Analysis, Comparison of Proteins

Andrew Torda, wintersemester 2009 / 2010, GST

From previous lectures

- we know about protein structures / coordinates
- we know how coordinates are collected

What kind of analysis would we like to do ?

- recognising common features
- classifying
  - (useful for prediction)

Philosophy

- ways to measure similarity between structures
- ways to find similar pieces / "motif"s
- most common motif ? Secondary structures

Next few weeks

- Secondary structure definitions
- Classifying protein structures
- Domains
- Supersecondary structure
- Protein similarity – sequence versus structure
- Sequence space
- Classifications – hierarchical
- Classifications – other
- Comparison of proteins

- touching on evolution, alignments, …

Secondary Structure Recognition

from coordinates
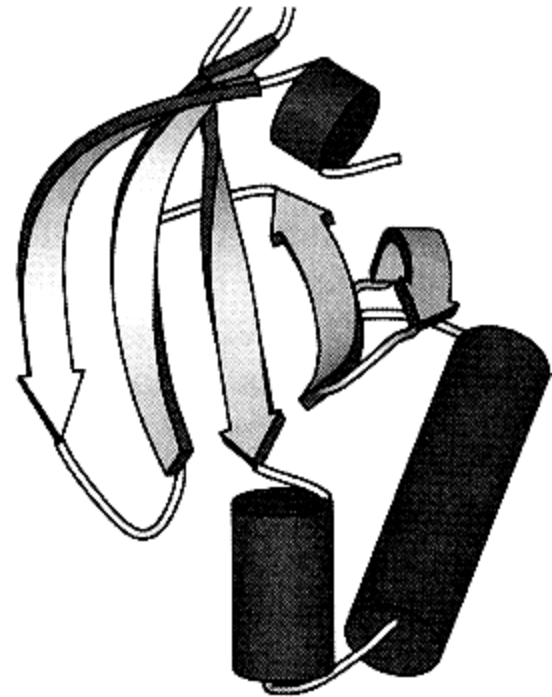
assumes structures
recognised

- how to define / recognise secondary structure ?

Defining Secondary Structure

- What do I want ?

	1	2	3	4	5	6	7	8	9	10	…
residue	A	C	A	D	L	V	A	W	W	A	…
sec struct	-	H	H	H	H	H	-	E	E	E	…

- at each residue, label as to secondary structure type
  - no ambiguity
  - labels at residues – not between !
- I do not want probabilistic answers (more soon)
- remember not all residues are in recognisable α-helix or β-sheet

Secondary Structure From Coordinates
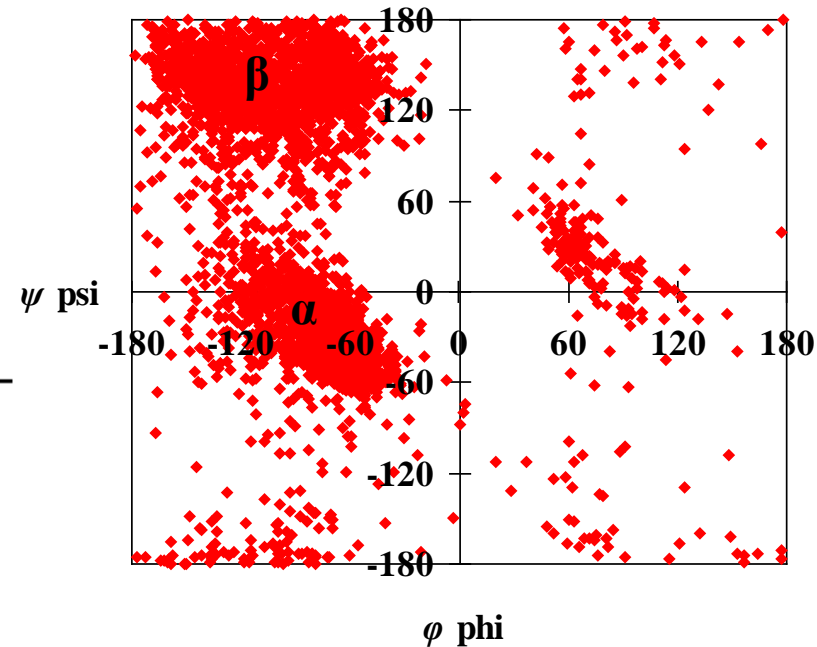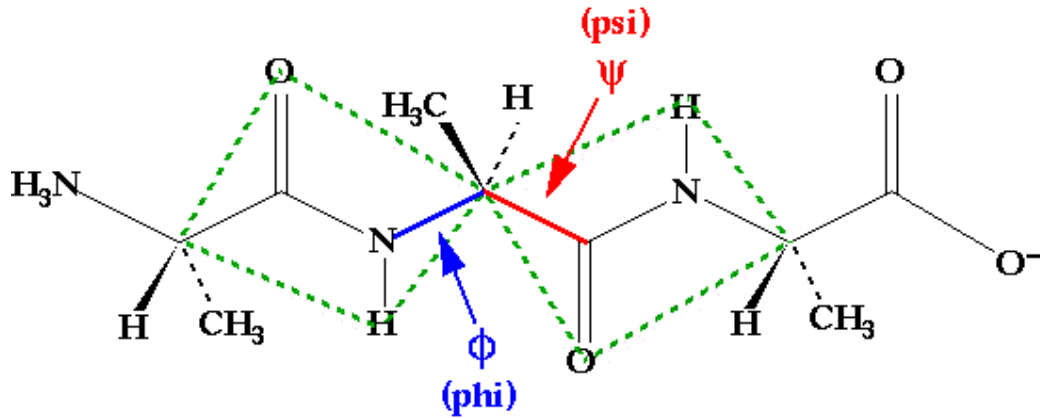
Start with *α*-helices

