

Classifying and comparing proteins

Plan

- why ?
- classifications
 - hierarchical vs pragmatic / empirical
 - continuous or clustered ?
- sequence similarity vs structure similarity
- example classifications
- comparison measures

Why ?

- Background – details later
 - evolutionarily close proteins - similar structures
 - evolutionarily remote proteins - may have similar structures
- function prediction / annotation
- interpretation
- structure prediction
 - examples

Function prediction (annotation)

Most common question

- gene (sequence) associated with disease – no idea of chemistry
- look for related sequences with known (annotated) function

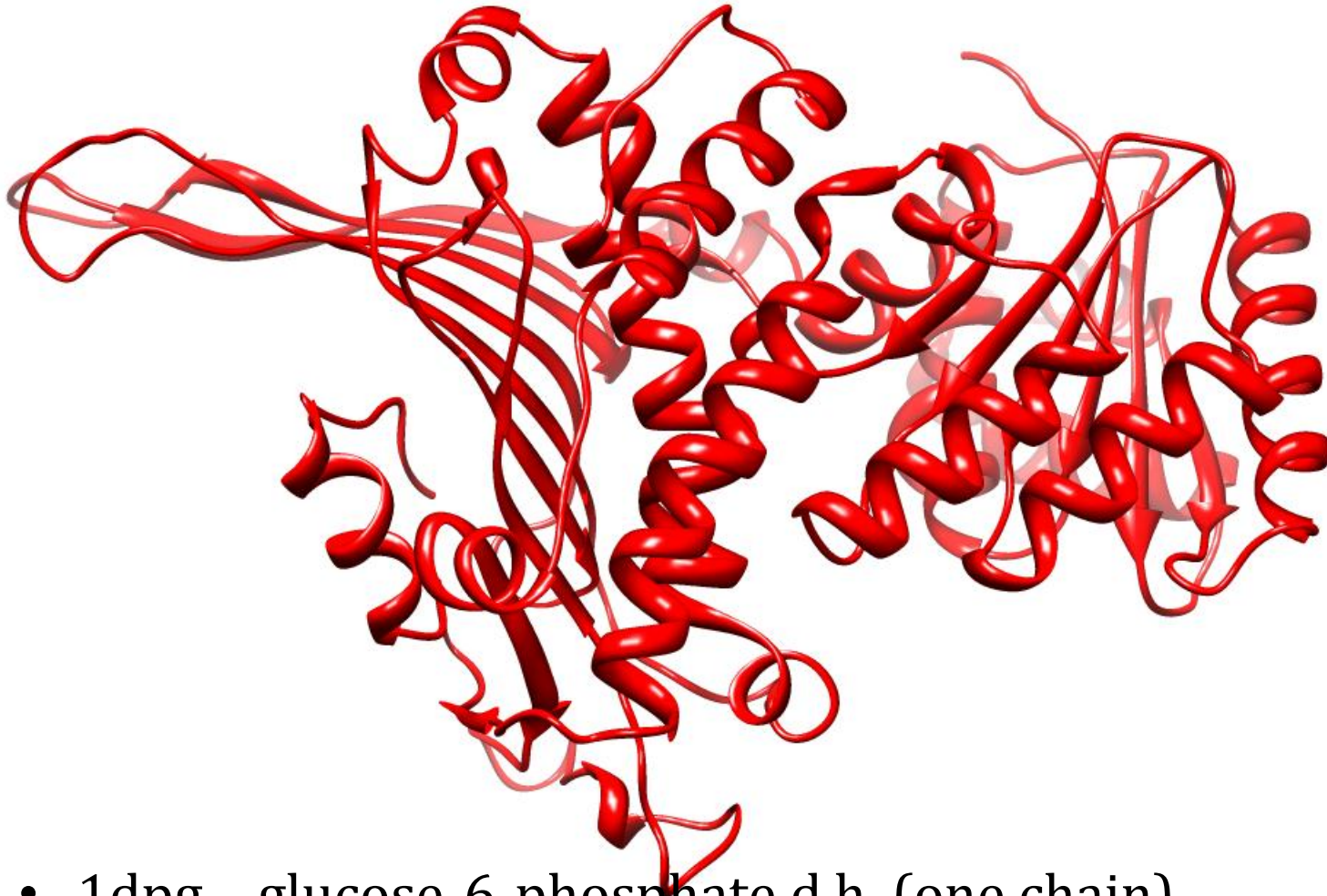
No answer ?

- structure available ? yes ..
- look for related structures in protein data bank with known function

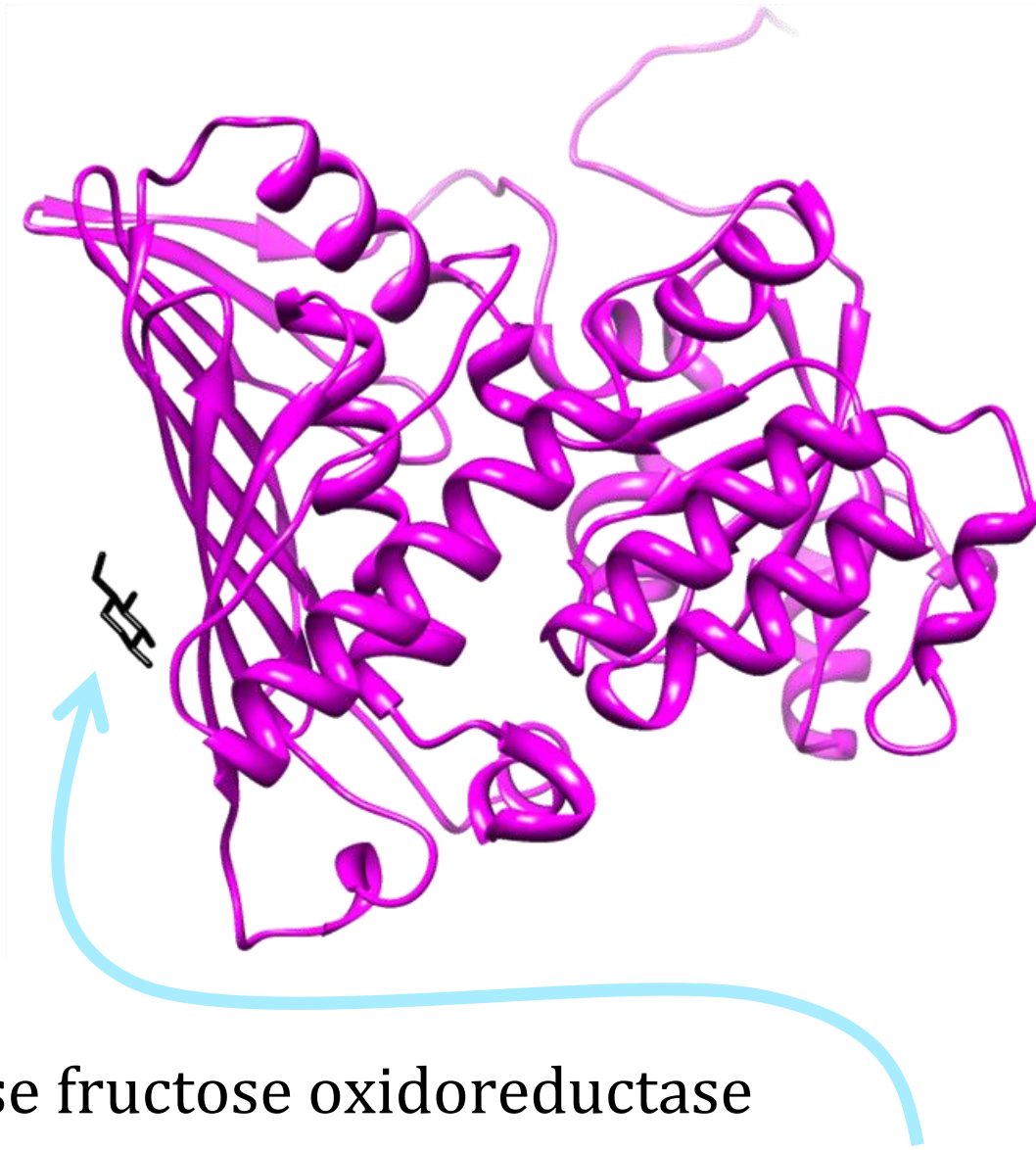
Interpretation of structures

- you know what your protein does
- you cannot crystallise it with reactants (substrates)
- you cannot see which residues are essential to function
- find a related structure which crystallises with its (maybe different) reactants
- example
 - 1dpg – oxidoreductase
 - acts on sugar, no idea where sugar binds

where does sugar bind in 1dpg ?



- 1dpg – glucose-6-phosphate d.h. (one chain)
- no idea where sugar binds / which residues are important
- well studied – never crystallised with sugar

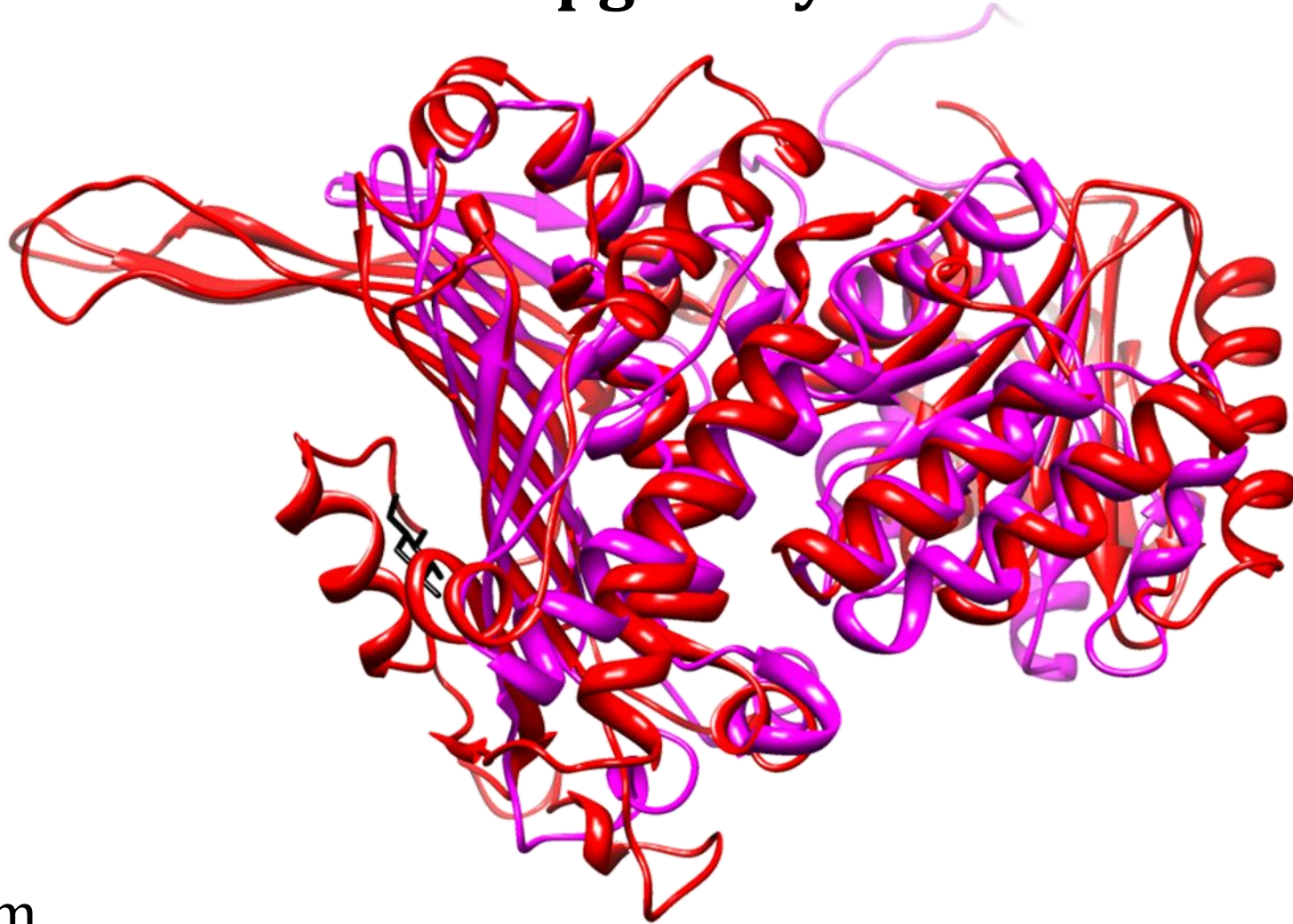


1ryd – glucose fructose oxidoreductase

- special – managed to crystallise with sugar

Transfer the reactant location...

1dpg & 1ryd



claim

- from structural similarity one knows which residues in 1dpg are important

Classification and structure prediction

How many possible protein structures are there ?

- astronomical

How many protein structures actually occur on earth ?

- 2×10^3 to 5×10^3

de novo / ab initio prediction ?

- search in giant space

Find most likely protein fold ?

