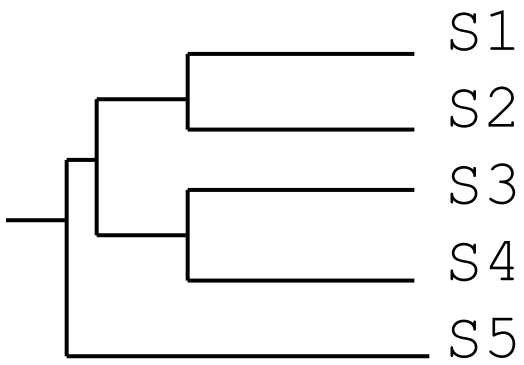
Progressive alignment - tree

```
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

- 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
```

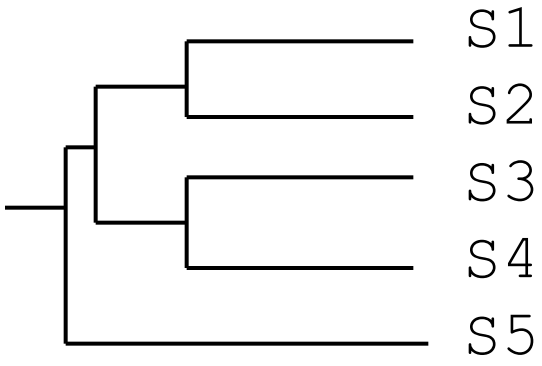
align av(S1,S2) with av(S3,S4)

```
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

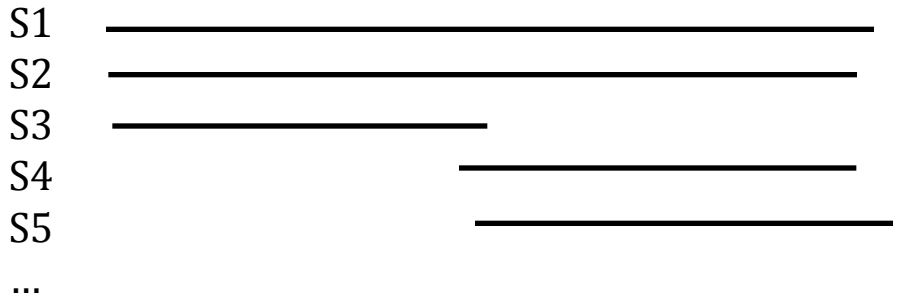


	S1	S2	S3	S4	S5
S1	-				
S2	.11	-			
S3	.20	.30	-		
S4	.27	.36	.09	-	
S5	.30	.33	.23	.27	-

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 from bacteria to man)
  - lots of little errors
- can break completely (domain example)

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

- more later

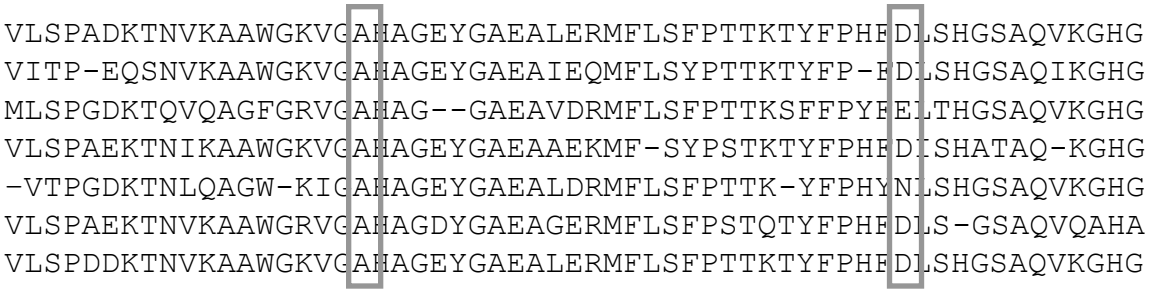
Measuring conservation / entropy

Entropy

- how much disorder do I have ? $S = -k \sum_{i=1}^{N_{states}} p_i \ln p_i$
- in how many states may I find the system ?

Our question

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



no disorder

much disorder

Calculate an "entropy" for each column

Entropy

- 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 ? $p_i = 1$ or $p_i = 0$

$S = \ln 1 = 0$ (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…

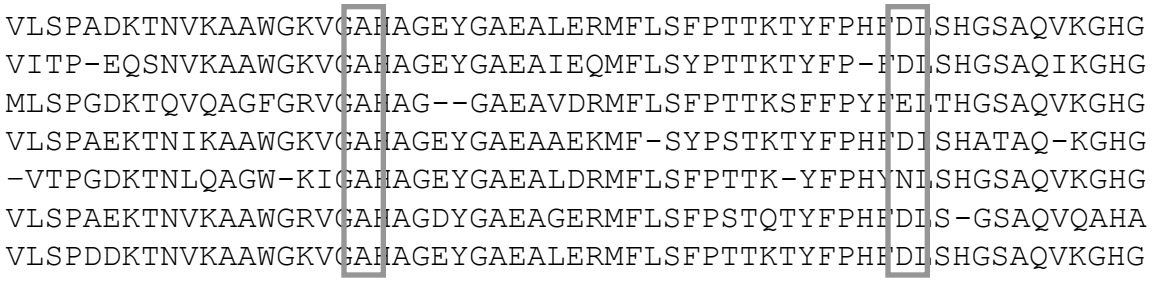
Entropy

- First column is boring
- Second

$p_{\mathrm{D}} = {}^{5}/_{7}$

$p_{\mathrm{E}} = {}^{1}/_{7}$

$p_{\mathrm{N}} = {}^{1}/_{7}$



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

$$\approx 0.8$$

Entropy from DNA

Exactly as for proteins

Will numbers be larger or smaller ?