- what do we know
  - look like helices
  - 3.6 residues per turn
  - H-bond pattern
    - N residue *i* to *i*+4
  - residue backbone angles
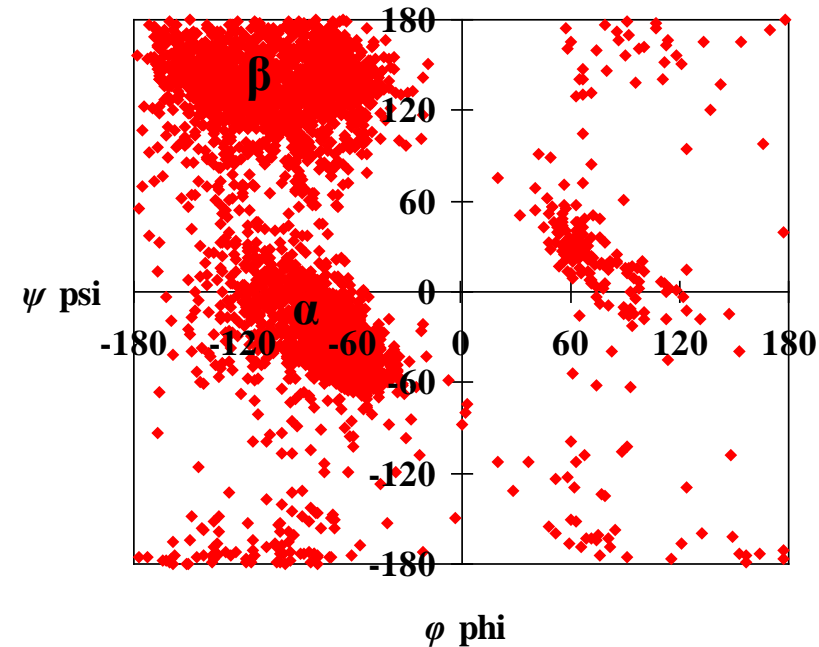
Using Backbone Angles
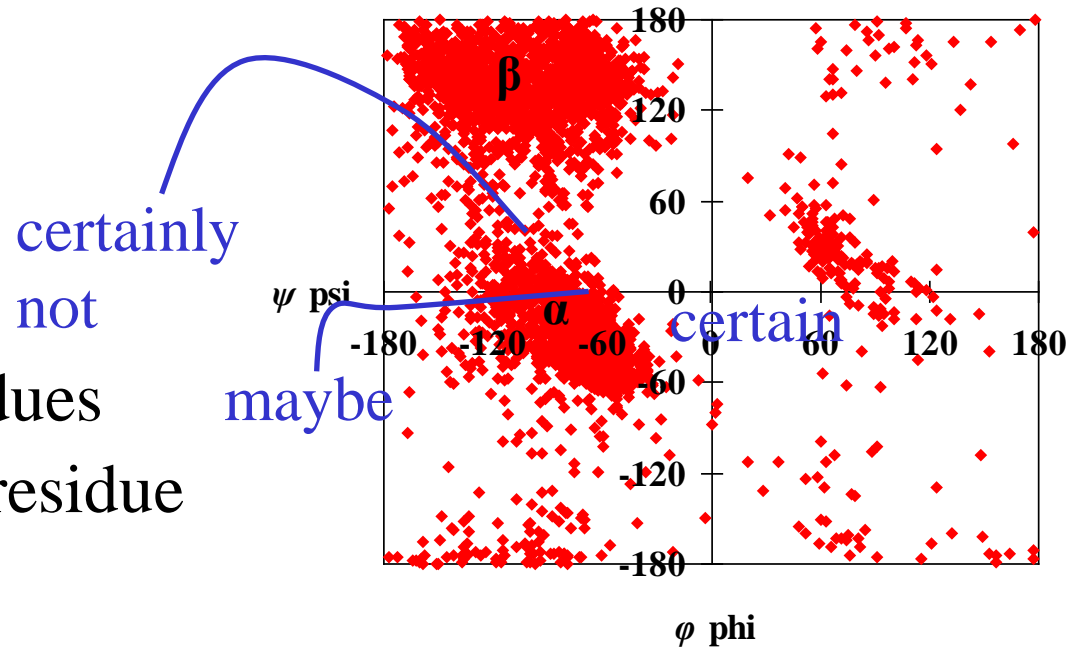
Given coordinates, easy to calculate

Problems
- what are my thresholds ?
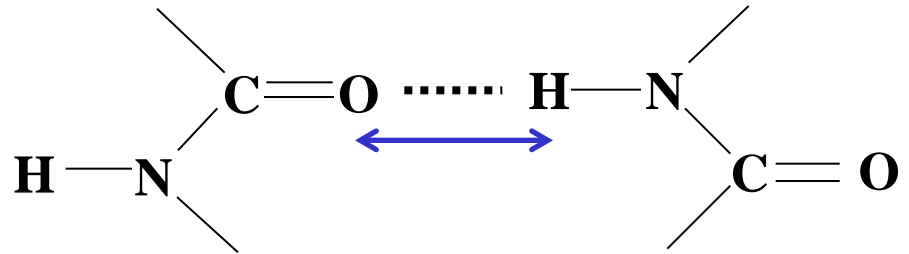- what if I see one residue with angle ?

Problems With Using Angles

- thresholds

- minimum number of residues
  - what if I see only one residue with perfect angles ?
  - not forming H bonds
  - need 3 or 4 residues

certainly not

maybe

certain

β

α

ψ psi

φ phi
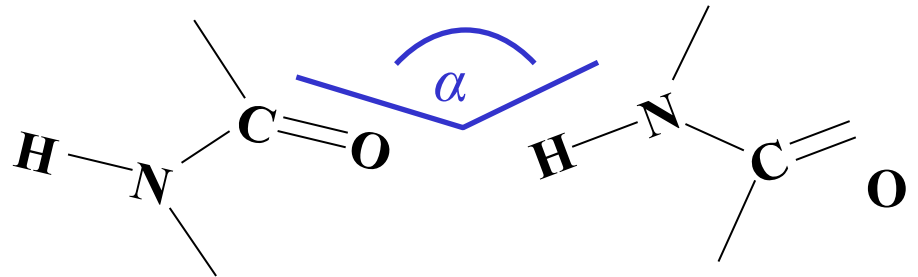
180
120
60
0
-60
-120
-180

-180 -120 -60   0   60   120   180

Maybe We Should Use H-Bonds

We have the coordinates

- should be easy to recognise all H bonds
- criteria ?
- distance $r(ON) < \approx 3.6$ Å

- angle ? ($\approx 120°$)

A practical definition
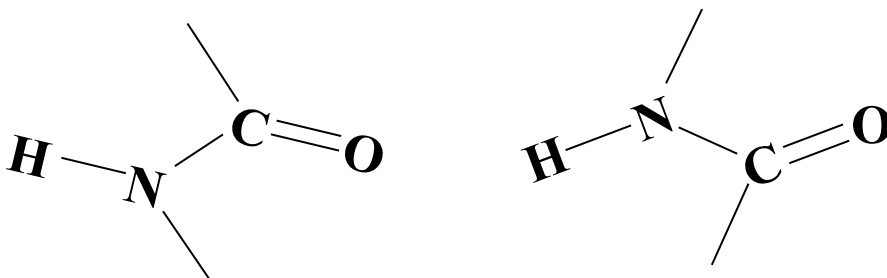
$$E = 332\, q_1 q_2 \left(\frac{1}{r_{\mathrm{ON}}} + \frac{1}{r_{\mathrm{CH}}} - \frac{1}{r_{\mathrm{OH}}} - \frac{1}{r_{\mathrm{CN}}} \right)$$

- require $E < -3$ (arbitrary !)
- note as r grows, E goes to 0

Kabsch, W. & Sander, C, Biopolymers, 22, 2577-2637, "Dictionary of Protein Secondary Structure…" (1983)

Problems with short helices

What if I see only 3 or 4 residues ?
- real helix has 2 H-bonds per residue
- what if I see one ?
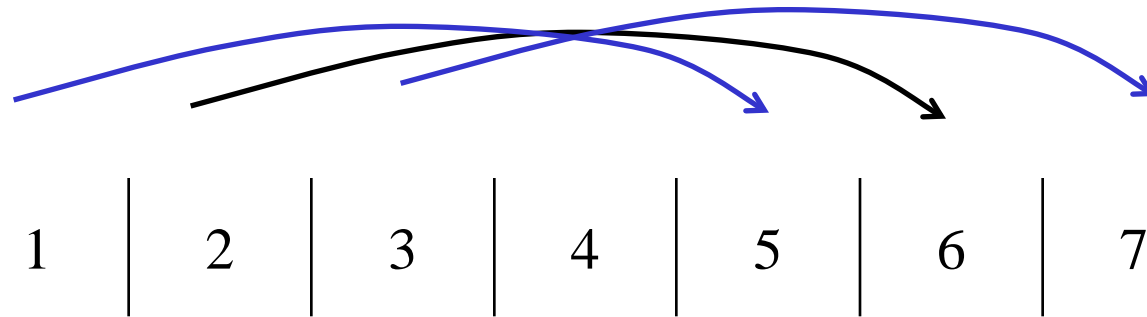
Compromise
- call this a turn (only has one H bond)
- α-helix definition
  - at least two consecutive (4 residue) turns

Useful definitions (α-helix)