- search amongst 10^3 to 10^4 structures
- find the class of your protein - crude structure prediction

Sequence vs structure similarity

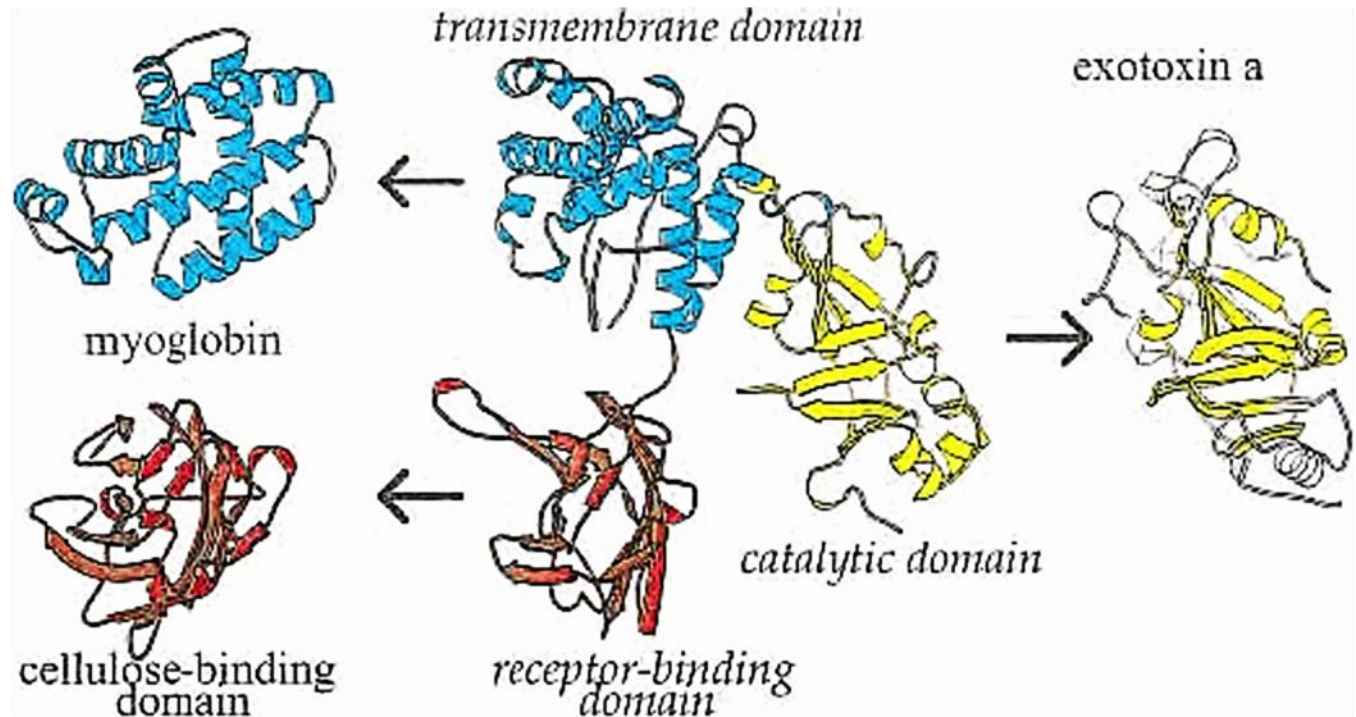
- Protein Databank $\approx 8.7 \times 10^4$
- 90 % sequence similarity $\approx 3.3 \times 10^4$ classes
- different shapes 2 to 5×10^3

- fewer classes when structure-based
- structure-based classes are larger

- speculations and explanations later
- now
 - domains
 - sequence space
 - hierarchical and non-hierarchical

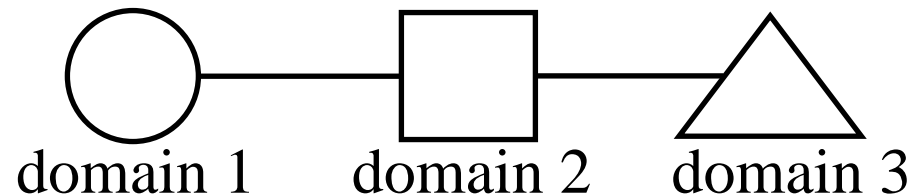
Domains

- Why mention ?
 - many groups work on domains, not whole proteins
- Reasons
 - many structures are labelled "a domain of protein X"
 - evolution – convincing picture (diphtheria toxin)



Domains – evolutionary viewpoint

- idealised view..
- claim / belief
 - evolution goes faster by mixing / swapping domains between proteins
- do we all agree on domains ?
 - 3 viewpoints

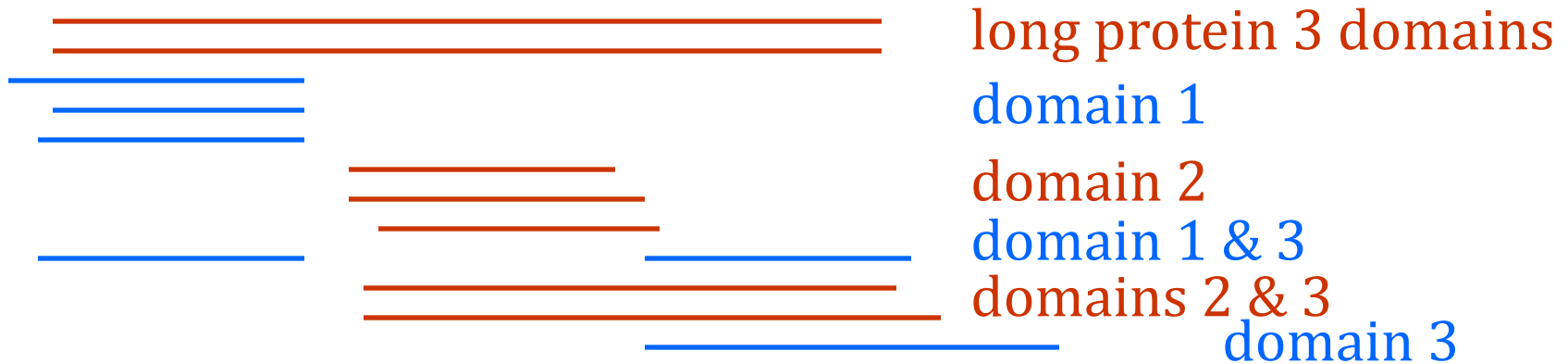


Domains in Biochemistry (view 1)

- Domains = parts of a protein with different functions
 - catalytic domain, regulator-binding domain...

Sequence level domains (view 2)

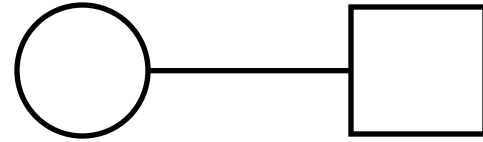
- Align a group of sequences



- appears to have 3 domains
- no reference to structures or function

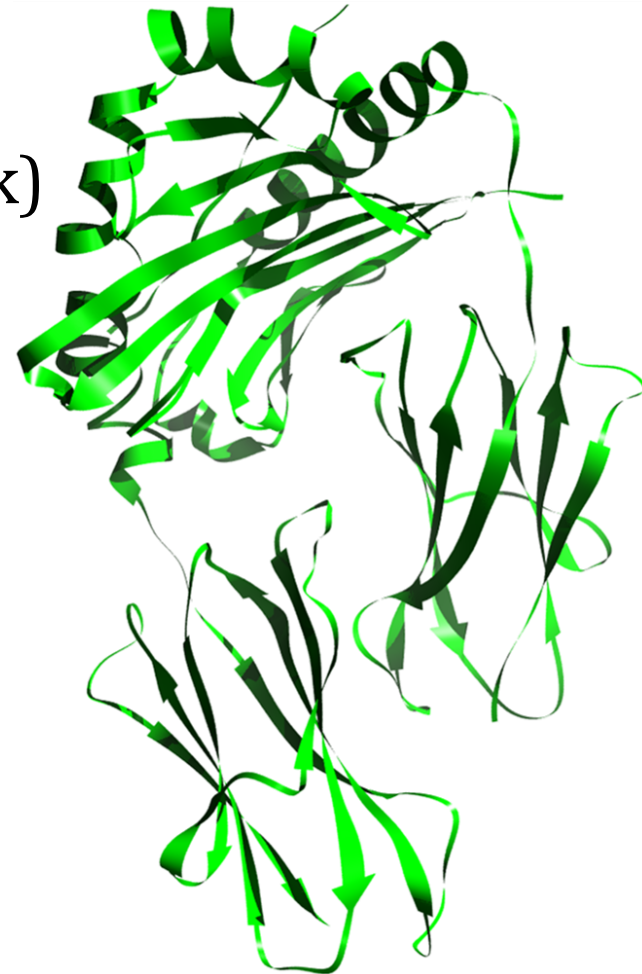
Domains in Structures (view 3)

- Many structures solved look like...



example: histocompatibility module (1iak)

- 3 domains

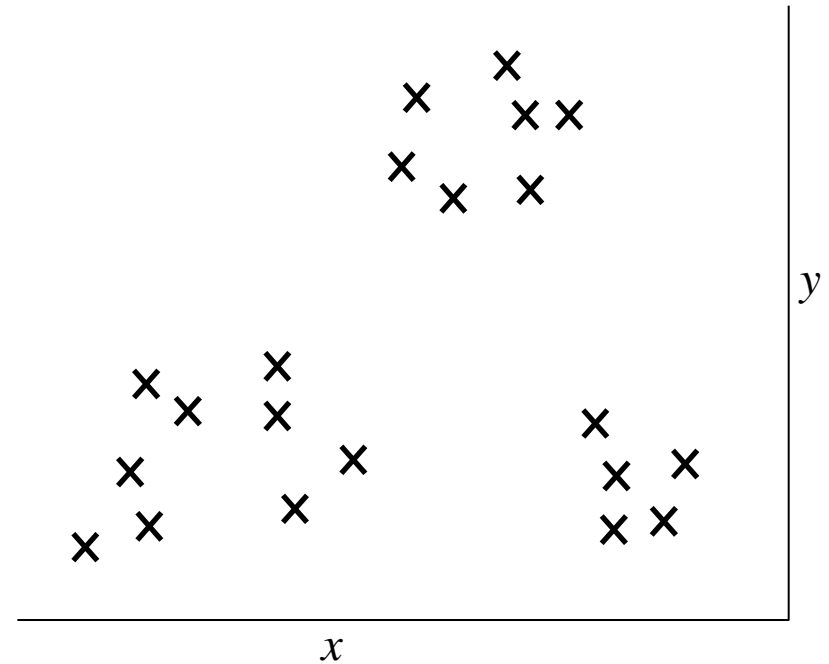


Domains for these lectures

- usually structure based
 - compact units
 - stable in solution (usually)
- can we really expect to classify proteins ?
- Next set of slides - one question
 - what is "protein space" or what would it mean ?

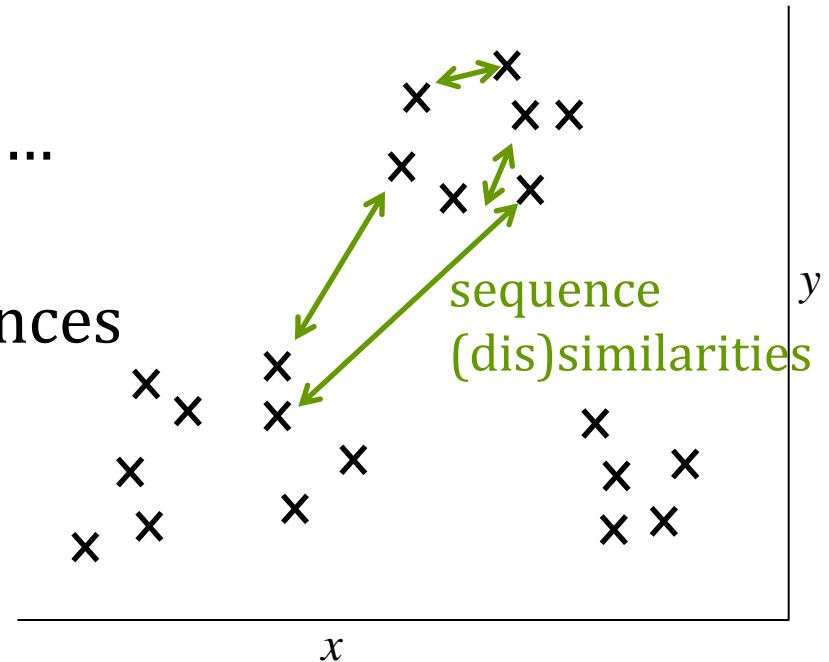
Protein classes / families

- questions
 - what do they mean ?
 - do you expect them ?
- meaning...
- each cross is a protein
 - what are x and y ?
 - two ways to answer
 - generic n -dimensional distances
 - example from sequence space



Spaces for proteins

- Sequence example
 - we can compare any two sequences
 - measure (dis)similarity
 - matches, similarity score, ...
- I have a matrix of $n(n - 1)$ distances
 - how would I go to x, y ?
 - how many dimensions?
- If I have similarities between objects
 - there is some implied $(n - 1)$ dimensional space
- a different way to have a sequence space