max possible entropy

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

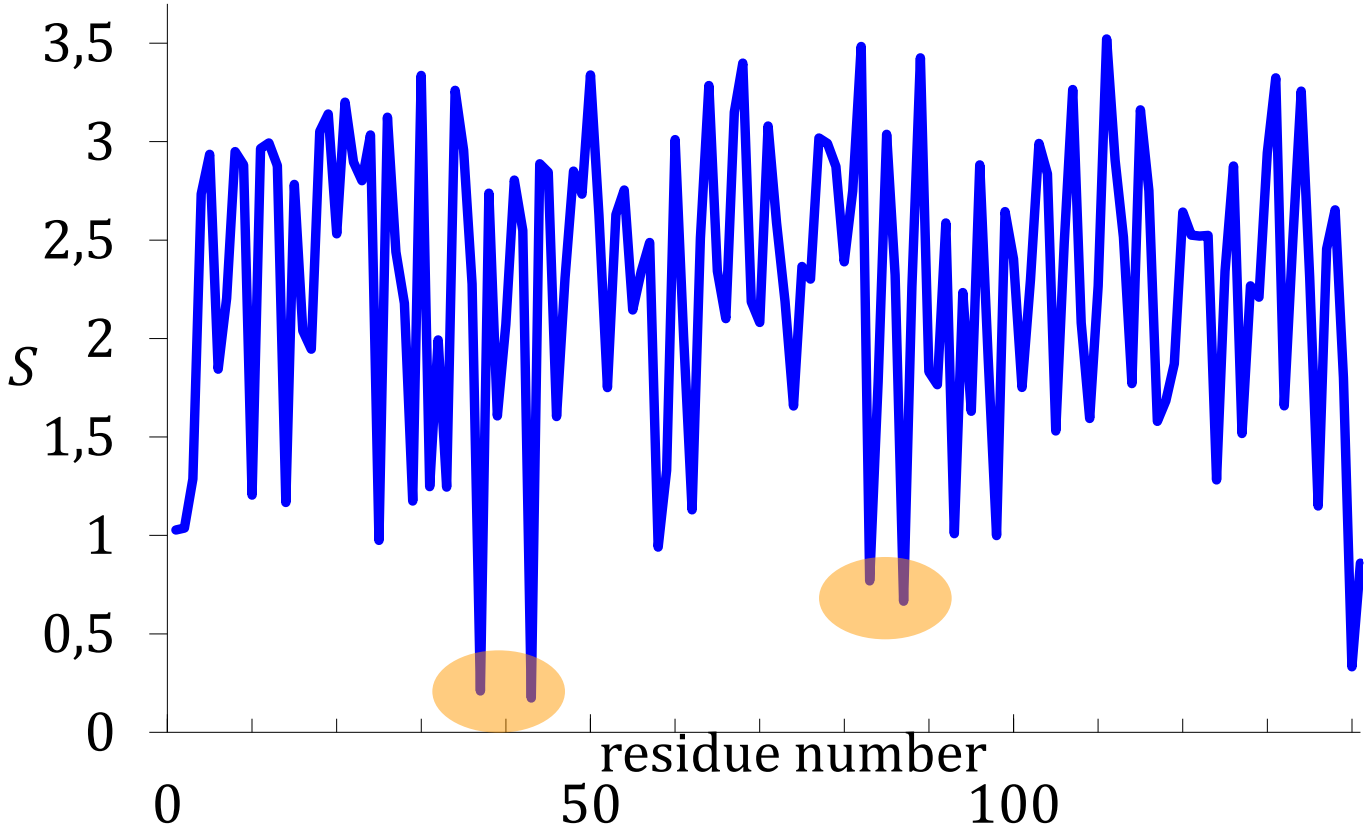
$$= -\ln\frac{1}{4}$$

$$\approx 1.4$$

example from start of this topic

Haemoglobin conservation

Look at residues 37, 43, 83 and 87



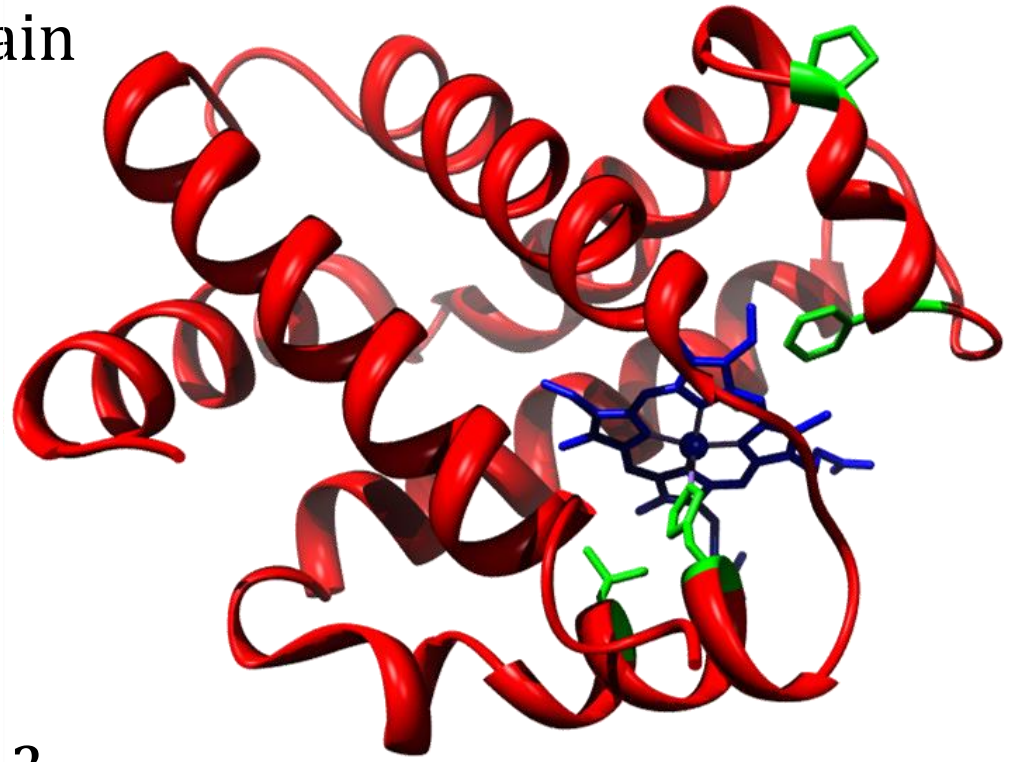
4 residues (maybe more) stand out as conserved

- 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

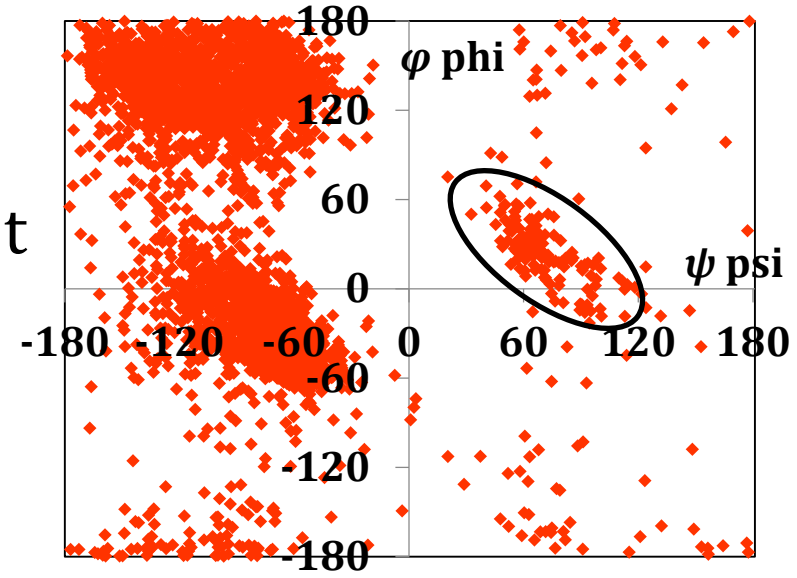
What is special about proline ?

- no H-bond 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

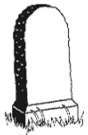
- what residue may I change to allow binding to a solid substrate ? (for biosensor/immobilized enzyme ?)
- try error prone PCR to select for new enzyme activity – which sites might I start with (active site) ?

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 
   - active site residues are 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

Where would you evolve to ?