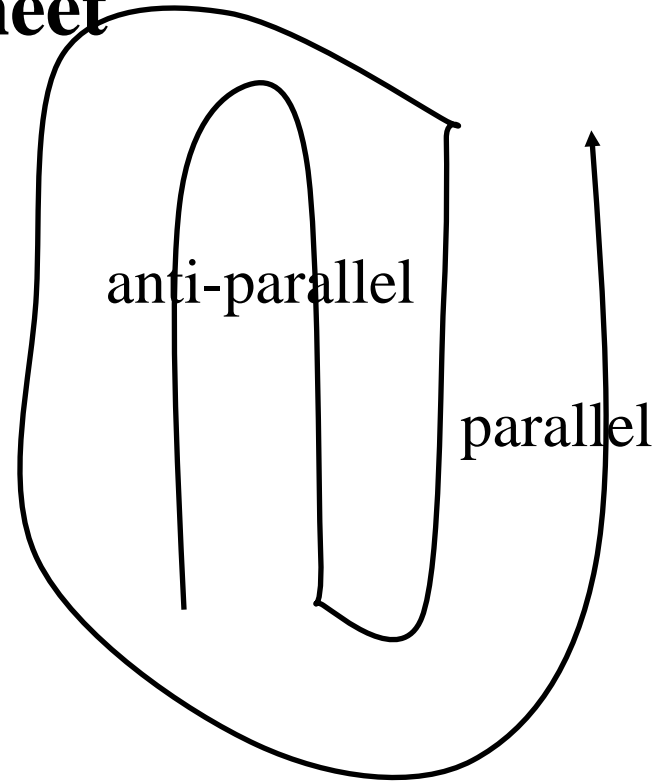- recognise an H-bond $i,i+4$ (either
    - r_{ON} + angle α or
    - general distance formula
    = turn
- two successive turns
    = minimal α-helix
- more overlapping helices
    = longer helix

- all we have done is an α-helix

a β-strand / sheet
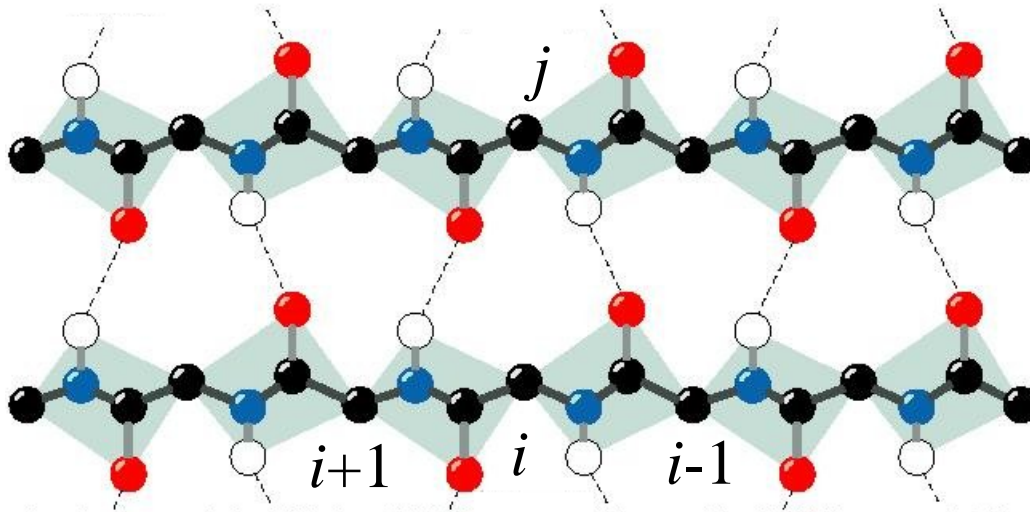
- much more difficult
  - parallel versus anti-parallel
  - H-bond neighbours not known
    - 5-109
    - 6-110
    - 7-111 … parallel
    or
    - 5-109
    - 6-108
    - 7-107…anti-parallel

- formalise this

anti-parallel

parallel

Defining a β-sheet

- start with a bridge
- parallel bridge
  - H-bond (i-1, j) & H-bond (j, i+1) or
  - H-bond (j-1, i) & H-bond (i, j+1)
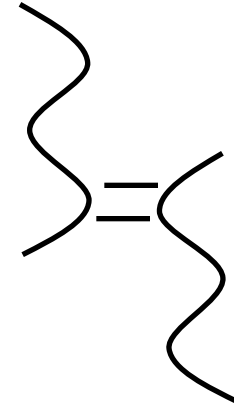
Defining a β-sheet

- start with a bridge
- parallel bridge
  - H-bond (i-1, j) & H-bond (j, i+1) or
  - H-bond (j-1, i) & H-bond (i, j+1)

- ladder = one or more consecutive bridges

- sheet = one or more consecutive ladders with shared residues

- similar definition for anti-parallel sheets
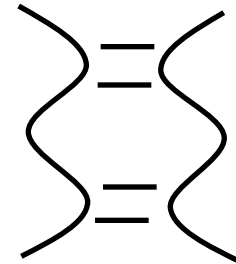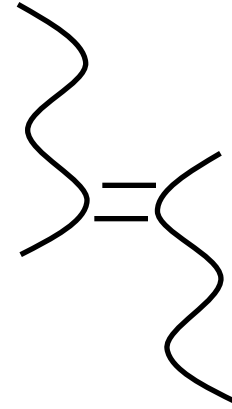
Are these problems real ?

Do thresholds matter ?

- do programs give the same answer ?
- if we use secondary structure for comparisons of proteins

- comparisons with experiment

Can we set perfect thresholds ?

- not all H-bonds are the same
- look at ψ-φ map, borders are not clear
  - mobility, finite energy
- coordinates have experimental error
- our programs generate worse coordinates (holes, distortions)
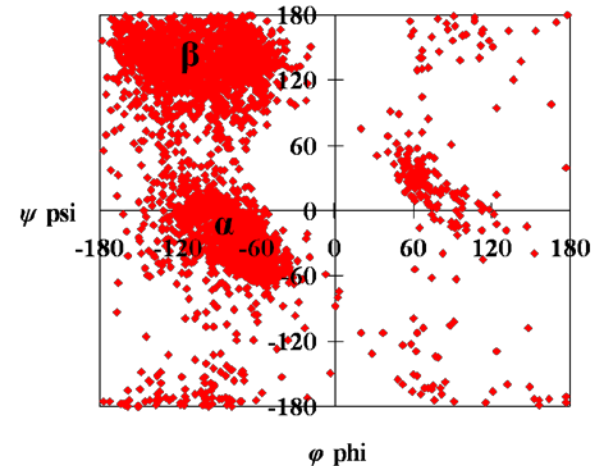
From secondary to higher levels Classification

- Why classify proteins ?
- Why recognise similarities
    - function prediction
    - structure prediction
        - vague idea of structure for mutagenesis, applications
- Why might this be useful ?
    - how many structures are there ?

How Many Protein Structures Are There ?

- Protein Databank $\approx 5.5 \quad 10^4$
- 90 % sequence similarity $\approx 1.7 \quad 10^4$
- different shapes 2 to 5 $\quad 10^3$
- implications for structure prediction ?
  - how many possible structures can we think of ?
    - exponential
  - how big is the real search space ?
    - really 10^3 to 10^4

Why So Few Structures

- discretization of space (makes it look smaller)
- physical reasons
  - compactness, stability
  - advantages of H-bonded conformations
- history / evolution
  - imagine all proteins evolve from some original molecule
  - evidence
    - theoretical – geometric constructions
    - chemical – construction of artificial protein(s)

Before Classifying

- earlier description of structure
- primary (sequence)
- secondary (α-helices, β-sheets, …)
  - supersecondary ?
- tertiary
  - arrangement of helices / sheets or
  - where atoms are in space
- quaternary…

- we need idea of domains, then supersecondary structures

Domains in Biochemistry

History / biochemistry / no structures

- invented story
- we have a big protein
  - catalyses A $\rightarrow$ B
  - C regulates it
- cleave protein (break with enzyme) to two parts
  - 1 still converts A $\rightarrow$B
  - 2 binds C
  - interpretation
    - catalytic domain
    - C binding domain
- more generally
  - different pieces of protein, responsible for different functions

Domain Concept Useful ?