Sequence Space

- convenient way to explain ideas of sequence similarity
- conventional spaces
 - 1D (x), 2D (x, y), 3D (x, y, z), 4D (x, y, z, w), ...
 - let us estimate how big a space or problem is
 - some problem – discrete space
 - how many variables do I have ? (a, b, c, ...)
 - how many values can each variable have ?
 - a 3 values, b 4 values, c 5
 - number of points in space = $3 \times 4 \times 5$
- protein sequences
 - each position can have 1 of 20 values
 - total number of sequences = $20 \times 20 \times \dots = 20^{N_{res}}$
 - like a space of N_{res} dimensions

Representing a Sequence

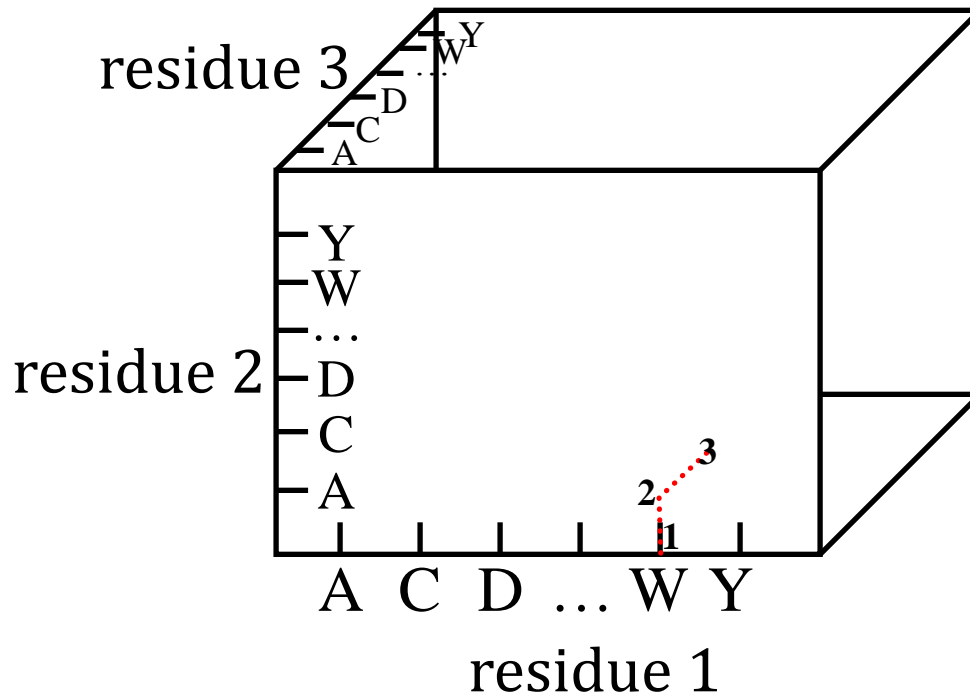
- protein sequence and structural coordinates

	1	2	3	4	5	6	7	...	N_{res}
x	1.2	2.3	...						10.3
y	2.4	3.5	...						11.1
z	1.7	2.9	...						15.5
seq	W	A	C	A	A	...			D

- consider the first three residues
 - WAC (for pictures only)

Finding a Sequence in This Space

- real diagram is a box of N_{res} dimensions
 - this one 3 dimensions

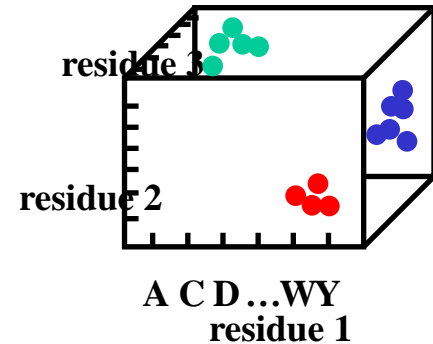


	1	2	3	4	5	6	7	...	N_{res}
x	1.2	2.3	...						10.3
y	2.4	3.5	...						11.1
z	1.7	2.9	...						15.5
seq	W	A	C	A	A	...			D

- looking for sequences...

Families in Sequence Space

- Similar sequences should land near each other
- How realistic ?
 - picture is a simplification
 - only works for $N_{seq1} = N_{seq2}$
 - very useful
 - distances between sequences
- Will return next semester

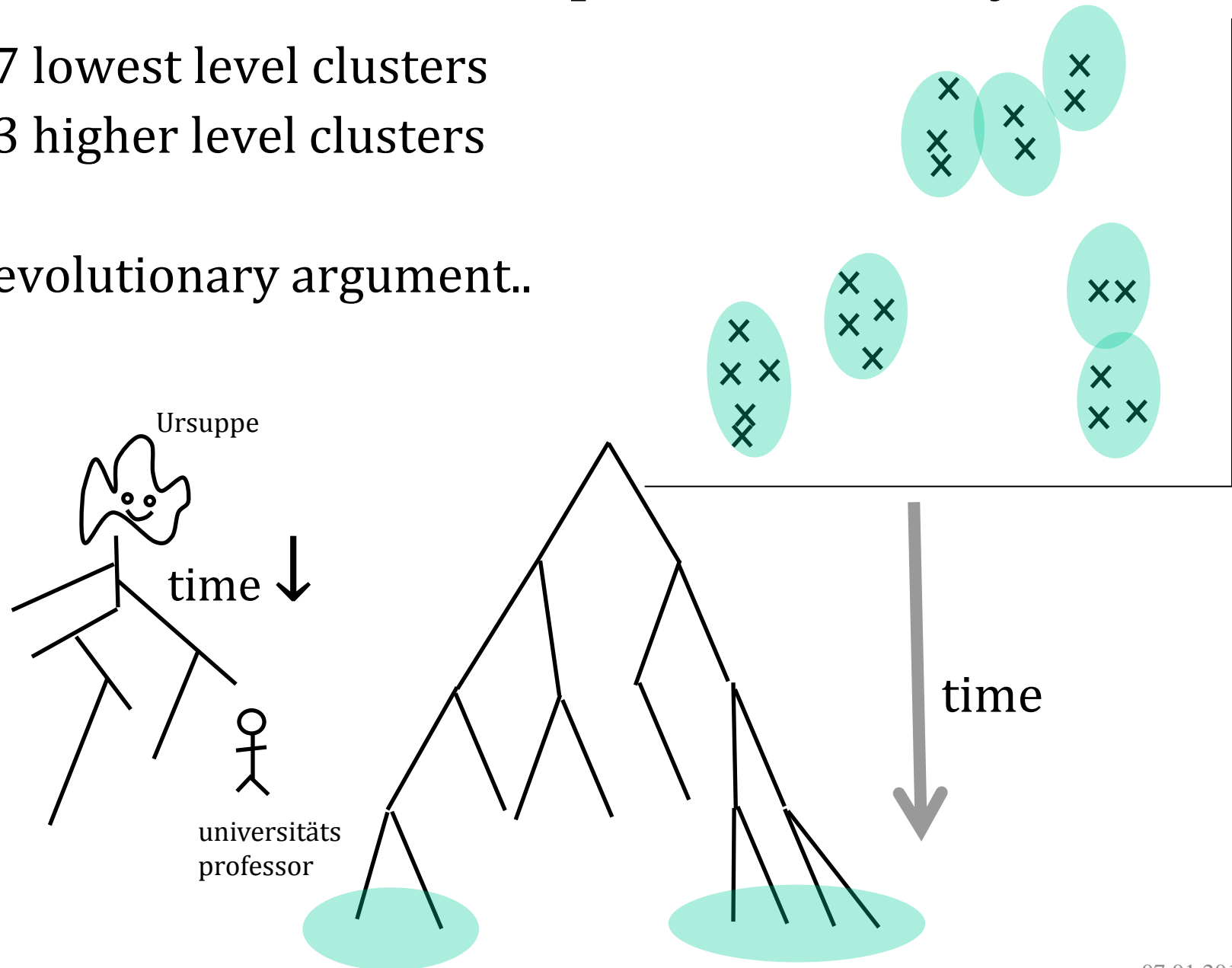


Summarise spaces

- we can imagine a space of proteins
 1. from similarities between points $(n-1)$ dimensional
 2. sequence space N_{res} dimensional

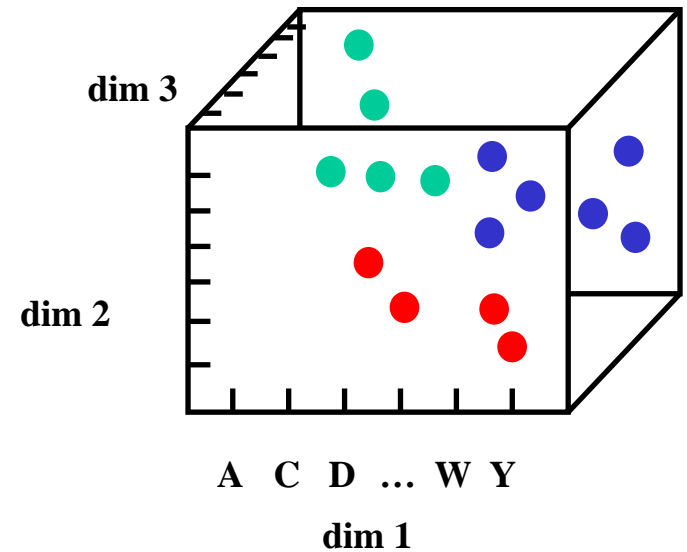
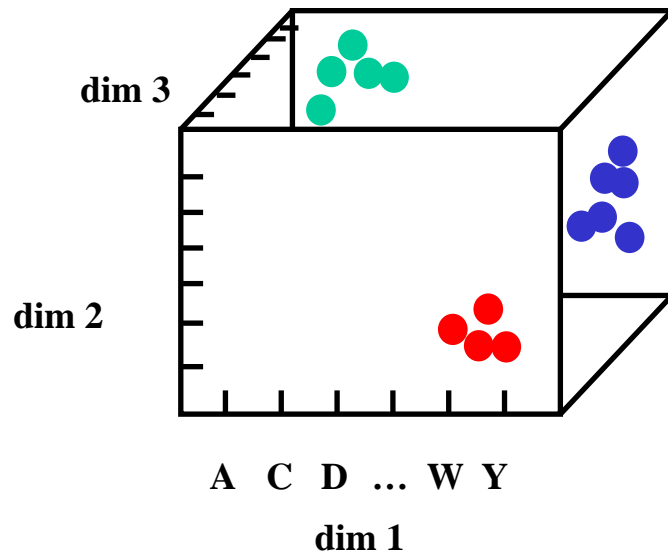
Should we expect a hierarchy ?

- 7 lowest level clusters
- 3 higher level clusters
- evolutionary argument..



Do we expect protein families ?

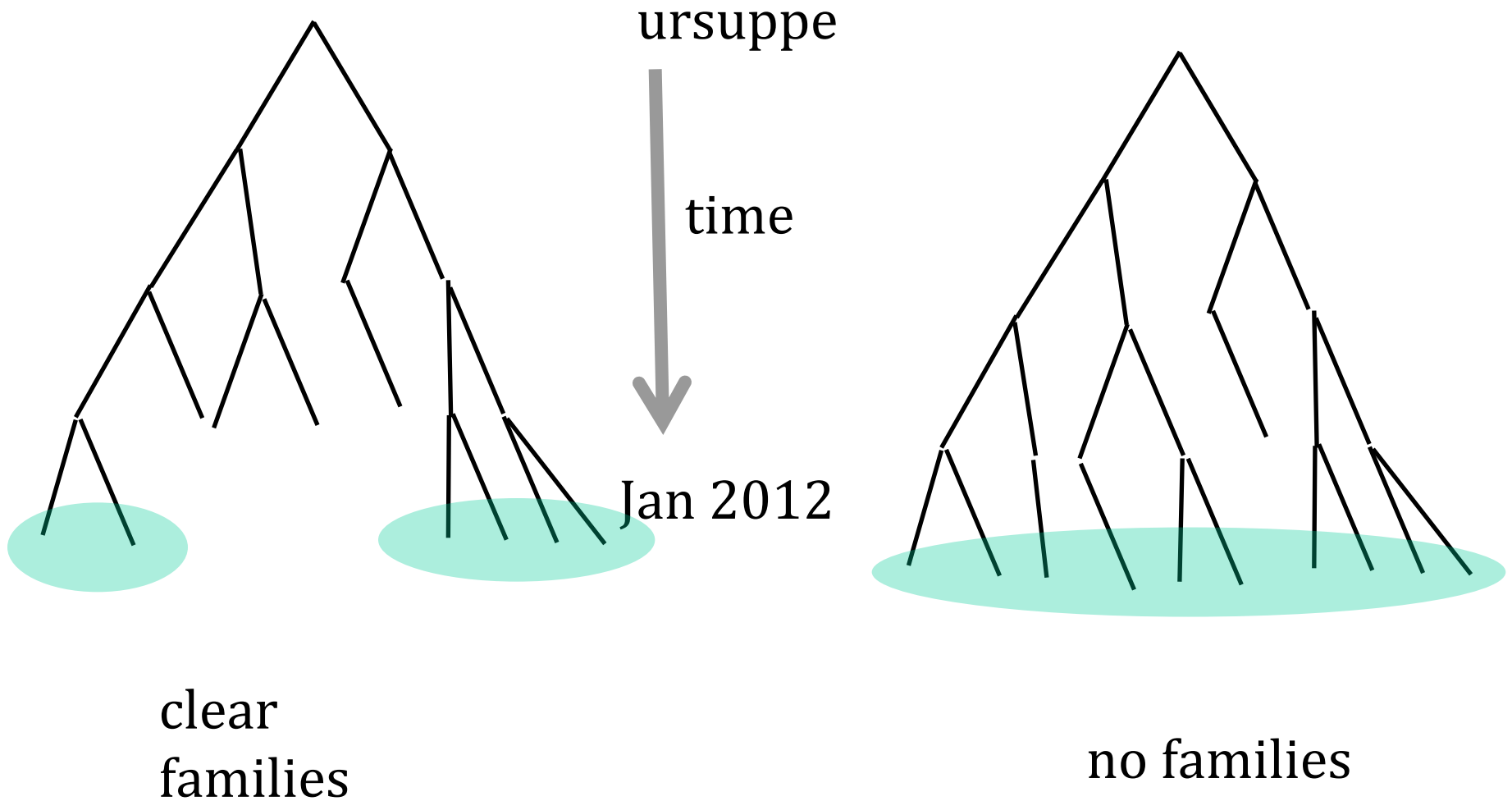
- No real answer
- we have an idea of spaces – sequence or structure based
- how are proteins distributed ?