1. very fragile
2. likely to survive mutations

Resistance to mutations…

Tolerance of mutations

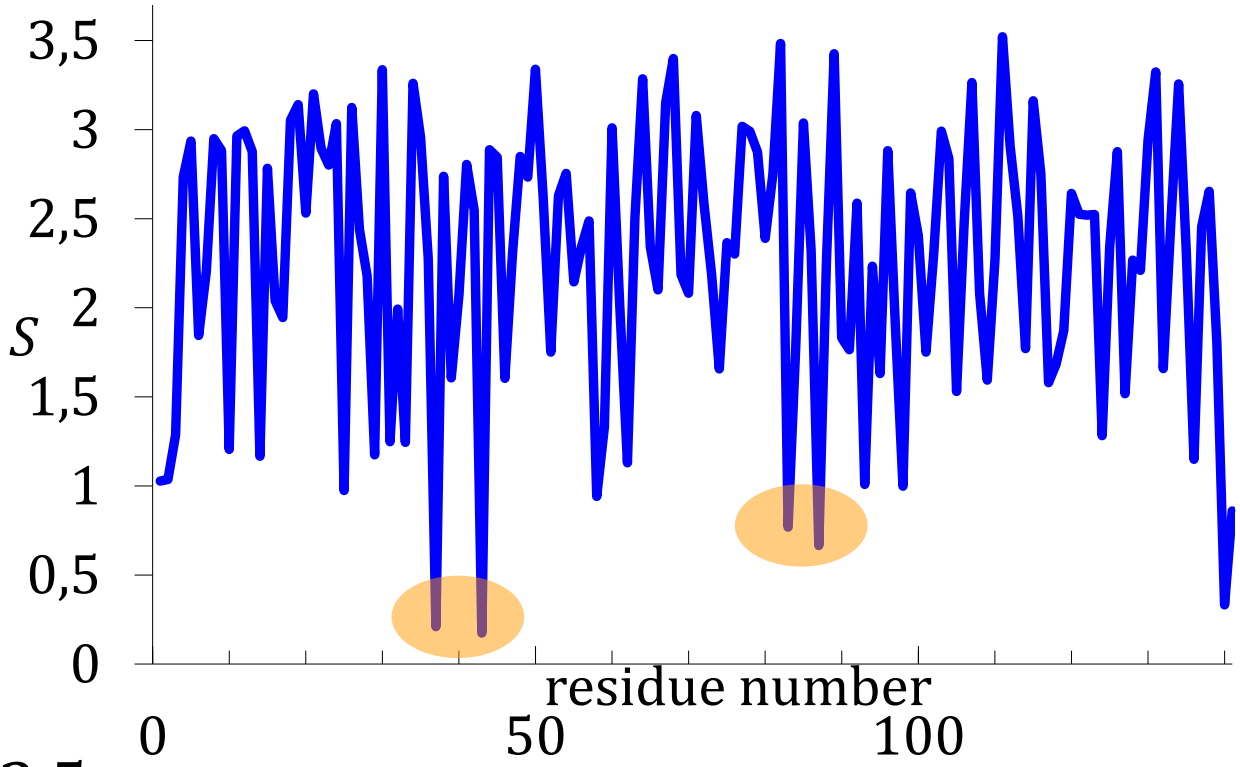
Boring answer

- some amino acids are similar to each other

Better answer

- it will be selected for
- you genes have a better chance of being passed on
- it is a Darwinian trait

Conservation – how meaningful ?



Earlier Folien…

- values from 0 to 3.5

What if I used more homologues ?

Conservation – how meaningful ?

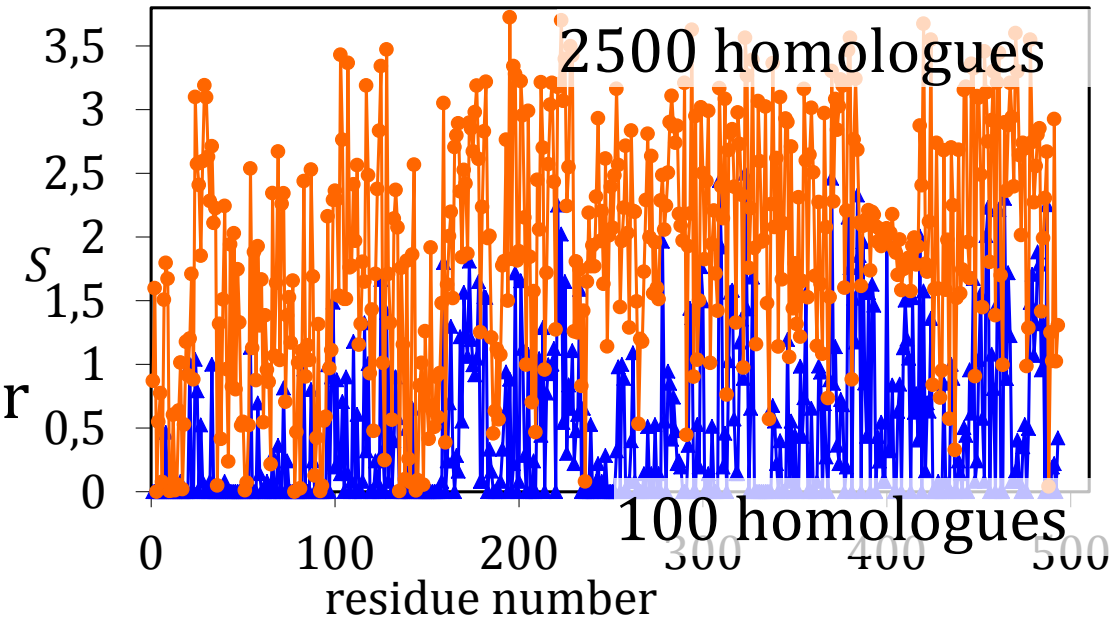
Example sequence (1ab4, DNA gyrase)

- find 100 close homologues (mostly > 80% similarity) – calculate conservation
- find 2500 close homologues (mostly > 50 % similarity) calculate conservation

Fewer sequences

- lots of conserved sites

- you can get the answer you want



Consequences - summarise

Significance of conservation

You read in a paper – residue 37 is conserved

- how many sequences did they look at ?
  - 10 ? bad 100 better 1000 better

- choosing the number of sequences lets you manipulate results

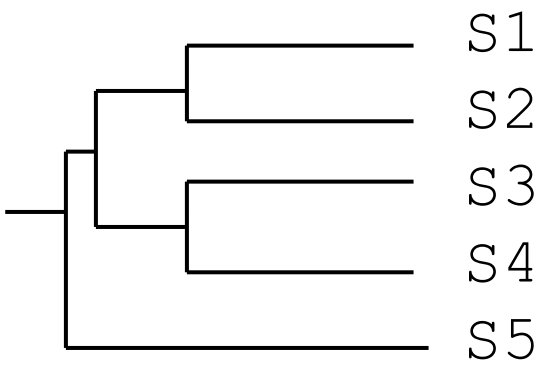
- statistically
  - have you sampled over enough sequences ?

Phylogeny / Evolution

The trees in text books are almost never perfect

One rarely knows the correct history

Problems..



Previously we had a "guide tree"

- did (S1,S2) and (S3,S4) share an ancestor but not S5 ?