- Many times a whole protein cannot be crystallised, solved by NMR

- attack protein with enzymes to break up

- look for activity in pieces, solve structures of pieces

domain 1   domain 2   domain 3 $\}$ each domain 50 to 250 residues
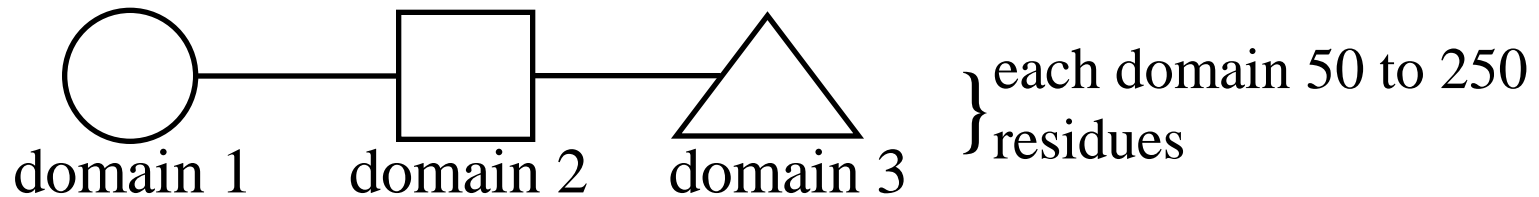
- literature / PDB full of "xxx domain of yyy"

- attractive ?

  - makes big proteins seem manageable

  - building block concept

    - attractive in evolutionary terms

Domains in Structures

- Many structures solved look like…
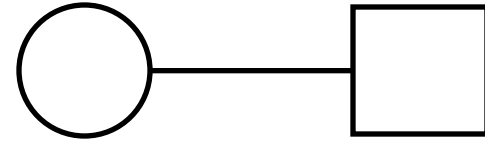
- histocompatibility module (1iak)
  - 3 domains + another protein

- are they always so clear ?
- porphobilinogen deaminase (1gtk)
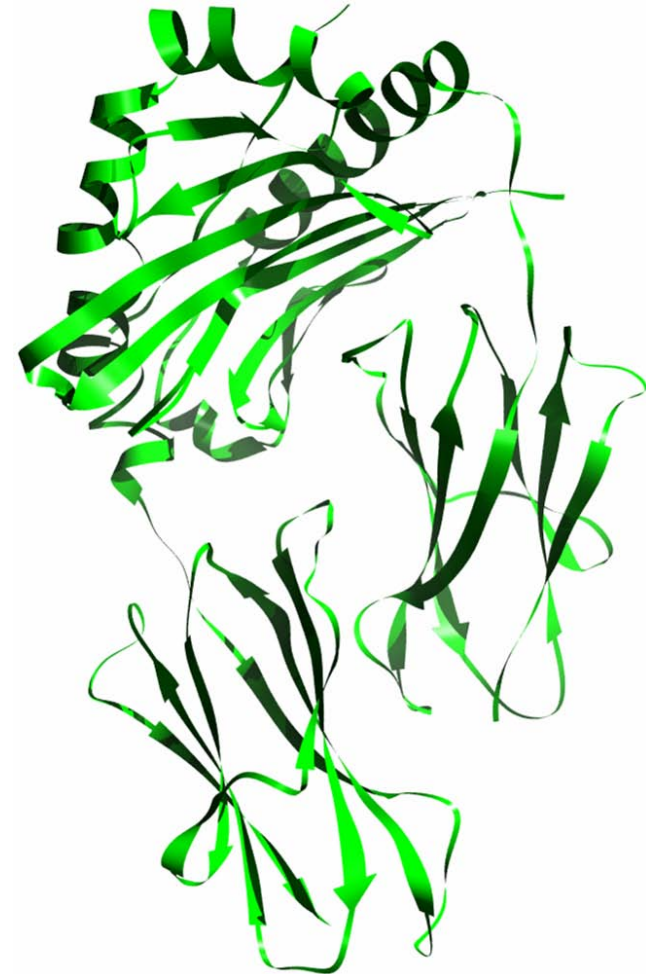
Domain definition version 3

Three reasonable definitions
- biochemistry
- structures
- look for conserved units in sequence comparisons

Domains for today

- compact structural units

Domains for classification

- structural classifications often domain based

Classifications In General

- 1. secondary structure
  - we see collections of residues and classify into recognisable types
- 2. different types of domain
  - soon

- 1b. supersecondary elements ?
  - are there some common small arrangements of α-helices, β-sheets ?

A Supersecondary Structure

- β-hairpin (β-turn-β) fits idea of common motif
  - described as built on secondary structure + specific H-bonds

Two-residue beta-hairpin turns.

Type I'   Type II'

White dots indicate hydrogen bonds.

The main difference between these two turns is the orientation of the peptide group between residues 1 and 2.

Two residue beta-hairpin turns.

Residue 1
Type I'
Residue 2

Residue 2
Type II'

Residue 1
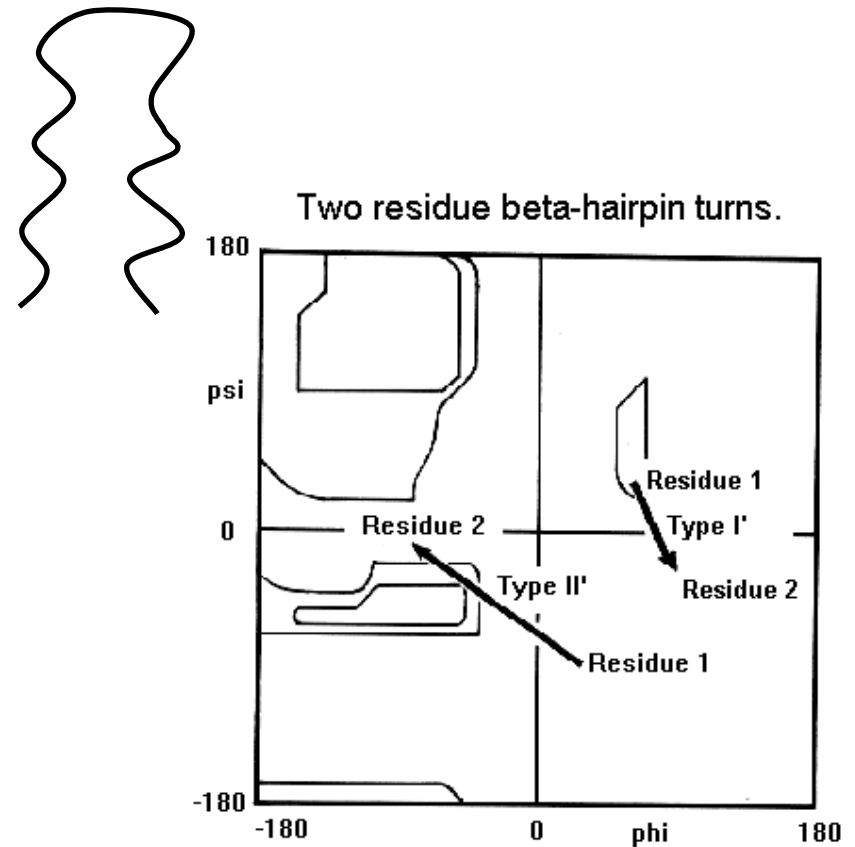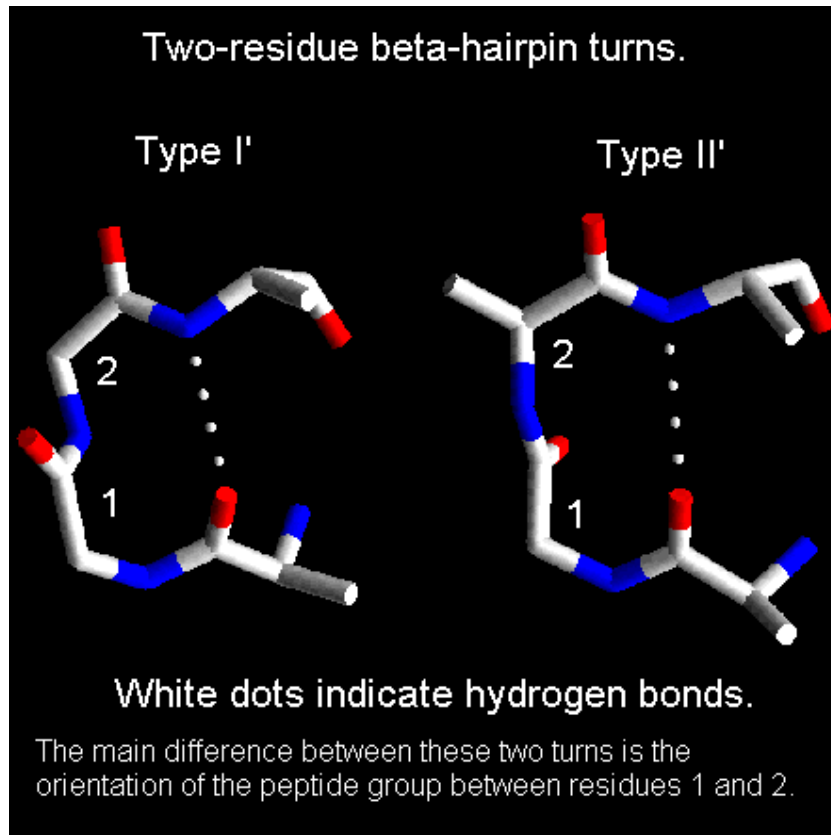
More Supersecondary Structures