- should you expect clusters ?

Evolution and phylogeny

- shape / density of tree of life



Questions for fun

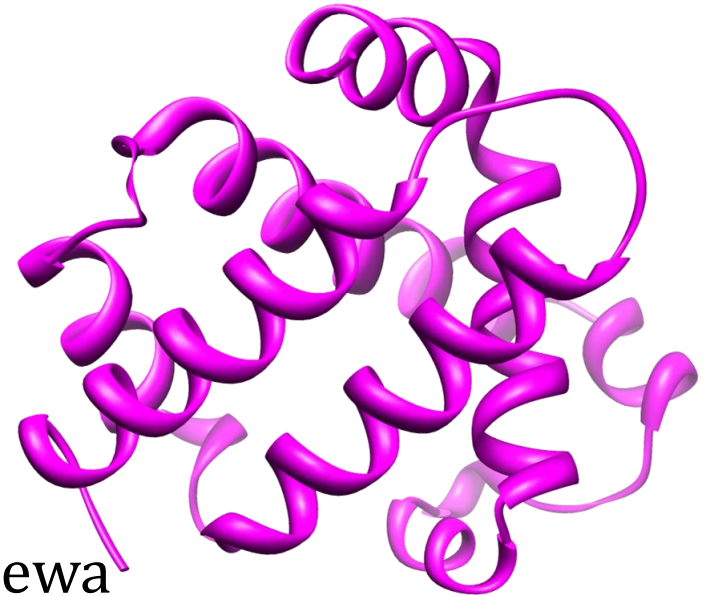
- Do we expect hierarchy ?
 - some people do
- Do we expect clusters
 - some people ..

Structure vs Sequence similarities

- more different than you might expect
- similar sequences
 - have not evolved for too long
 - expect similar structures
- other way round ? Examples



1ecd
erythrocyruorin



1ewa
dehaloperoxidase

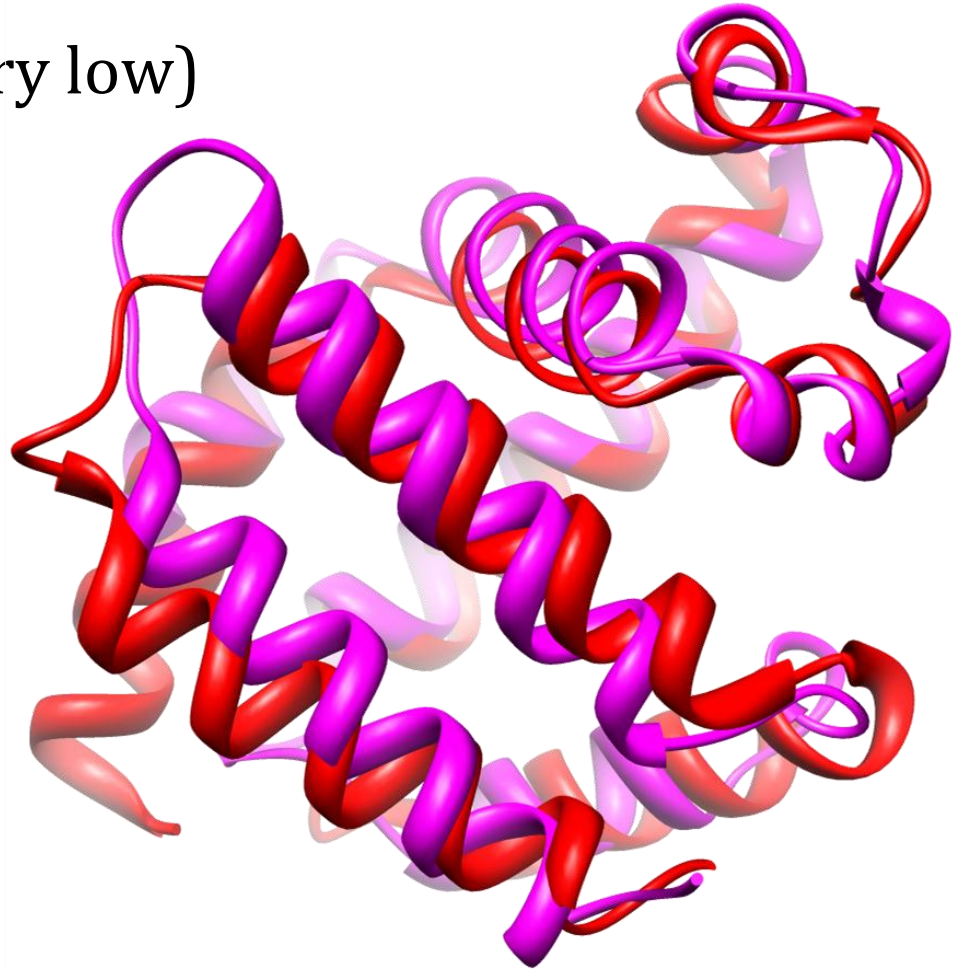
very different sequences

1ecd & 1ewa

- 17% sequence identity (very low)
- structures almost identical

Is this an exception ?

- 100's of examples
- totally normal
- play with our server



<http://public.zbh.uni-hamburg.de/salami/>

Example family

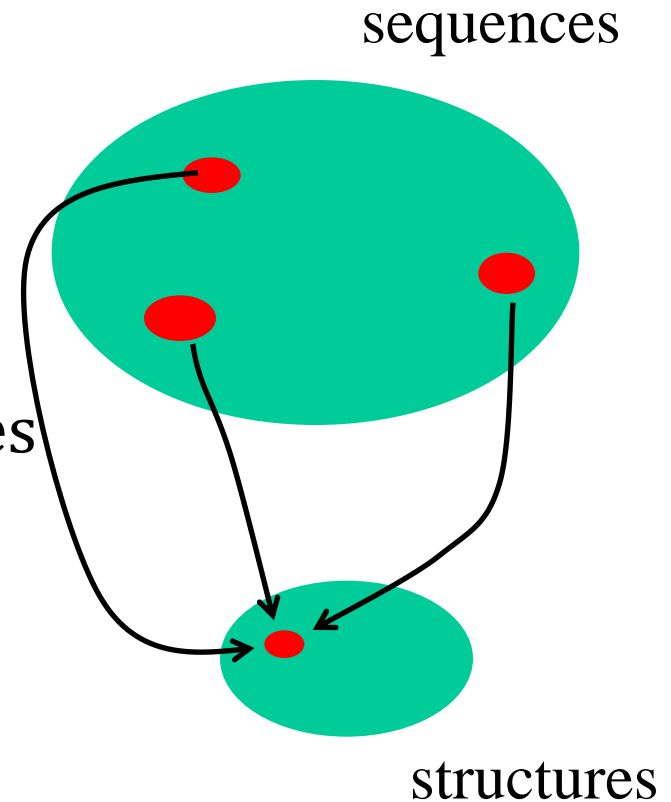
Example, neighbours of 1cun chain A

- look at sequence identity (% id)
- alignment length (lali = number of residues)
- root mean square diff in Å

No	Chain	%id	lali	rmsd	Description
1	1cunA	100	213	0.0	ALPHA SPECTRIN
2	1hciA	24	111	1.6	ALPHA-ACTININ 2
3	1ek8A	12	106	4.4	RIBOSOME RECYCLING FACTOR
4	1oxzA	9	91	2.5	ADP-RIBOSYLATION FACTOR BINDING PROTEIN GGA1
5	1eh1A	8	102	4.6	RIBOSOME RECYCLING FACTOR
6	1hx1B	5	105	3.1	HEAT SHOCK COGNATE 71 KDA
7	1dd5A	8	103	4.7	RIBOSOME RECYCLING FACTOR
8	1lvfA	9	98	2.6	SYNTAXIN 6
9	1bg1A	9	99	2.3	STAT3B
10	1hg5A	5	98	3.0	CLATHRIN ASSEMBLY PROTEIN SHORT FORM
11	1hs7A	14	92	2.5	SYNTAXIN VAM3
12	1dn1B	10	101	2.7	SYNTAXIN BINDING PROTEIN 1
13	1ge9A	6	108	4.6	RIBOSOME RECYCLING FACTOR
14	1fewA	8	125	3.5	SECOND MITOCHONDRIA-DERIVED ACTIVATOR OF
15	1qsdA	4	90	2.4	BETA-TUBULIN BINDING POST-CHAPERONIN COFACTOR
16	1e2aA	6	95	2.8	ENZYME IIA
17	1i1iP	7	95	3.3	NEUROLYSIN
18	1fioA	8	100	2.6	SSO1 PROTEIN
19	1m62A	8	81	2.8	BAG-FAMILY MOLECULAR CHAPERONE REGULATOR-4
20	1k4tA	6	147	25.8	DNA TOPOISOMERASE I

Structure vs Sequence

- there are 1000's of such families
- summarise
 - similar sequences
 - similar structures
 - very different sequences
 - similar or different structures
- why ?



Structures < Sequences... Why ?

Evolution

- many small changes
- if structure changes, function breaks, you die
- sequences change as much as possible within this constraint

chemistry

- sequence does determine structure
 - many sequences could fit structure (more next semester)

Surprising ?

- consider near universal proteins
 - 100's millions years evolution, function largely preserved

Classifying by sequence

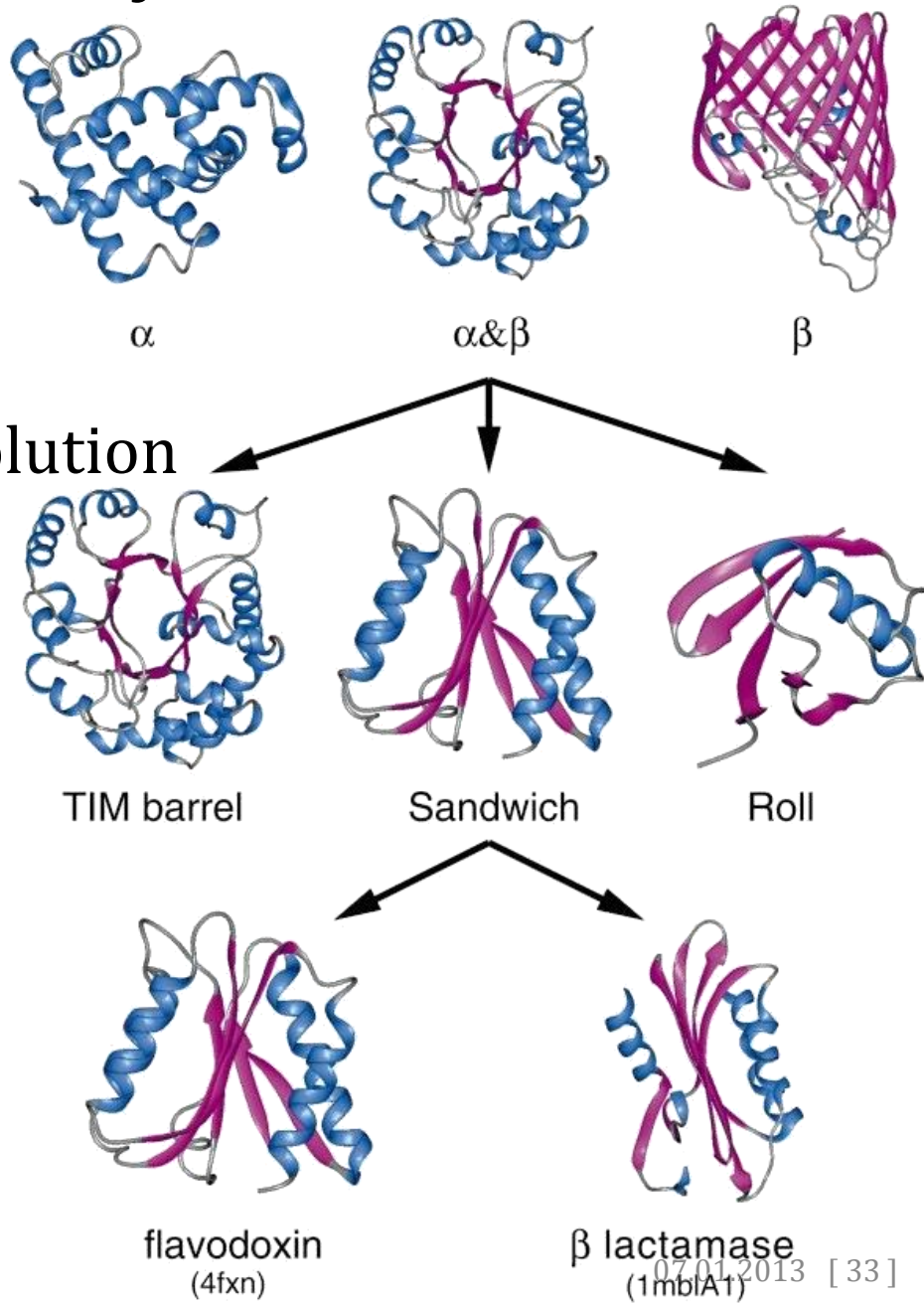
- forget hierarchy (for now)
- tools
 - any alignment program (blast, fasta, clustal, ...)
- method
 - survey all proteins in the protein databank
 - collect all pairs $> x \%$