- 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 → C → G two events (although looks like A→G)

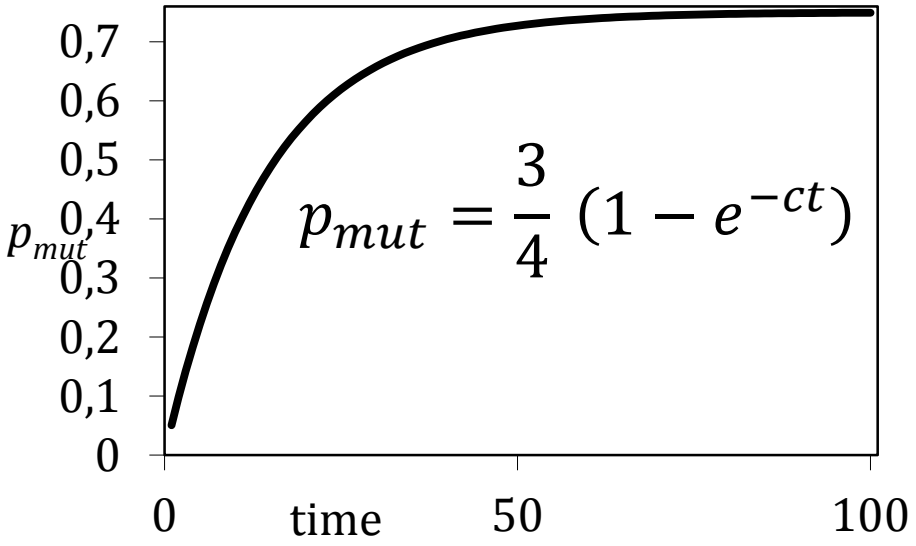
A → C → G → C → A looks like zero mutations

If I have infinite time

- all bases / residues equally likely
- p_{mut} =3/4 = 0.75 (DNA) or p_{mut}=19/20 (protein)

Mutation probability

- time units are arbitrary
- how would I estimate time ? (for DNA)
- $t \propto -\ln\left(1 - \frac{4}{3} p_{mut}\right)$
- p_{mut} ? count $\frac{n_{mut}}{n_{res}}$
- work in relative time



For short times, p_{mut} changes fast

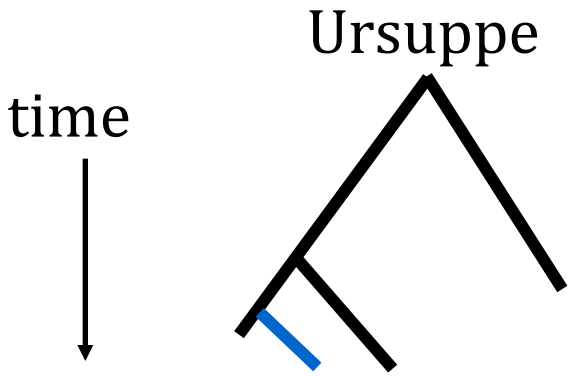
- for small *t*, distances will be more reliable
  - as will be alignments

Is this enough for phylogeny ?

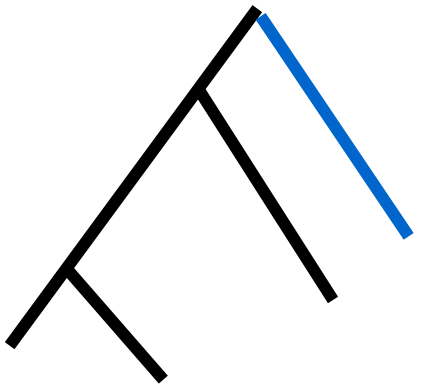
- what about reliability ?

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

Backwards mutations ?

- not really a problem (Klausurerfahrung)

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

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

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…

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 ?

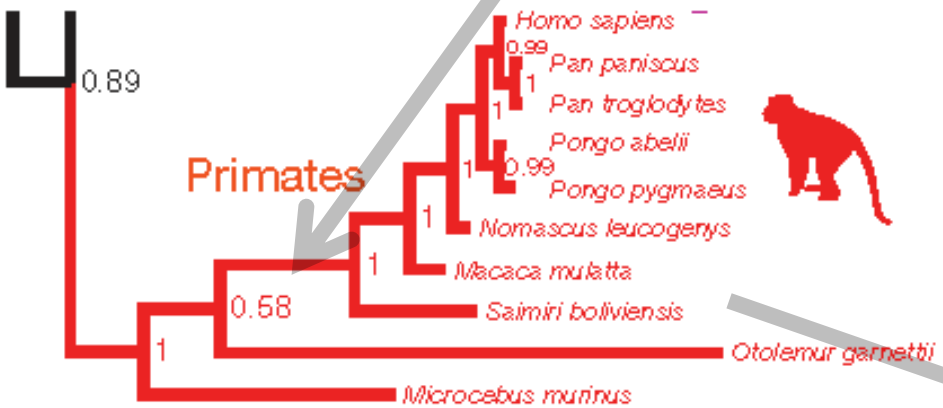
Monster example

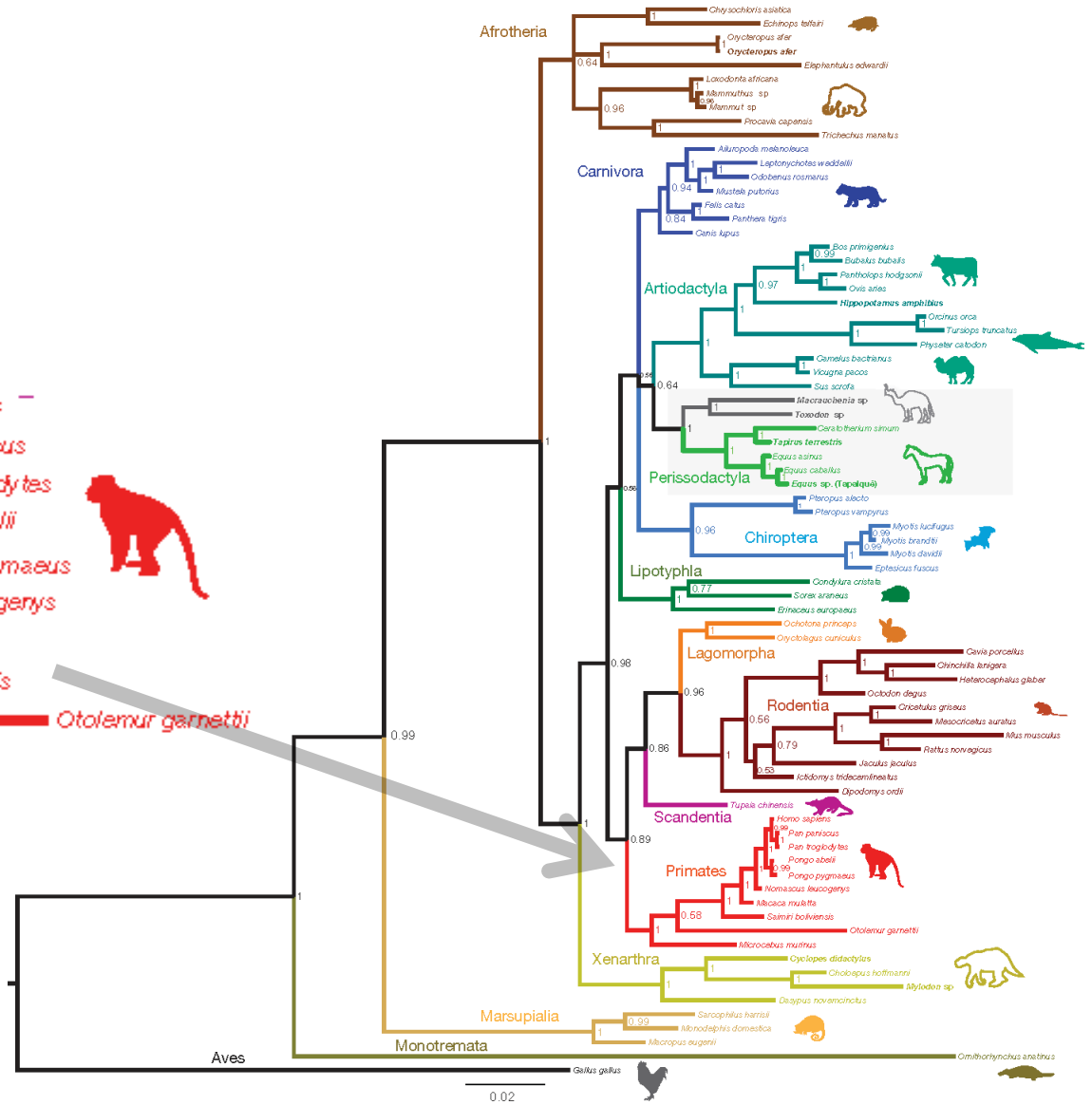
- generate 1000 trees
- for each sub-tree
  - see how often it is present

- example from nature

Monster calculation

- Took a long time
- look at this number





Welker, F., ... MacPhee, D.E. Nature, 522, 81-84 (2015) Ancient proteins resolve the evolutionary history...