- helix-turn-helix (α-X-α)
  - DNA binding proteins
- helix-longer_loop-helix
  - Ca^{++} binding
- …

Who cares ?

- repeated patterns / motifs suggests there are smaller number of structural units to recognise
- modularity appeals
- functional association
- conforms to some ideas on protein folding (more next semester)

Why I Do Not Like Supersecondary Structure

- ideal picture…
  - primary structure arrangement →
  - secondary structure / arrangement →
  - supersecondary structure →
  - tertiary structure or domains
- implies supersecondary structure is useful hierarchical element
  - not really used !

Sequence vs Structural Similarity

Background

- in the real world we usually have sequence information first
- want to make guesses about protein structure

I have two aligned protein sequences

- are they structurally similar ?

Old rule

- > 25 % sequence similar – similar structures
- < 20 % cannot tell
- 20 % < x < 25 % - "twilight zone"

Is this universally valid ?

Sequence Similarity → Structure

Take a set of pairs of proteins

- find those which are not structurally similar
- look at sequence similarity

- old rule is not valid
- 50 residues
  - > 30 % seq
- 150 residues
  - > 20 %

- rule:
  - sequence similarity (length dependent) very good indicator of structural similarity

Rost, B., Protein Eng. 1999, 12:85-94, "Twilight zone of protein sequence alignments"
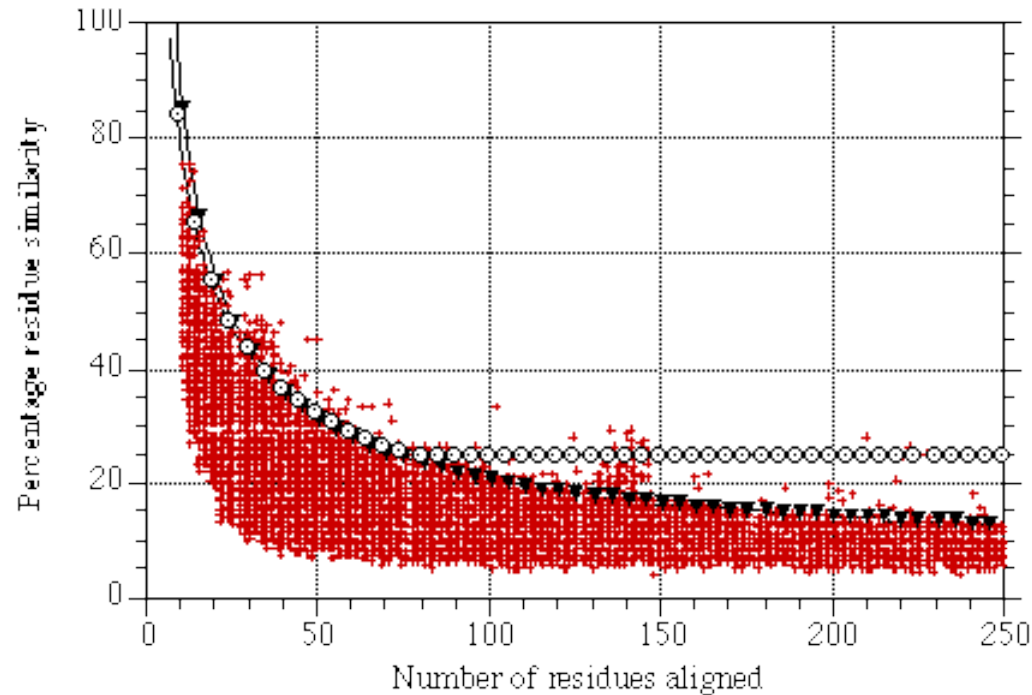
Using Sequence Similarity

- consequence
  - I could try to categorise proteins based on sequence
- tools
  - any alignment program (blast, fasta, clustal, …)
- method
  - survey all proteins in the protein databank
  - collect all pairs > x % (or use more sophisticated threshold)

- result (jan 2009)

similarity	num clusters
90 %	20 002
70%	17 490
50%	14 906

- much more than 2 to 5 $\times$ 10^3 ?
- maybe some of my classes are not really different

Sequences Not Similar

- Sequences similar ? .. similar structure
- Sequences different ?
  - ??