similarity	num clusters
90 %	30 321
70%	26 171
50%	22 050

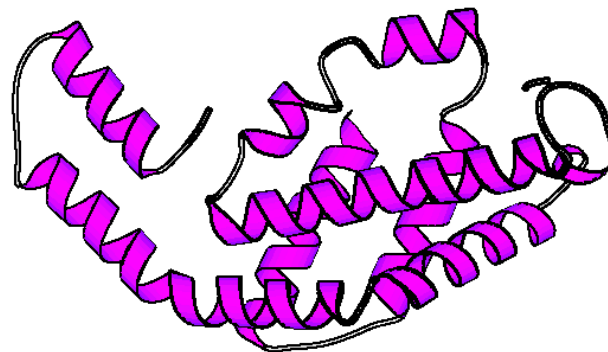
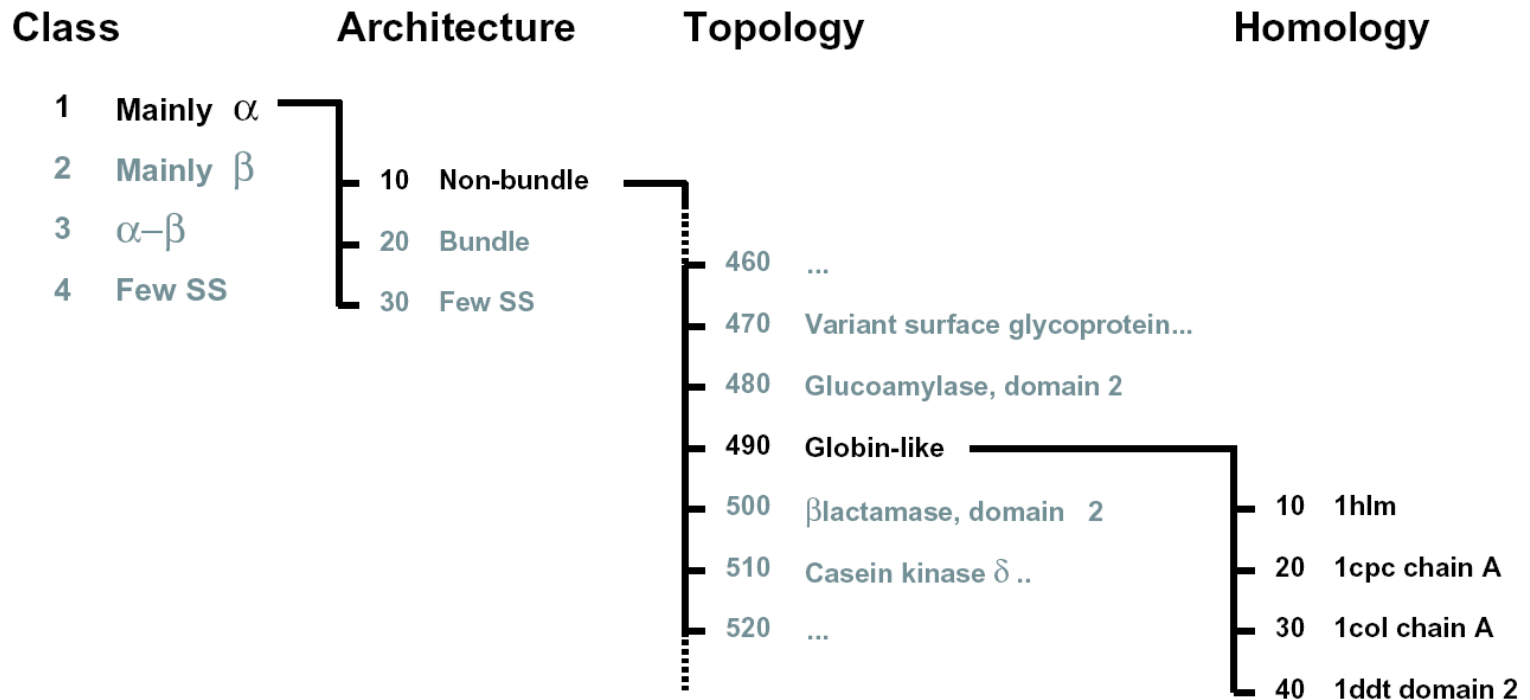
- result (jan 2013)
- how many structure classes ? 2 to 5×10^3 ?
- some sequence classes are not really different from each other
- now.. examples of structure based classifications

Imposing a Hierarchy on Proteins

- parts may correspond to evolution
- top level ?
- How useful and applicable ?
 - examples



Example from "CATH"

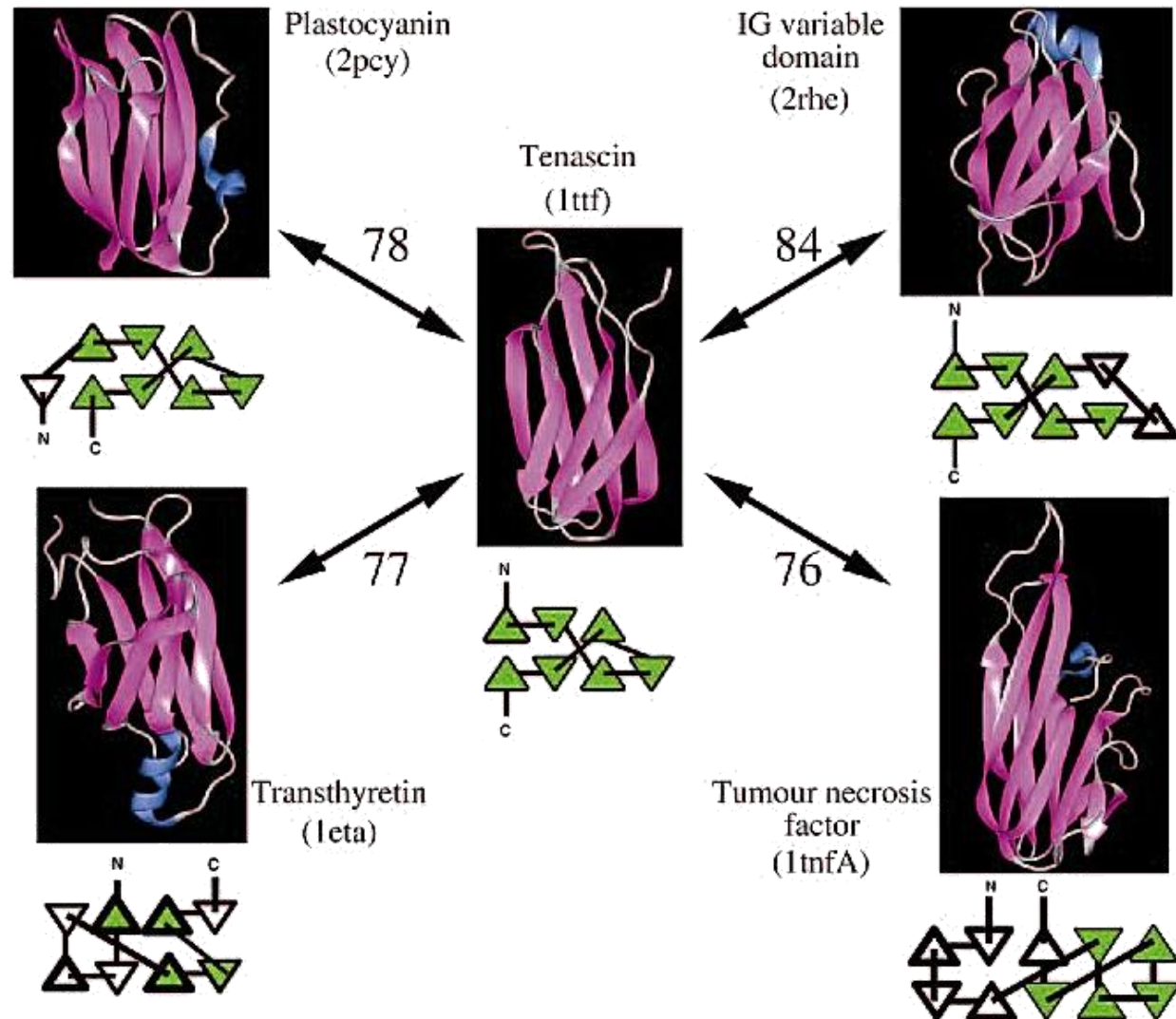


1.10.490.20

Mainly α .Non-bundle.Globin-like.1cpc chain A

Evolution and Classification

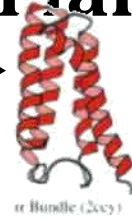
- can we interpret structures in evolutionary terms ?
- sometimes
- for more remote proteins – not really possible
- some typical figures from a literature classification



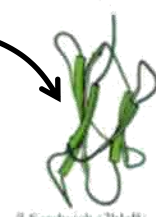
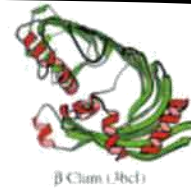
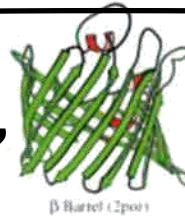
Lots of families

α -helix bundles ?

- ≈ 226 domains,
- 3 % surveyed structures

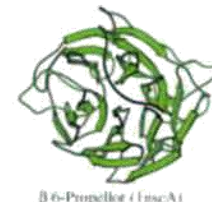
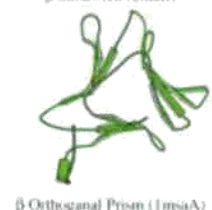
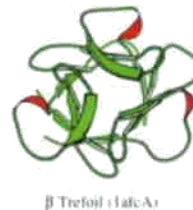
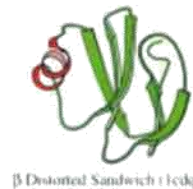


β -sandwich ≈ 1236 domains,
15 %



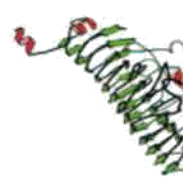
some families ?

- < 0.01 %



Interesting...

- some families very popular, some not

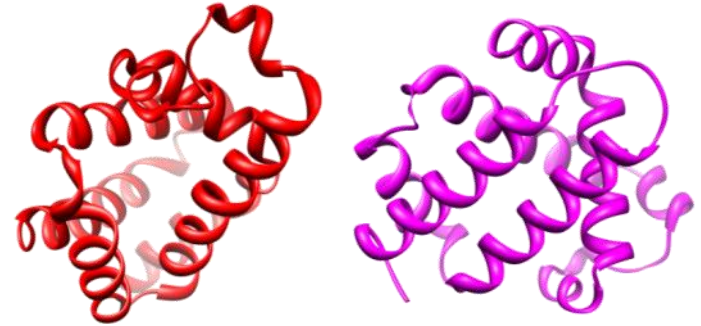


Some families populated more than others ?