DNA or protein sequences ?

The issues

- regulatory regions, RNA genes
- synonymous mutations (common – only seen in DNA)
- non-synonymous mutations (amino acid changes)
  - more information D⇌E, I⇌L⇌V, ..

Alignment reliability

- proteins
  - uses codon structure (implicitly)
  - better, amino acid similarity, I⇌L⇌V is not bad
- DNA
  - less information

DNA or protein sequences ?

	protein	DNA	time
synonymous changes	no	yes	short
a.a. changes	yes	no	longer
frame shifts	no	yes	
non-coding regions	no	yes	

Very short time

- use DNA

Longer time

- use proteins

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 1000s of sequences
- more complicated methods – Frau Dobler's Vorlesungen

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

Sequence and structure similarity

Claim from before

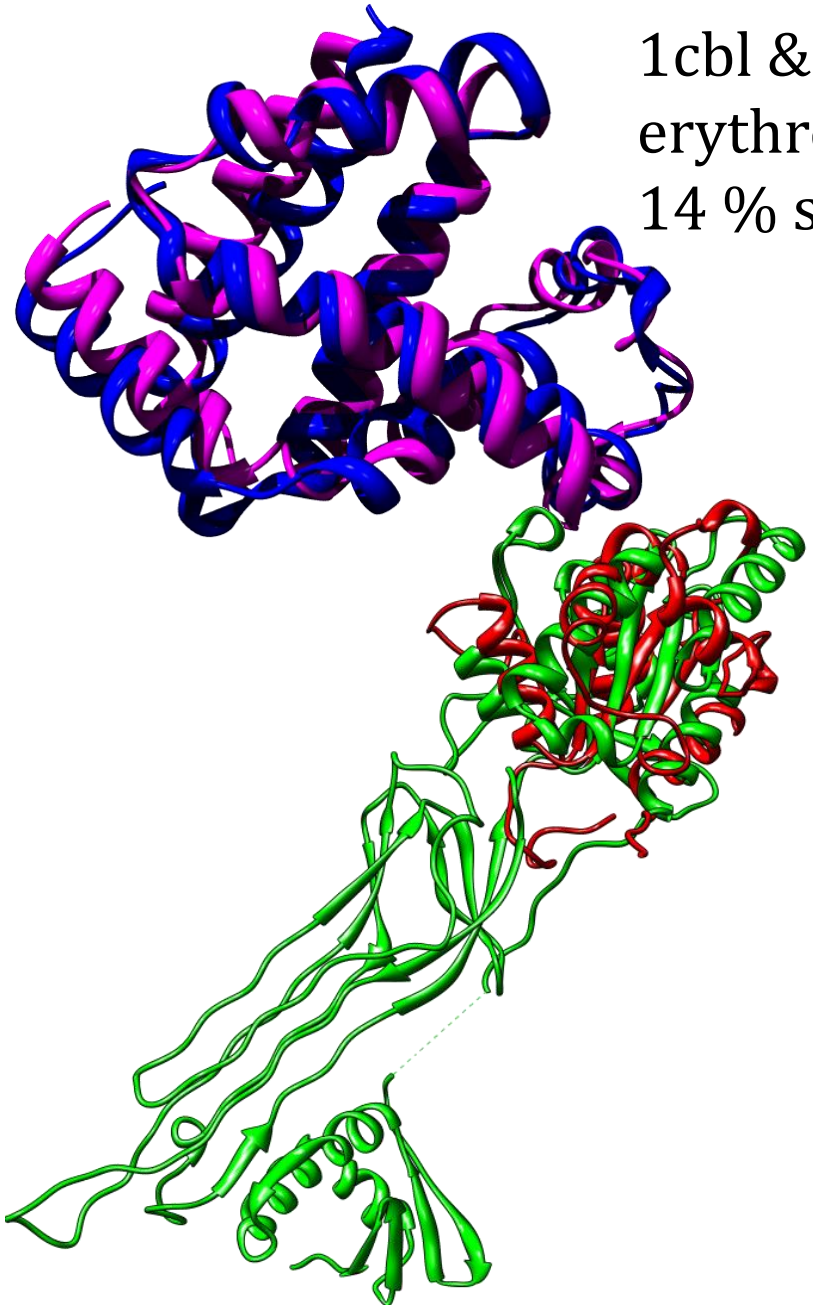
- if two sequences are similar
  - they are related – structures are similar

Question

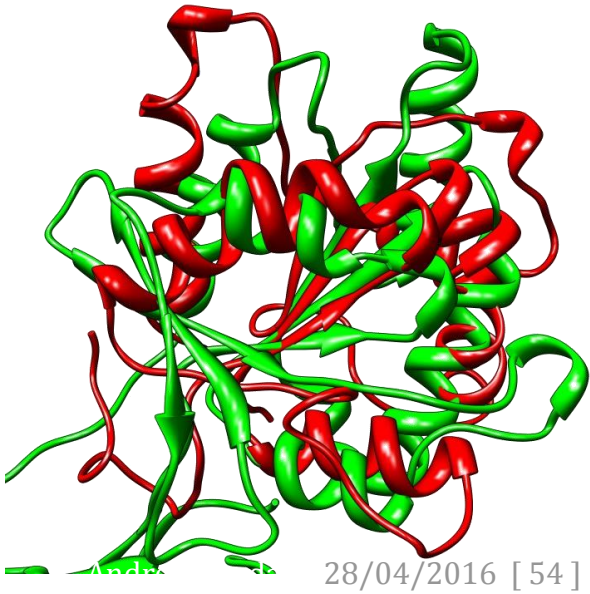
- if two sequences are different
  - are their structures different ?