- Example
  - 100's examples

1ecd, 1mbd
no significant
sequence identity
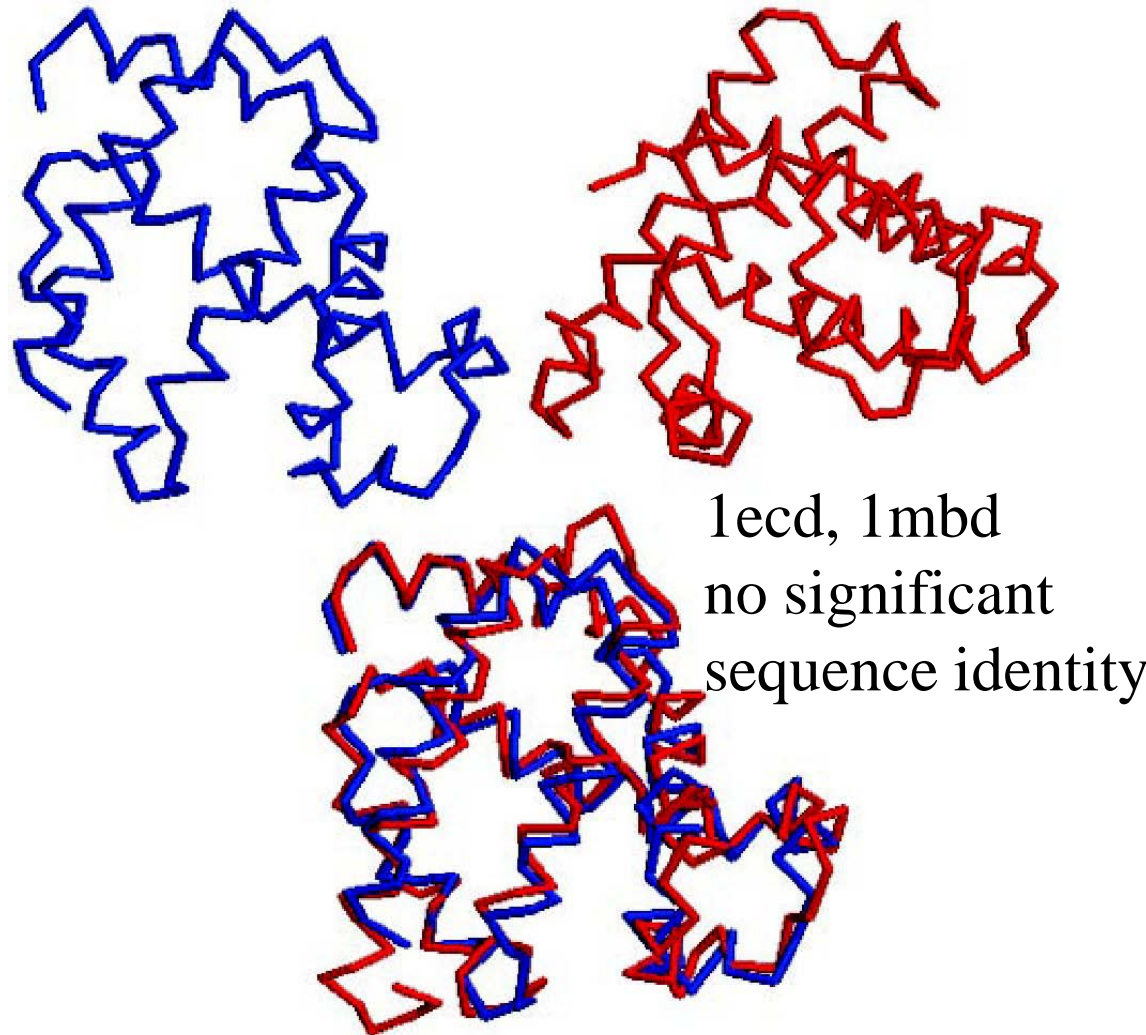
An Example Family