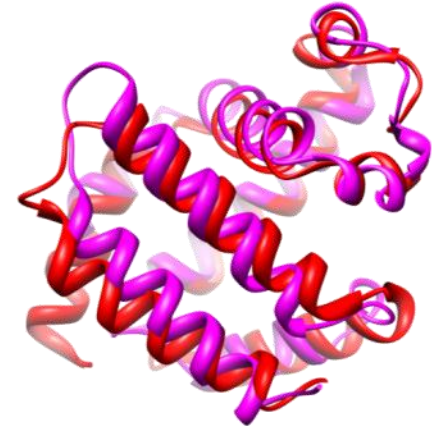
- more next semester
- are some structures more stable ?
- are some older in evolutionary terms ?
- can some "accommodate" more sequences / tolerate more mutations
- reflection of physics ?
- biases ? PDB has
 - mainly soluble, globular proteins which crystallised
 - very few membrane-bound proteins

Forget Evolution

- Is the hierarchy really justified ?
 - at low levels maybe
 - at higher levels ? ($\alpha, \alpha/\beta, ..$)



- better to discover relationships automatically
- Imagine I can compare arbitrary proteins
- have some measure of similarity
- use this to classify
- Huge problem
 - proteins are different sizes and shapes
 - how to compare ?



Summary

- Classification would be useful
- Given a distance (dissimilarity) one can invent a space for sequences or structures
- not known if it
 - exists
 - is hierarchical
- sequence vs structure similarity
 - different sequences can fold to same structure
- imposing a hierarchy on protein structures – very *ad hoc*
- one can forget hierarchy – simply use a clustering method
 - one will need a measure of similarities
 - big topic...

Protein Structure Comparison / Numerical

Most common protein structural question

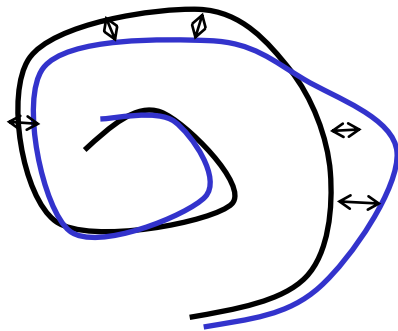
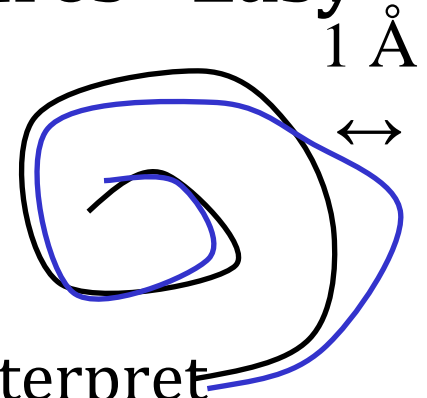
- how much has my protein moved over a simulation ?
- how similar are these NMR models for a structure ?
- how close is my model to the correct answer ?

- more difficult
 - how similar is rat to human haemoglobin ?

- two cases
 1. same protein, same number of atoms
 2. different proteins
- first
 - measures for easy cases

Numerical Comparison of Structures - Easy

- what units would we like ?
 - scale of similarity (0 to 1.0) ?
 - comparison of angles
 - distance / Å ? most common / easy to interpret
- looks a bit like the average difference between coordinates
- consider analogy with standard deviation / variance



From Standard Deviation to RMSD

Analogy with comparing a set of numbers

- get average (mean) $\bar{x} = N^{-1} \sum_{i=1}^N x_i$

- standard deviation $\sigma = \left(N^{-1} \sum_{i=1}^N (x_i - \bar{x})^2 \right)^{1/2}$

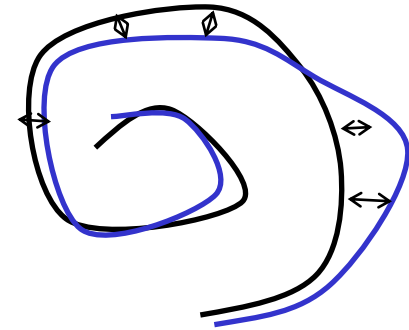
- apply this to coordinates of r and r'

$$rmsd = \left(N^{-1} \sum_{i=1}^N |\vec{r}_i - \vec{r}'_i|^2 \right)^{1/2}$$

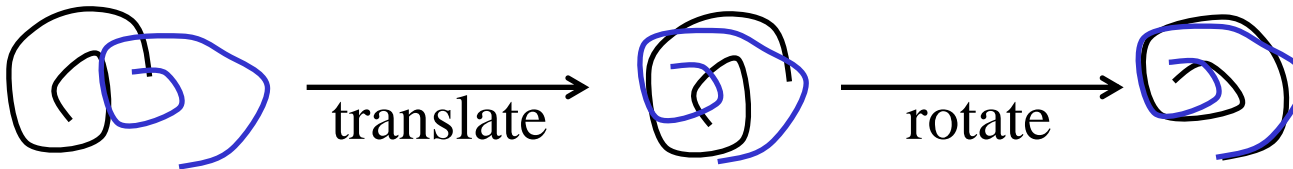
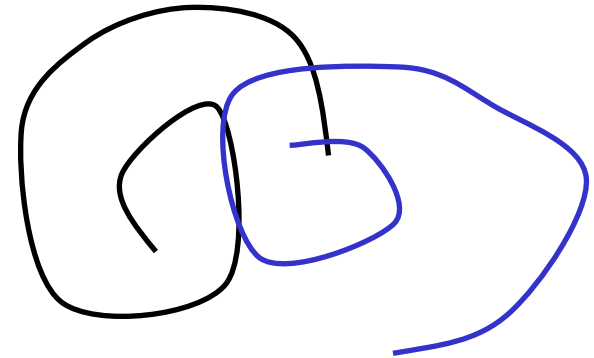
- rms / *rmsd* / RMSD = root mean square difference

Calculating rmsd

- $rmsd = \left(N^{-1} \sum_{i=1}^N |\vec{r}_i - \vec{r}_i'|^2 \right)^{1/2}$
- start at one end
- difference between pairs of atoms
- $|\vec{r}_i - \vec{r}_i'|^2 = (x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2$



- huge problem..
 - coordinates are normally...
- what to do ?



Translation and Rotation

translation

- c.o.m. = centre of mass $\vec{r}_{com} = \left(\sum_{i=1}^N m_i \right)^{-1} \sum_{i=1}^N \vec{r}_i m_i$
- subtract difference vector $\vec{r}_{diff} = \vec{r}_{com} - \vec{r}'_{com}$

rotation

- rotation matrix to minimise

$$rmsd = \left(N^{-1} \sum_{i=1}^N |\vec{r}_i - \vec{r}'_i|^2 \right)^{1/2}$$

summary

- translate
- rotate
- apply formula
- still not finished

Which Atoms ?

What tells me the shape of a protein ?

- backbone trace

What happens if you include all atoms ?

- bigger *rmsd*
- normal choice
 - C^α
- sometimes
 - N, C^α , C
- all atoms ?
 - when a model is very close



Still not finished with simple *rmsd*

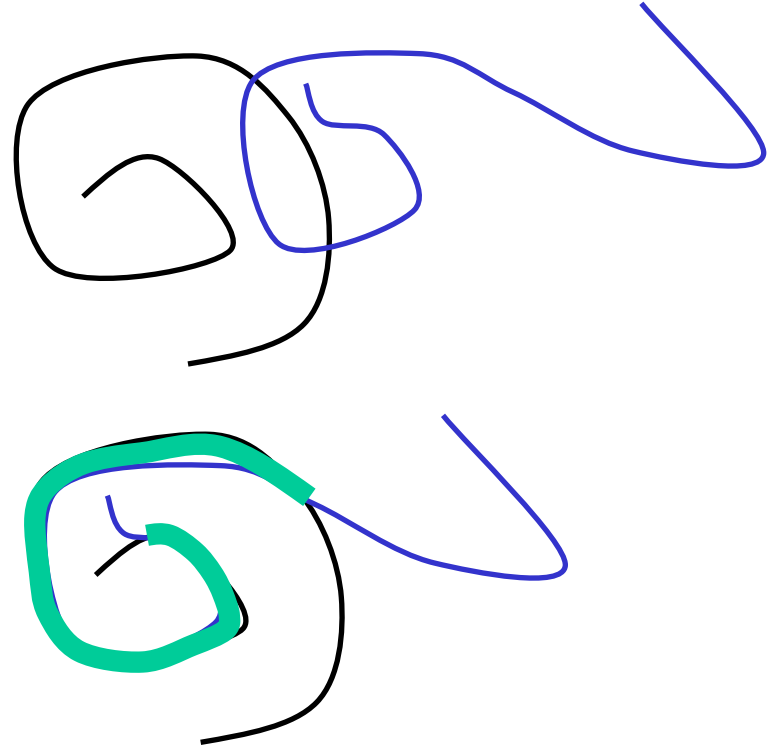
Parts Of Proteins

Two models of a molecule

- mostly very similar
- is *rmsd* a good measure ?

Identify similar parts

(method used in chimera)



define

```
superimpose ({r},{r'},{d}) {  
    translate ({r},{r'},{d)  
    rotate  ({r},{r'},{d)  
}
```

where **{*d*}** is some subset of sites

Selection of Interesting Atoms

Define a threshold like **thresh** = 2 Å

```
{d} = { | $r_i - r'_i$ | } i=1..N
```

```
sort {d}
```

```
diff = rmsd ( { $r_i$  }, { $r'_i$  } )
```

```
while (diff > thresh) {
```

```
    remove largest d
```

```
    superimpose ( { $r$  }, { $r'$  }, {d} )
```

```
    recalculate distances
```

```
    diff = rmsd ( { $r$  }, { $r'$  }, {d} )
```

```
}
```

```
if (diff < thresh)
```

```
    return {d}, diff
```

```
else
```

```
    return broken
```

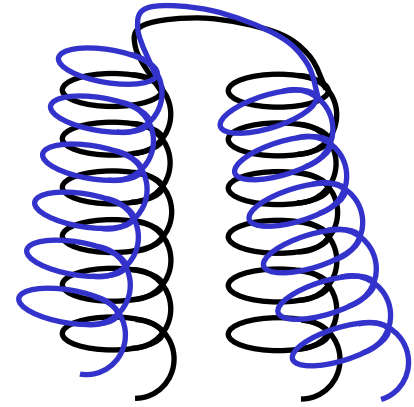
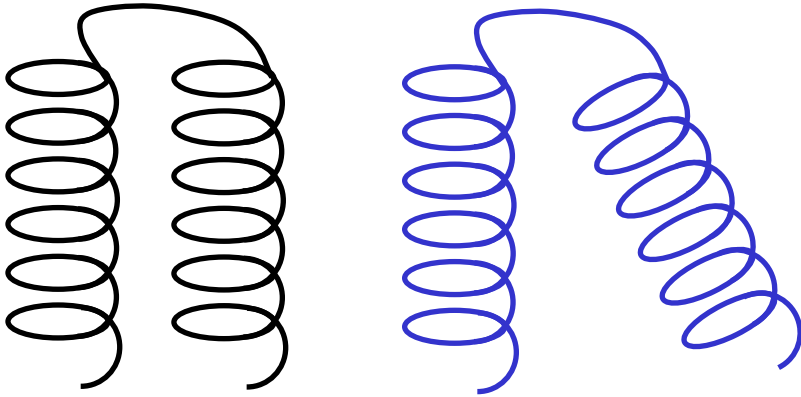
Result ? a subset of interesting atoms

Subsets of Atoms

- Originally, quantify structural differences as Å *rmsd*
- Alternative quantity implied
 - number of residues used for *rmsd* below threshold
- implicit rule
 - as number of atoms ↓ calculated *rmsd* ↓

Why Not Use *rmsd*

- helices identical, fold identical
 - *rmsd* ?



- big *rmsd*, but structure has hardly changed
- do not see that helices are identical
- more problems

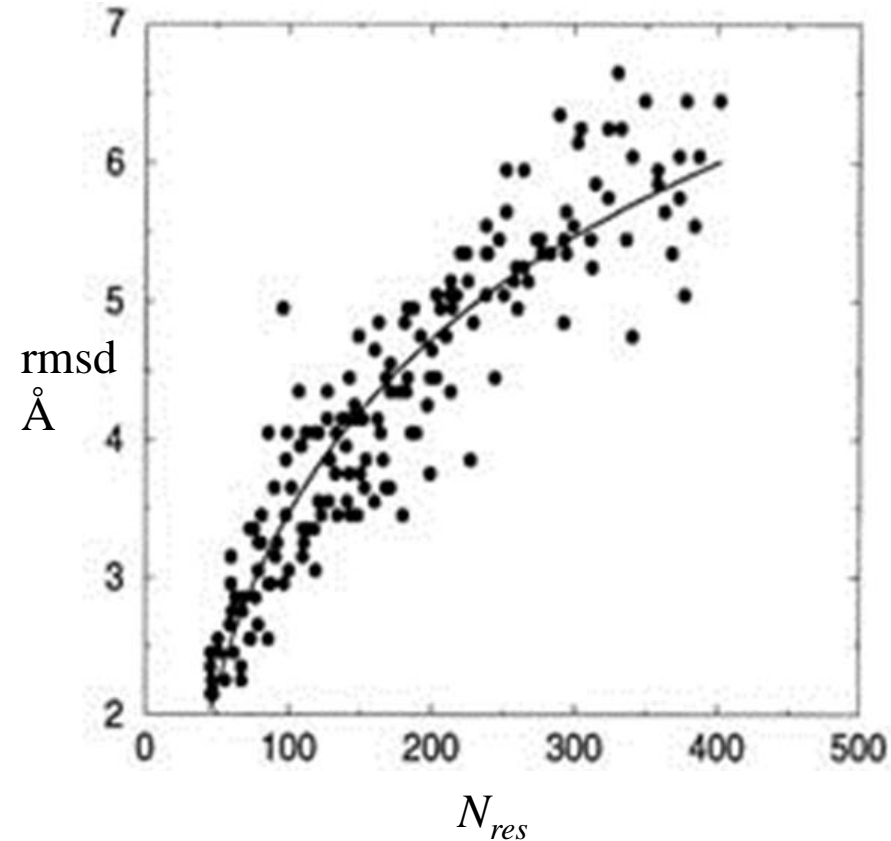
- superposition requires rotation, affects all atoms

Size dependence

- Two proteins with 5 Å *rmsd* – similar or not ?
- Consider proteins of different sizes
 - maximum difference with $N_{res} = 50$ or $N_{res} = 100$?
 - consider random structures with $N_{res} = 50$ or $N_{res} = 100$
- for small proteins 5 Å *rmsd* may be bad
- for large proteins 5 Å *rmsd* may be almost identical
- extends to comparisons of small molecules
 - ligands / medikamente...
- What would one expect for random structures ?...