Remote similarities

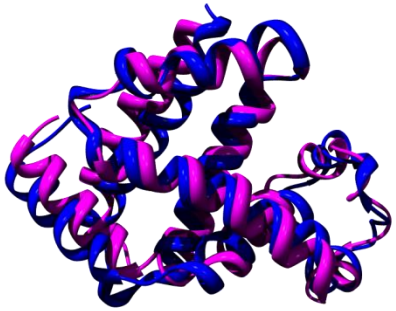
1cbl & 1eca (haemoglobin & erythrocruorin)
14 % sequence id



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



No sequence similarity – similar structures

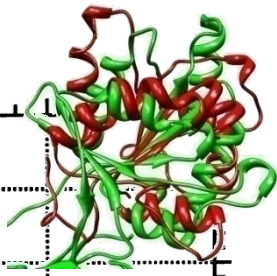


Are these rare ?

- 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

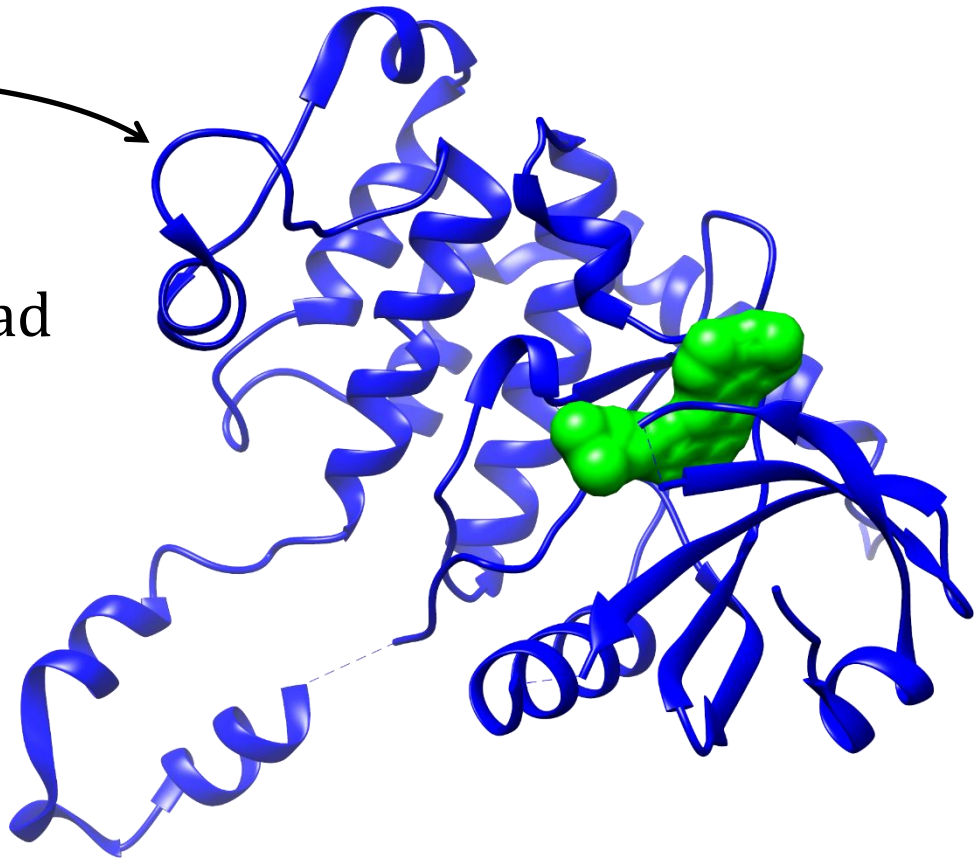
Why can sequence change

View of molecular evolution…

change here

residue changes ? OK

structure changes ? Bad

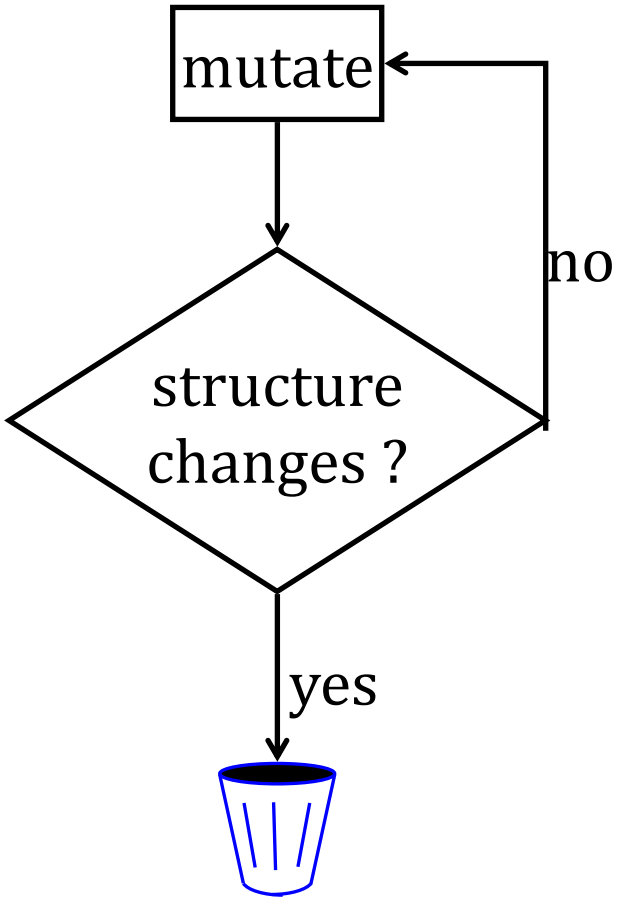


2wtj (transferase) + inhibitor

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

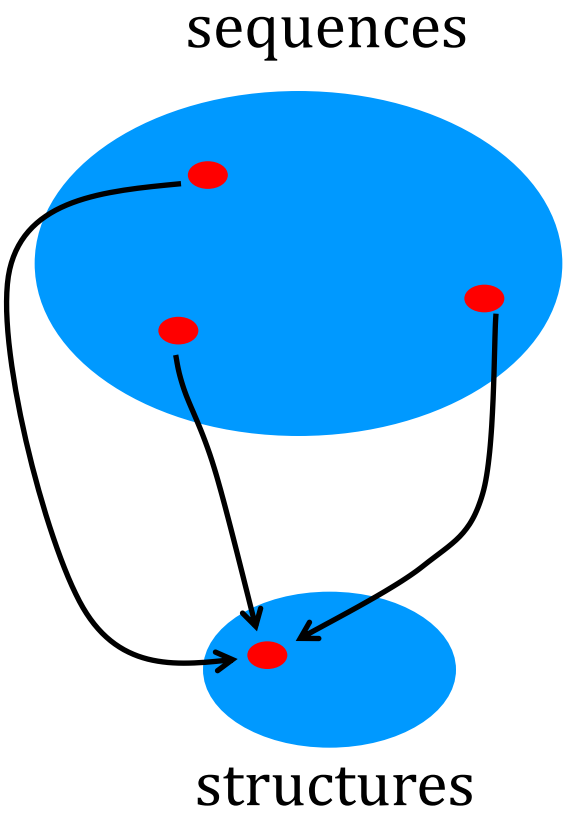
- evolution will find many sequences
  - compatible with structure
  - compatible with function
- how else would we see this ?