- 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 T
```

DIVERSION 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
  - 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 $= 20 \times 20 \times … = 20^{Nres}$
  - like a space of N_{res} dimensions

Representing a Sequence
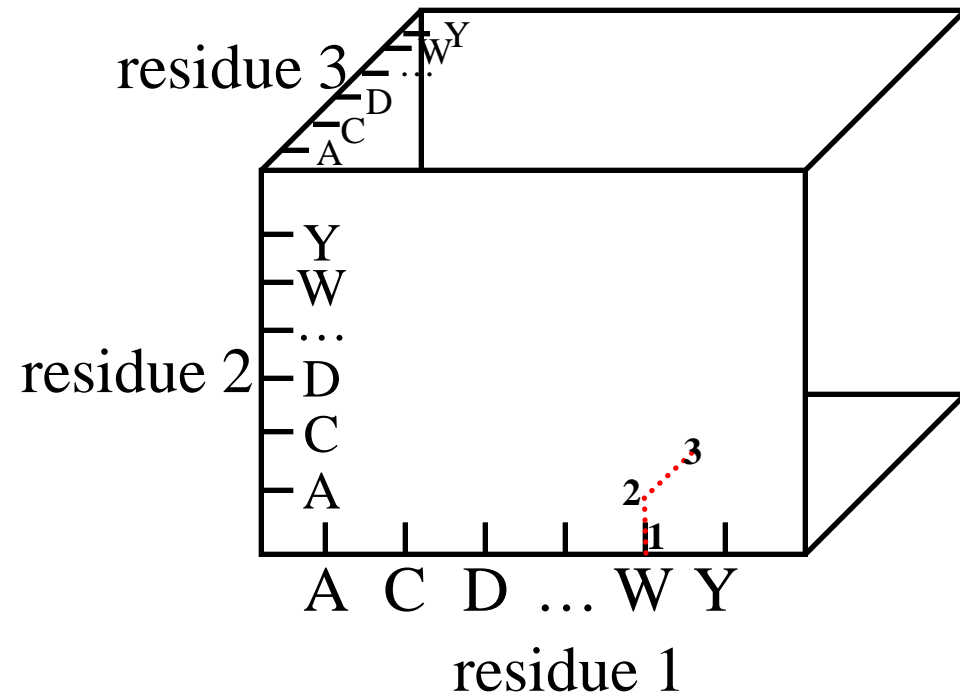
- protein sequence and structural coordinates

	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**

residue 3

residue 2

residue 1

- 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

residue 3

residue 2

A C D …WY
residue 1

- Will return next semester
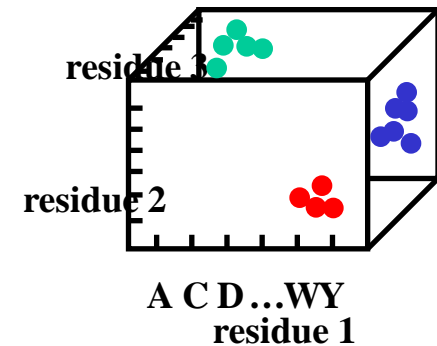
Structure vs Sequence

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

sequences
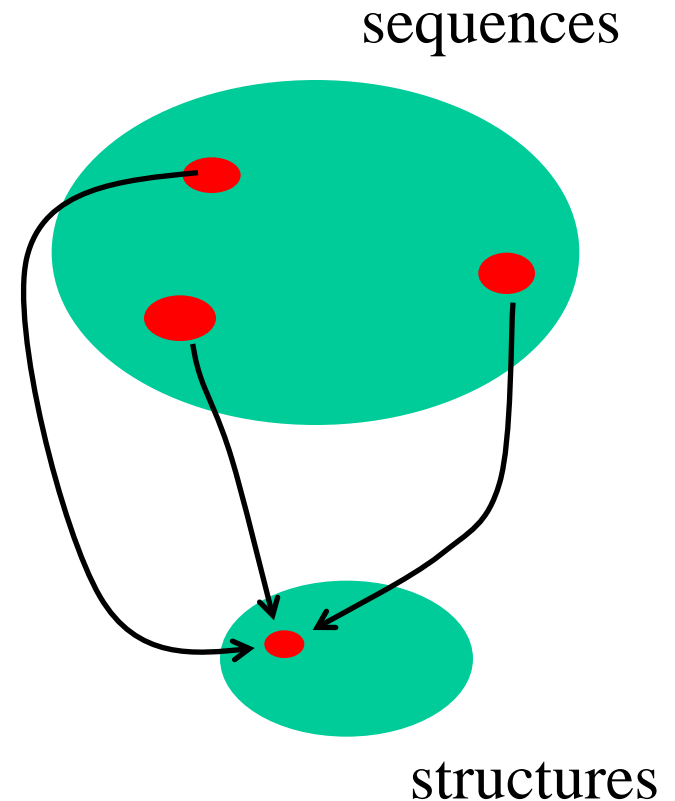
structures

Structures < Sequences… Why ?

Evolution 1

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

Evolution 2

- maybe some cases of convergent evolution
- impossible to prove

Consequences of sequence based categorisation

- we will have different classes, but really same protein shape

Surprising ?

- consider near universal proteins
  - 100's millions years evolution, function largely preserved
- chemistry
  - sequence does determine structure, many sequences could fit structure

Back to Classification
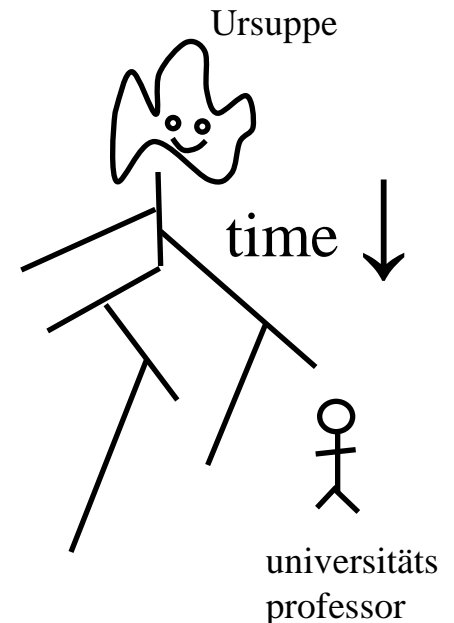
Sequence classification

- good, reliable, similar class = similar structure
- not enough to find all similarities
- need for structure based methods
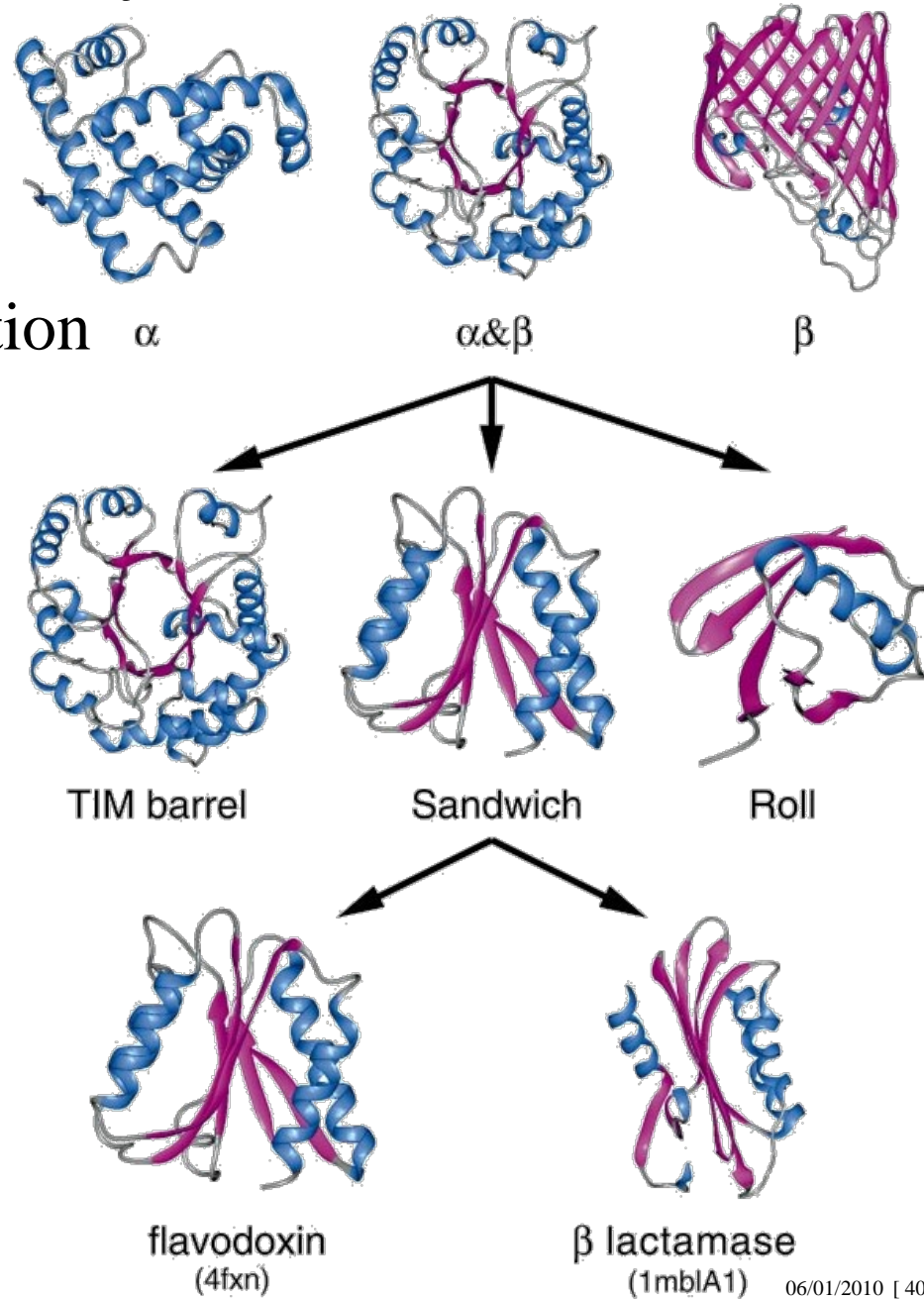
Philosophies

1. evolution

2. just classify proteins
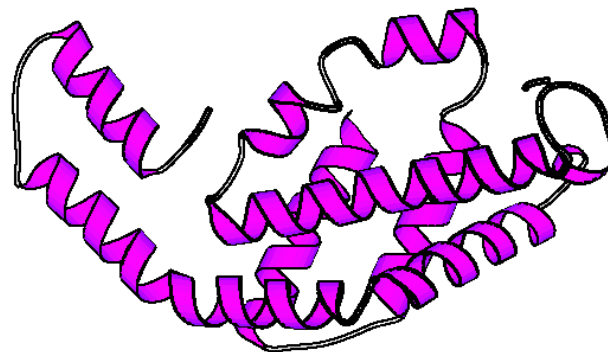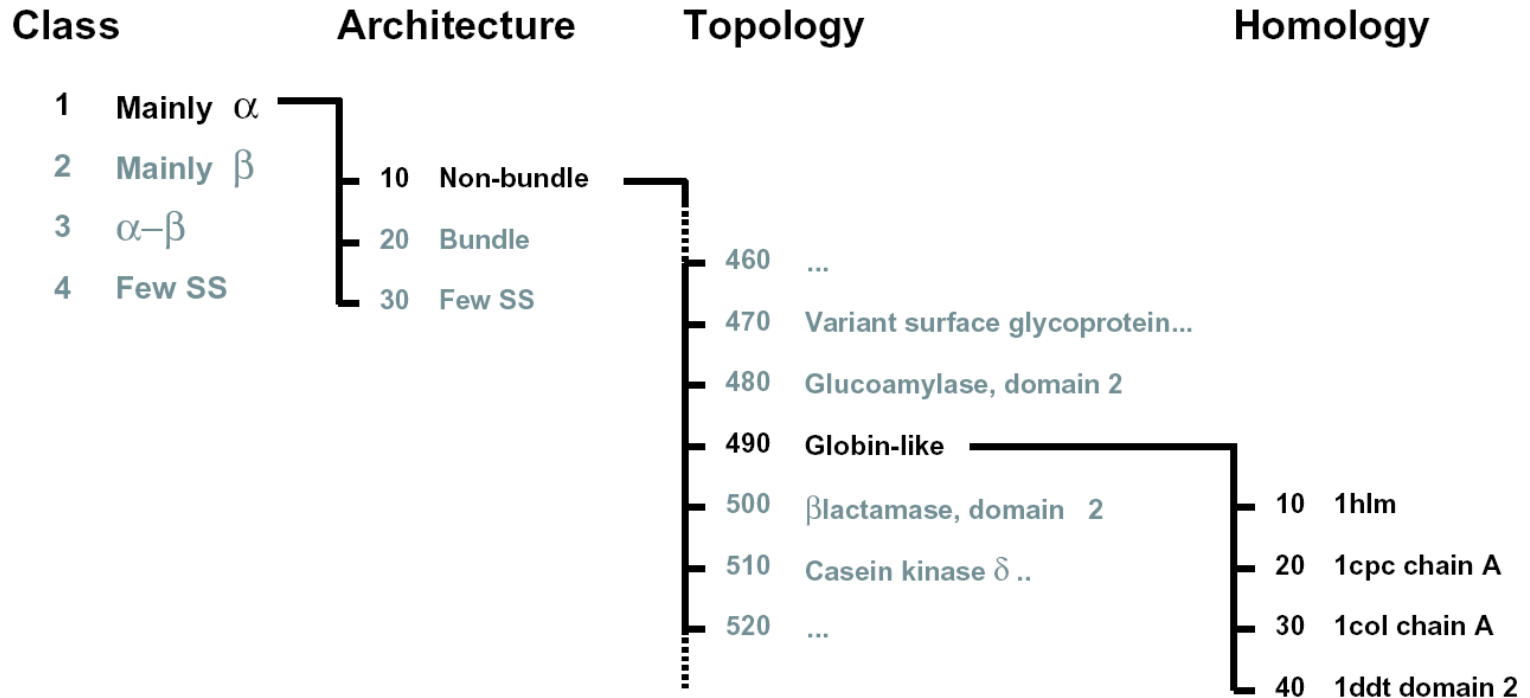
Evolution

- diagram →
- we expect a hierarchy

Ursuppe

time ↓

universitäts professor

Imposing a Hierarchy on Proteins

- parts may correspond to evolution
- top level ?

- How useful and applicable ?
  - examples

α   α&β   β