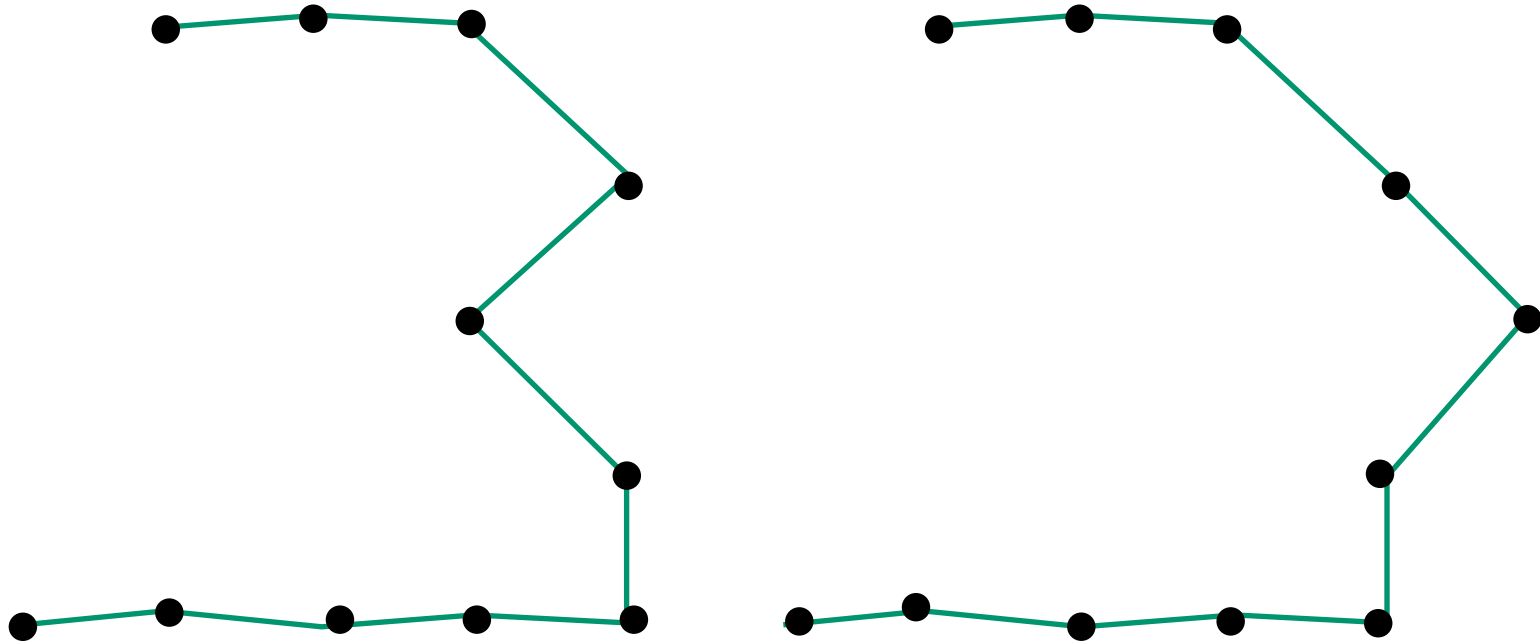
Size dependence

- a survey of random protein comparisons
 - several similar surveys in literature
- can find result from compact polymer theory (Florey)



rmsd size dependence

- good rule
 - $rmsd_{interesting} = a + b(N_{res})^{1/3}$ for some constants a, b
- problems with *rmsd* measure – alternatives
 - angles ? OK – angles compensate for another



- distance matrices ...

Distance Matrices With Numbers

Another characteristic of structures

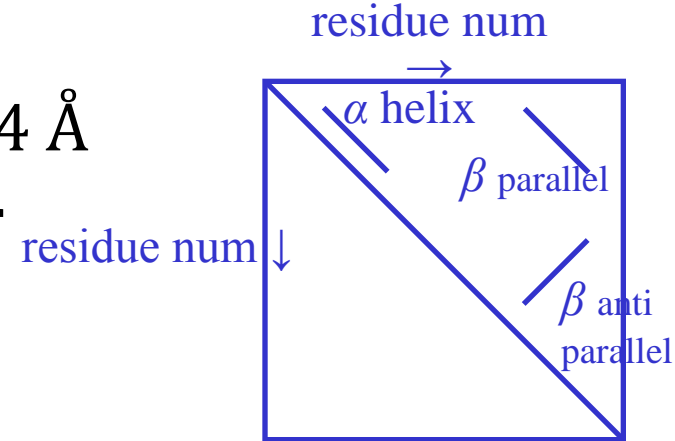
- C^α distance matrices
- simply measure the distance between C^α atoms

	1	2	3	4	5	6	7	...		N
1	0	3.8	6	7	...					
2		0	3.8	5	...					
3			0	3.8	4.5	...				
4				0	3.8					
5					0	3.8				
6						0	3.8			
7							0	3.8		
...								0	3.8	
									0	3.8
N										0

Distance Matrix for Recognising Structure

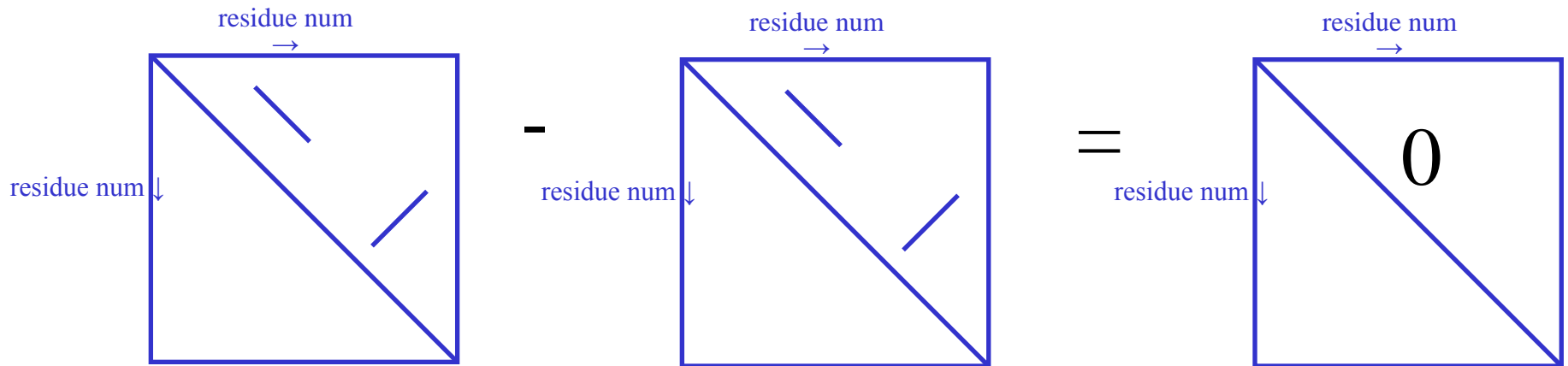
One way to summarise a structure

- plot C^α distance matrix, points below 4 Å
- can make α -helices and β -sheets clear



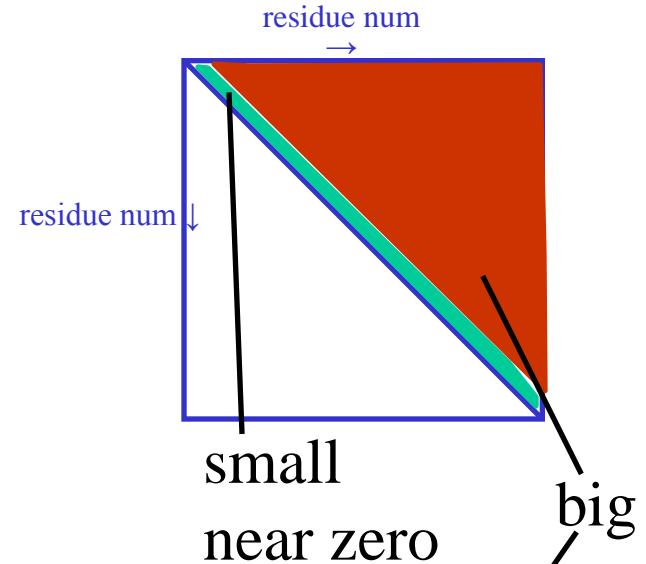
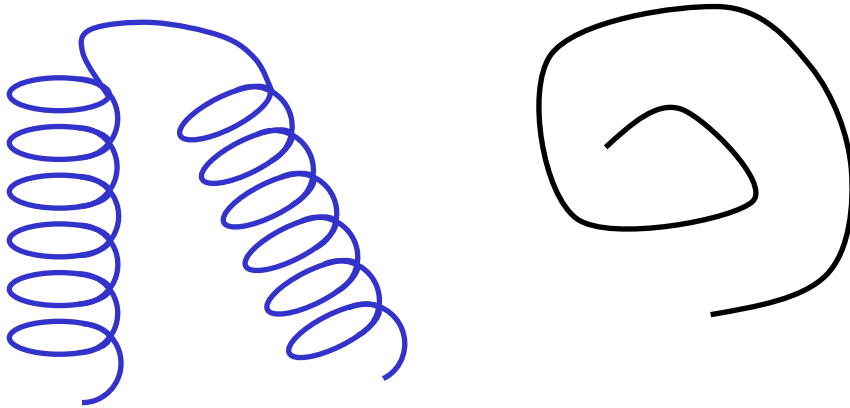
Distance matrix for comparing structures

- take two similar proteins
 - look at the difference of distance matrices

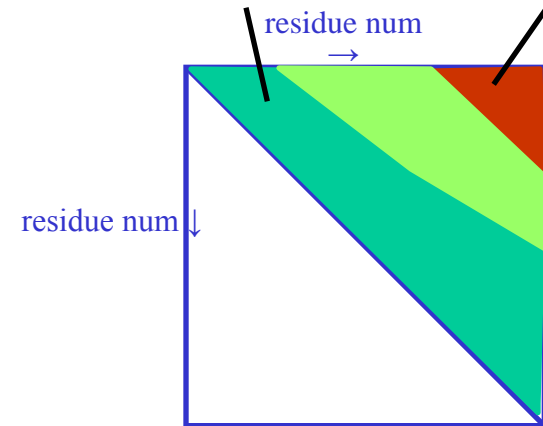
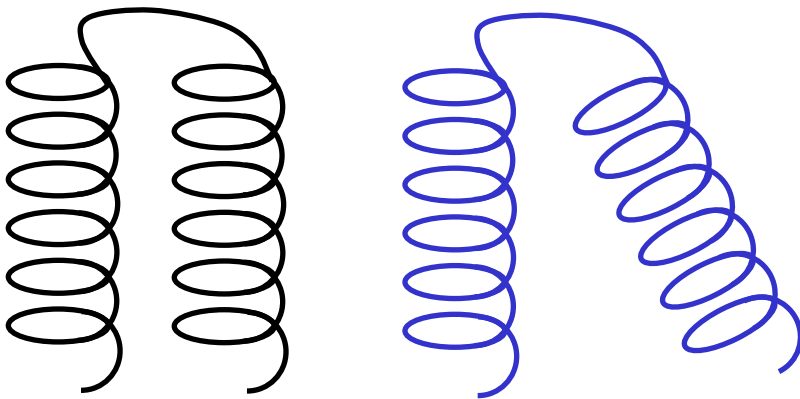


Comparing Distance Matrices

- consider two very different structures



- two related structures



- pictures are better than any single measure, but...