Sequence vs structure evolution

Sequence and structure space

- sequence space is larger
  - many different sequences map to similar structure

- sequence evolves faster than structure



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

Sequence vs structure similarity

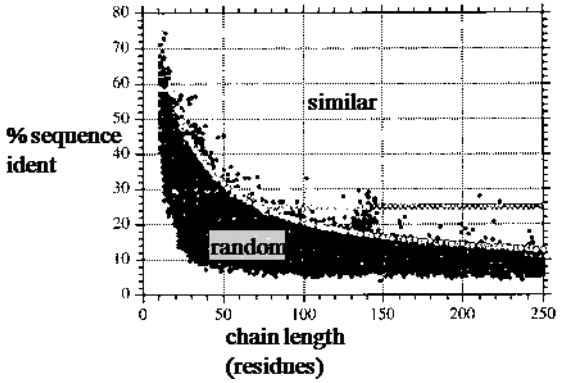
When comparing proteins

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 ?
  - coming soon

practical consequences …

Little summary

Multiple sequence alignments

- for conservation
- first step to phylogenies

Phylogenies

- not as reliable as the pictures imply

Structure vs sequence evolution

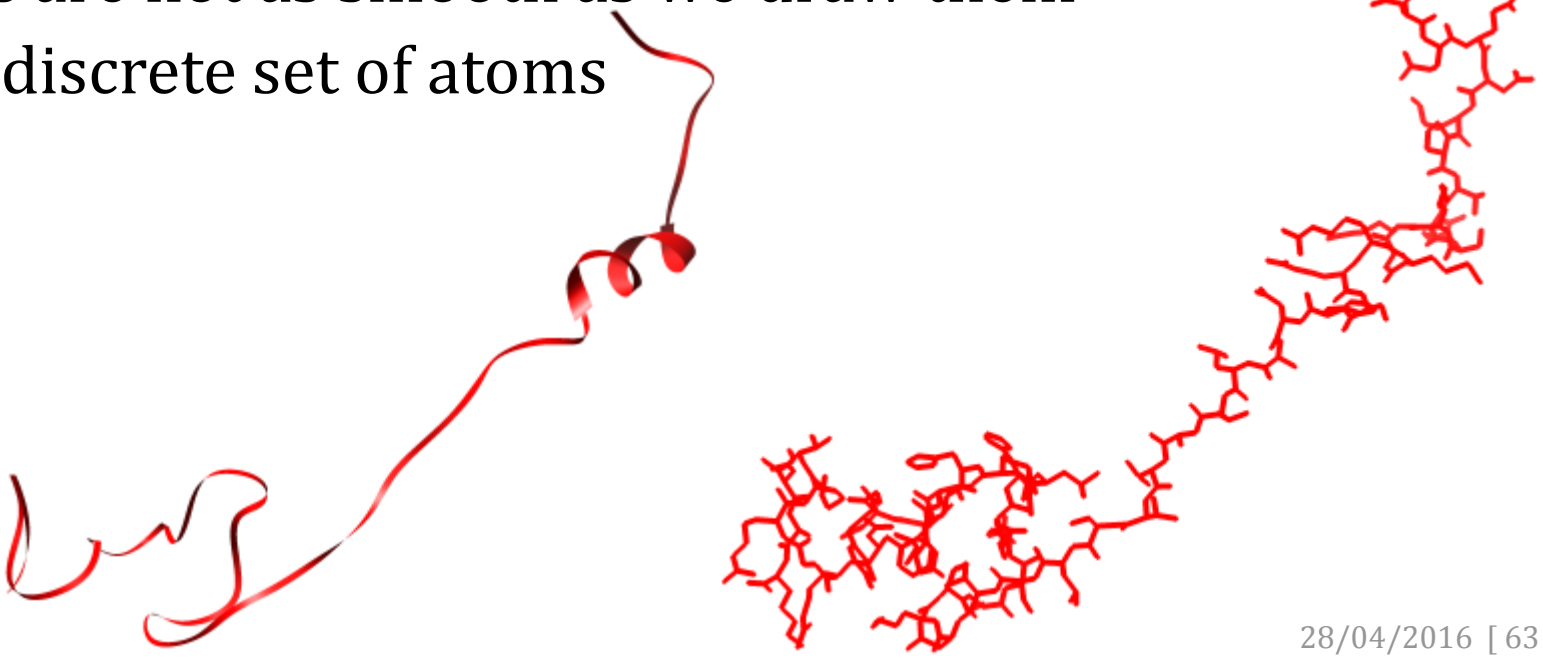
- sequence changes faster
- sequence similarity means a closer evolutionary relationship
  - functional similarity

Comparing structures

- what are protein coordinates ?
- comparison

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
- ≈ 1.2 ×10^5 structures
  - huge redundancy (> 500 T4 lysozyme)

- X-ray crystallography ≈ 85 %
- NMR ≈ 14 % (more in smaller proteins)

Protein Databank – biased ?

10^5 structures – a good summary of the world ?

Maybe not

1. Chemistry – we see proteins

- that could be crystallised
- that can be expressed
- protein 2 is easier if protein 1 has a known structure

2. Sociology

- human proteins
- "model organisms"
- disease-causing proteins / therapeutic targets

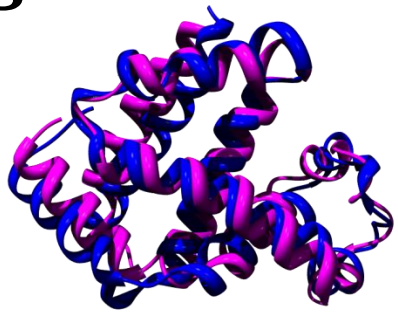
the bias

Lots of proteins that are

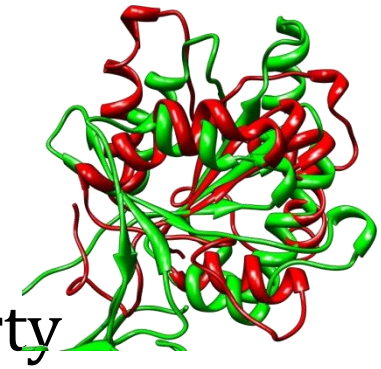
- smaller
- soluble / globular (few membrane bound)
- stable / can be handled in laboratory
- not toxic – can be expressed
- similar to ones in the databank
- from humans and special projects

Comparing coordinates

These are very similar



These are clearly related, less similar

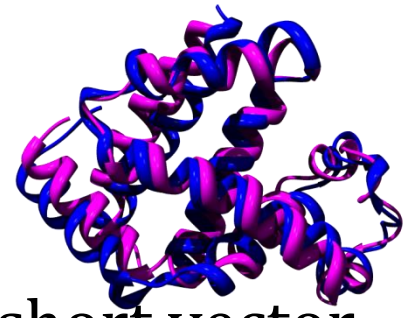


- we want to put numbers on this property

First some notation

- We have spoken of x, y, z coordinates. Easier..
  - vector $\vec{r}$ or $\vec{r}_i$ for atom 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

- 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

$$\sigma_x = \left(\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 \right)^{\frac{1}{2}}$$

Something similar for coordinates

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

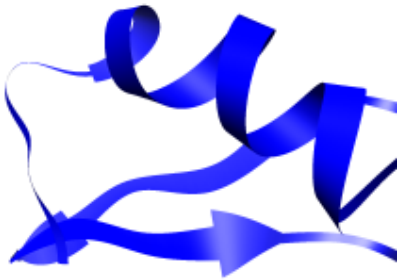
- where proteins a and b have N_{res} residues
- rmsd is “root mean square difference”

- complications

Before calculating rmsd

Two very similar proteins

- coordinates are in different orientations
- not on top of each other



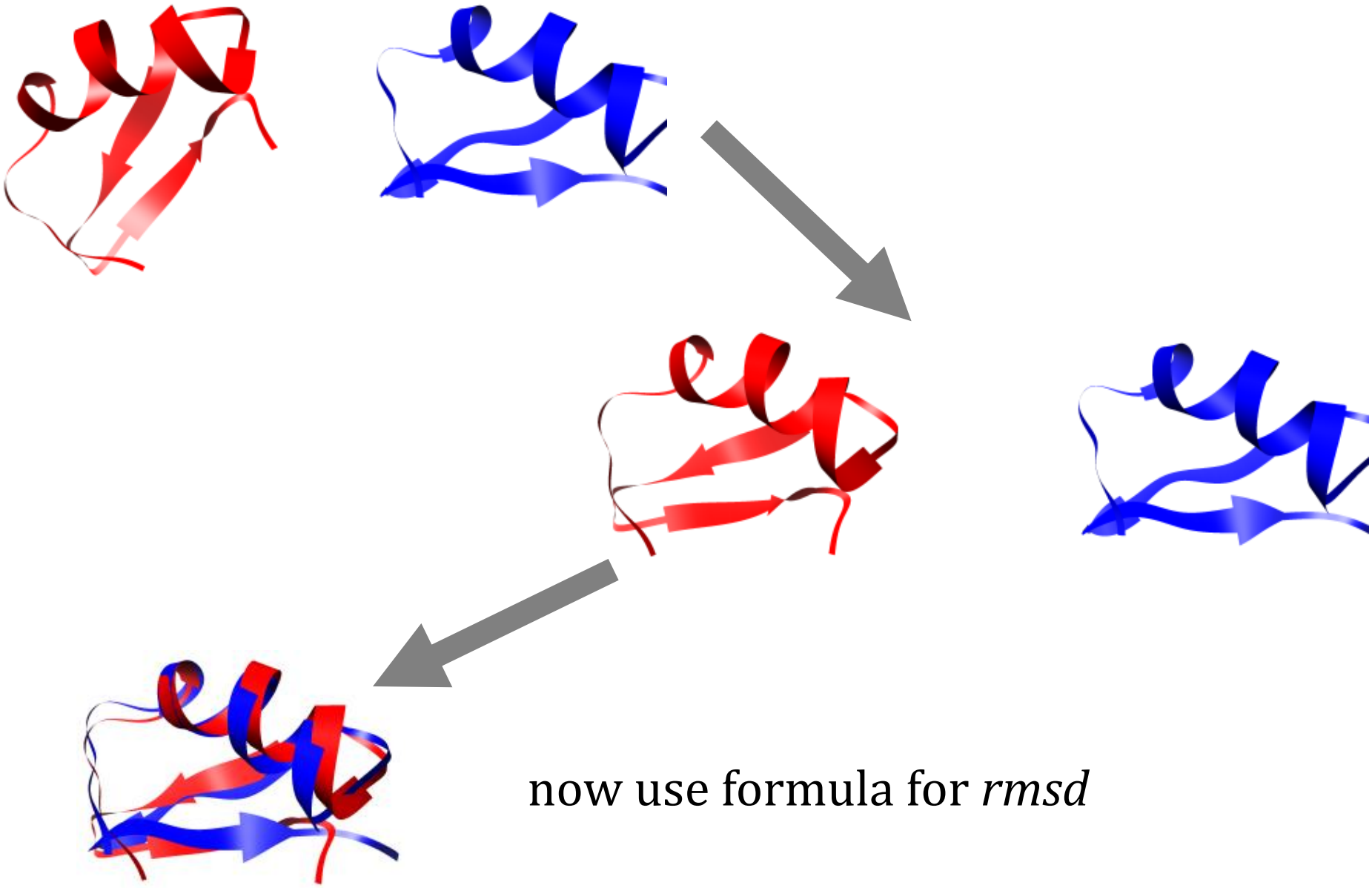
What are the orientations of files in PDB ?

- totally arbitrary

Consequence

- put the proteins on top of each other

Superposition of coordinates



now use formula for *rmsd*

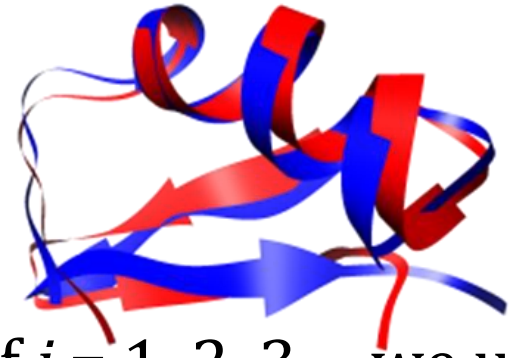
Meaning of *rmsd*

Before calculating *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

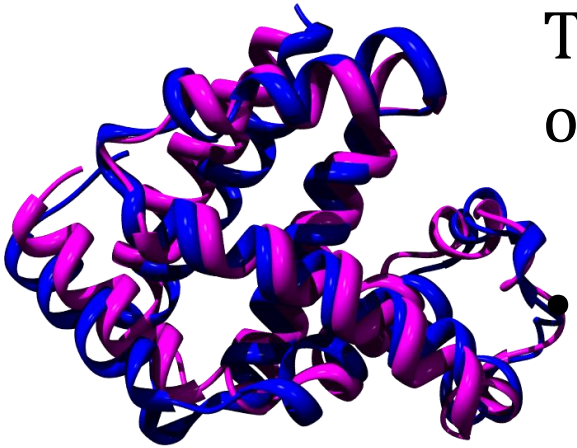
Difficulty with *rmsd*

These two proteins have the same number of residues



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

If i = 1, 2, 3, .. we use residue 1, 2, 3 in both proteins

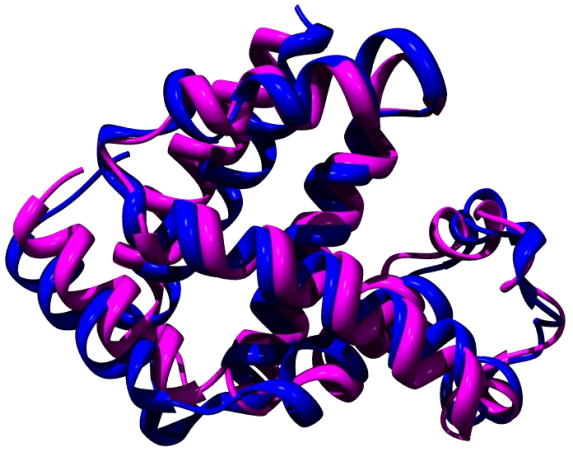


These two proteins have different numbers of residues

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

rmsd different sized proteins

- make a list of residues in each protein
- just work with corresponding residues (amazingly difficult)



rmsd summary

- formula – bit like standard deviation
- needs translation and rotation
- size dependent (just remember)
- difficult for proteins of different sizes

Summary of comparisons

Closely related proteins

- sequence alignments
- multiple sequence alignments
  - conservation, phylogeny

Structures

- see similarity even when sequences are very different
- necessary for drug design / Wirkstoffentwurf
- explaining patterns of conservation
- predictions of which residues are near/far from active site