From Distance Matrices to Single Number

For lots of comparisons, single number is more convenient

- root mean square (*rms*) difference of distance matrices

- distance between C $^{\alpha}$ atoms i and j $d_{ij} = |\vec{r}_i - \vec{r}_j|$

- *rms* of distance matrices measure is

$$rms = \left(\frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j>i}^N (d'_{ij} - d_{ij})^2 \right)^{1/2}$$

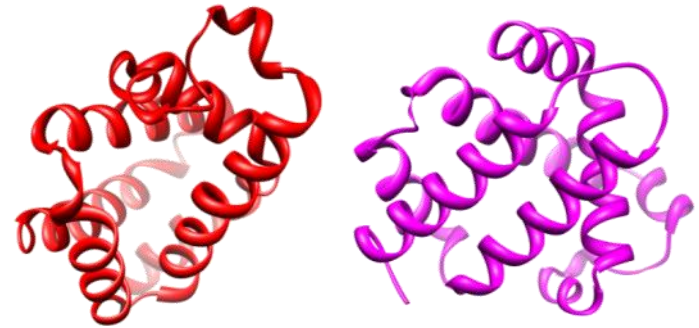
- just like all other *rms* quantities
 - normalised over top half of matrix

Summary – Comparing Models / Structures

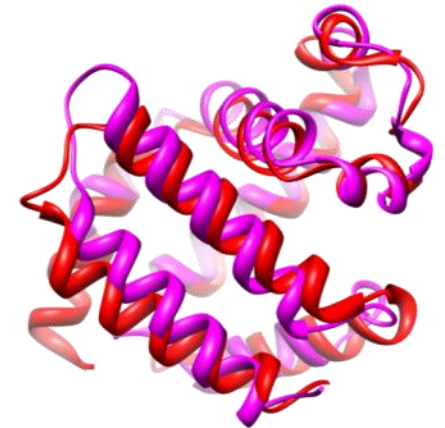
- *rmsd*
 - most popular
 - requires superposition (translate + rotate)
 - can be fooled by "hinge" movements
 - size dependent
- to look at the shape of a molecule use C^α or backbone atoms
- numbers in Å have a physical meaning
- to look for the common core of a structure, find a subset of backbone
- other measures may be better than *rmsd*
- weakness of all measures
 - a single number can never capture all information

Comparing Proteins – different sizes

- compare red and blue proteins
- if we know which residues match
 - easy (use any *rms* formula)
- which residues match ?
 - sequence alignment ?



protein 1	A	C	D	W	Y	T	R	P	K	L	H	G	F	D	S	A	C	V	N
protein 2	A	C	D	W	W	T	-	P	K	V	H	G	Y	D	S	A	C	V	N



- **green** residues – mismatches (no problem)
- **pink** residues – ignore
- is this useful for similar proteins ? very (rat vs human haemoglobin)
- for very different proteins ? no

Comparing Very Different Proteins

Sequence alignment vs identity

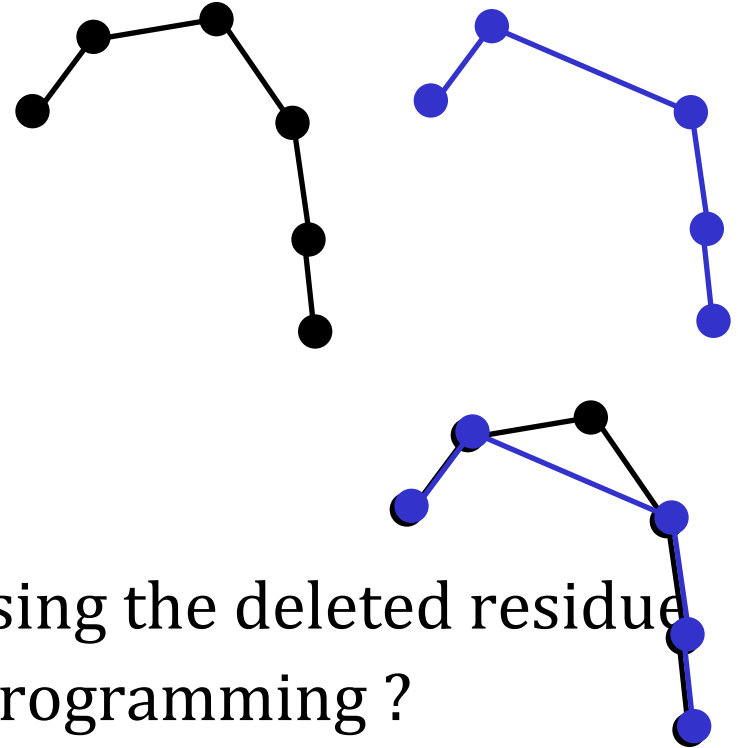
- as identity ↓, errors ↑

Consequence

- methods needed
 - operate on C^α
 - do not require sequence

How difficult ?

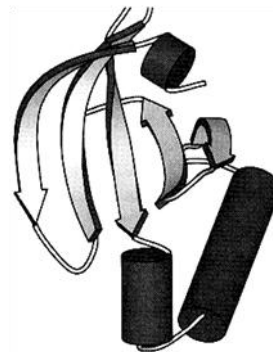
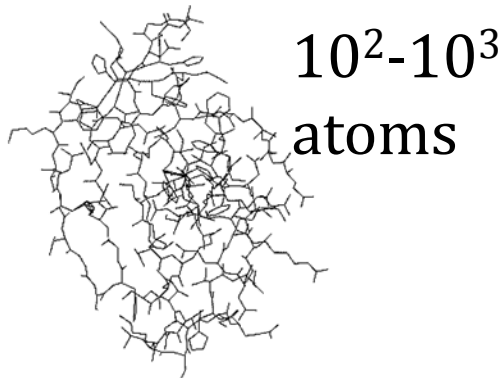
- superposition requires recognising the deleted residue
- can we use standard dynamic programming ?
 - no
- gap/insertion at any position, any length
 - combinatorial explosion



Strategies For Comparing Different Structures

1. use secondary structure

- Combinatorial explosion is the problem
 - reduce size of problem
 - use elements of secondary structure

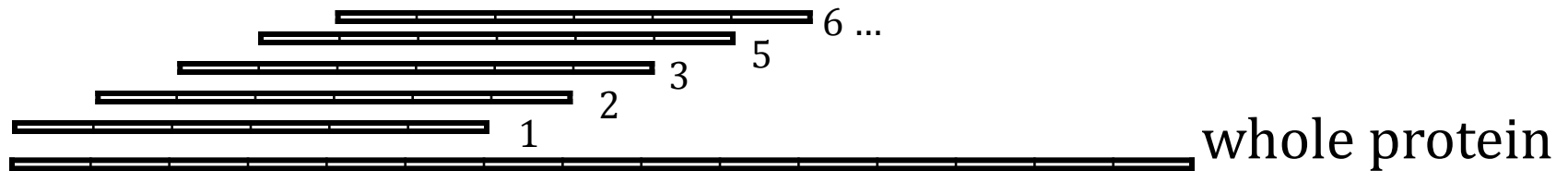


about 8 units

- define secondary structure
- search for superposition
- for each residue
 - find closest C^α in partner structure
 - use the set of matching residues to calculate *rmsd*

2. Peptide fragment strategy

- more general version of idea on previous page
- basis of most popular methods
- Ingredients
 - break protein into overlapping fragments (length 6 or 8)
 - protein is no longer a string of residues nor a whole structure

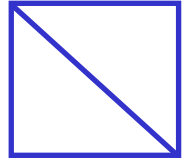
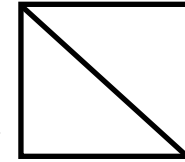


- each fragment is a little distance matrix



Fragment Based Comparison

- any two distance matrices can be compared
- two proteins length N and M can now be compared...



protein 2 fragments

	1	2	3	4	5	...		$N-7$
1	1.3	1.0	2.0	0.9	...			
2	2.7	2.3	0.5	...				
3	5.5	4.4	...					
4	0.1	0.5	0.3	3.3	4.2	...		
5	1.9	4.4	5.5	0.3	3.3	...		
6	4.4	1.6	1.7	5.0	2.3	...		
...	4.1	3.1	3.3	4.4	0.2	3.3	...	
$M-7$	5.2	1.1	0.1	5.5	4.4	0.1	3.3	0.1

protein 1 fragments →

- imagine *rmsd*
- this is now like a sequence comparison problem

Finding Equivalent Fragments

- find optimal path through matrix
- classic dynamic programming method like sequence comparison

	1	2	3	4	5	...		N-7
1	1.3	1.0	2.0	0.9	...			
2	2.7	2.3	0.5	...				
3	5.5	4.4	...					
4	0.1	0.5	0.3	3.3	4.2	...		
5	1.9	4.4	5.5	0.3	3.3	...		
6	4.4	1.6	1.7	5.0	2.3	...		
...	4.1	3.1	3.3	4.4	0.2	3.3	...	
N-7	5.2	1.1	0.1	5.5	4.4	0.1	3.3	0.1

- like sequence comparison
 - find optimal path through matrix
 - classic dynamic programming method (N & W, S & W)
 - uses gap penalties

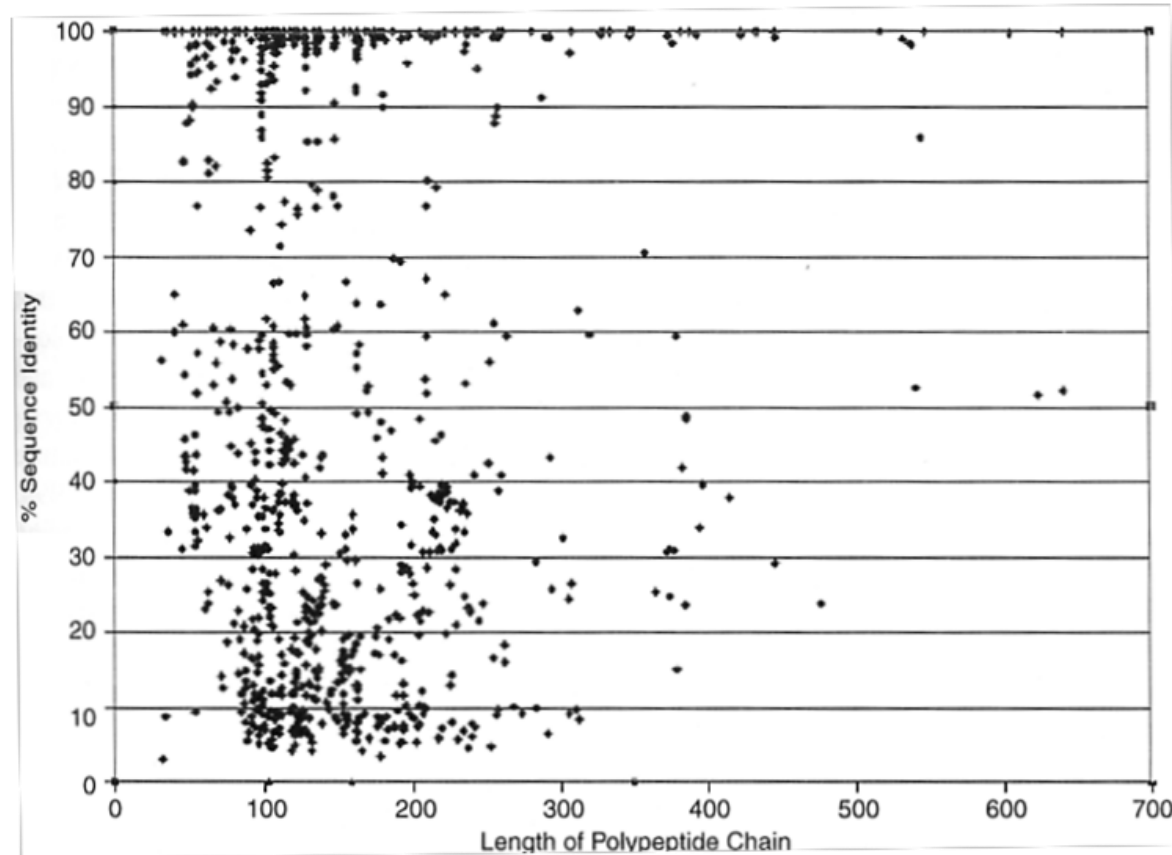
Comparing Different Size Protein Structures

- Break protein into overlapping fragments
- fragments can be compared to each other via distance matrices
- align like sequences
- from aligned fragments, get list of aligned residues
- using aligned residues, calculate *rmsd*, *rms* of overall distance matrices

How Important Are These Similarities ?

- survey 1 000 proteins
- find structurally similar pairs
- plot sequence identity

may not be found by
sequence methods



Summary of All Protein Comparisons

Classification of proteins

- could be done by sequence, better by structure

Structure comparison

- for one protein
 - selection of atoms
- for different proteins
 - requires list of matching atoms
- for similar proteins
 - can use pairs from sequence alignment
- for often dissimilar proteins
 - pure structure based method

Summary of everything

- classification is appealing
- very different answers using sequence or structure
- even if we believe in evolution
 - complete hierarchical scheme may be artificial