TIM barrel   Sandwich   Roll

flavodoxin
(4fxn)

β lactamase
(1mblA1)

CA Orengo AD Michie, S Jones, DT Jones, MB Swindells, JM Thornton, Structure, 1997, 5,1093–1108

Example from "CATH"

Class	Architecture	Topology	Homology
1 Mainly α	10 Non-bundle	460 ...	10 1hlm
2 Mainly β	20 Bundle	470 Variant surface glycoprotein...	20 1cpc chain A
3 α–β	30 Few SS	480 Glucoamylase, domain 2	30 1col chain A
4 Few SS		490 Globin-like	40 1ddt domain 2
		500 βlactamase, domain 2	
		510 Casein kinase δ ..	
		520 ...	

1.10.490.20

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

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

α Bundle (2ccy)   α Non-Bundle (1cca)   α Few SS (2ifo)
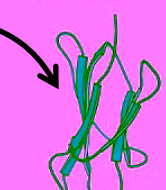
β Ribbon (1tpm)   β Single Sheet (1hro)   β Roll (1pht)

β Barrel (2por)   β Clam (3hcl)   β Sandwich (2BfaB)

β Distorted Sandwich (1cdq)   β Tretoil (1afcA)   β Orthogonal Prism (1msaA)

β Aligned Prism (1vmoA)   β 4-Propellor (1hxn)   β 6-Propellor (1nscA)

β 7-Propellor (2bbkH)   β 8-Propellor (3aahA)   β 2-Solenoid (1tsp)

CA Orengo AD Michie, S Jones,DT Jones, MB Swindells,JM Thornton, Structure, 1997, 5,1093–1108

Why are 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

- is this a reflection of physics ?

- no – PDB is very biased
  - mainly soluble, globular proteins which crystallised
  - very few membrane-bound proteins

# Supersecondary Structure

- Was supersecondary structure helpful here ?
- members of a given family probably have common supersecondary motifs.
  - not all proteins can be generated as a collection of motifs

Evolutionary interpretation

- given a classification – does it reflect evolution ?
  - maybe

Evolution and Classification

- for very similar proteins, easy
- more remote ?
  - maybe
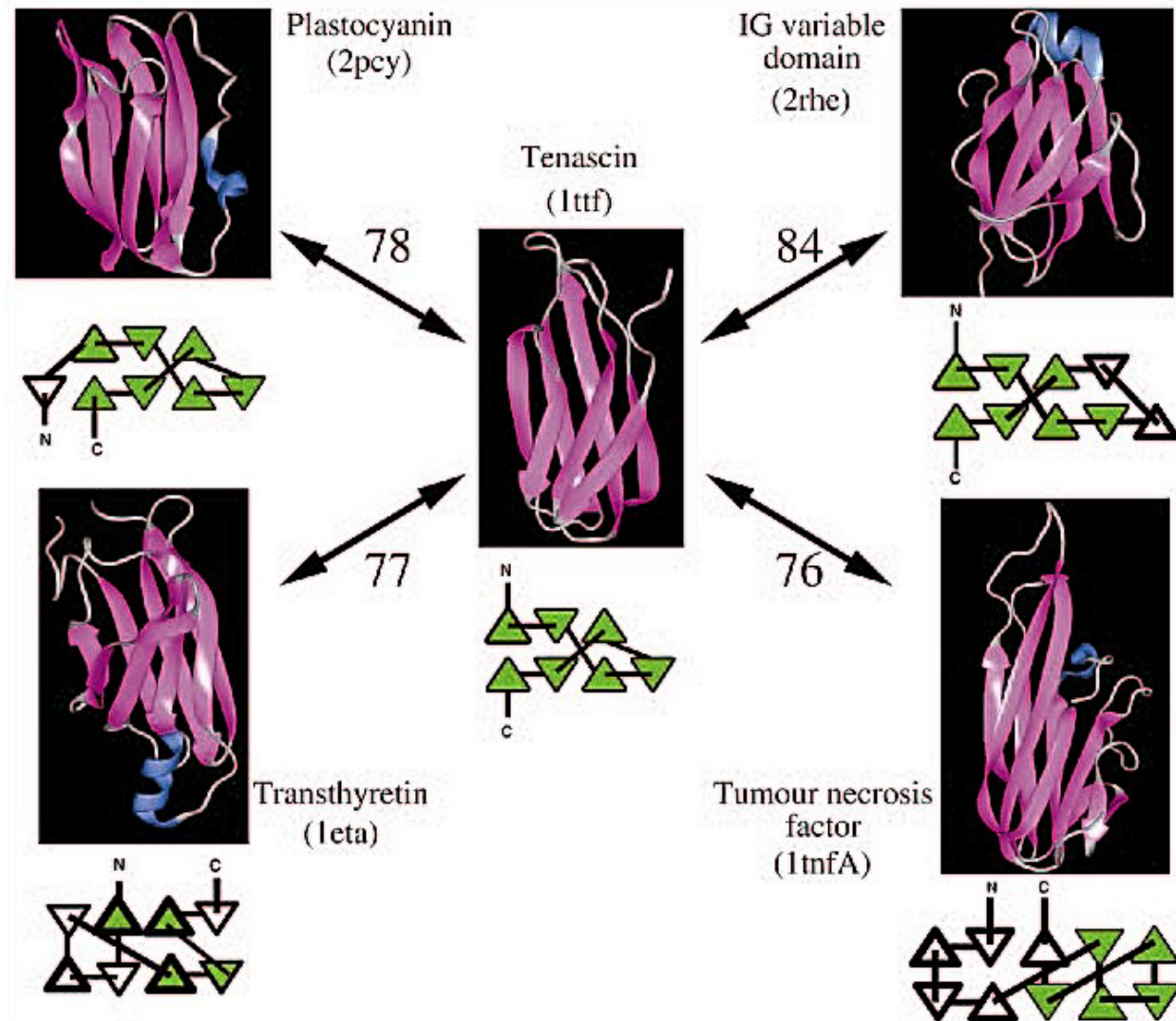
Forget Evolution

- 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 ?
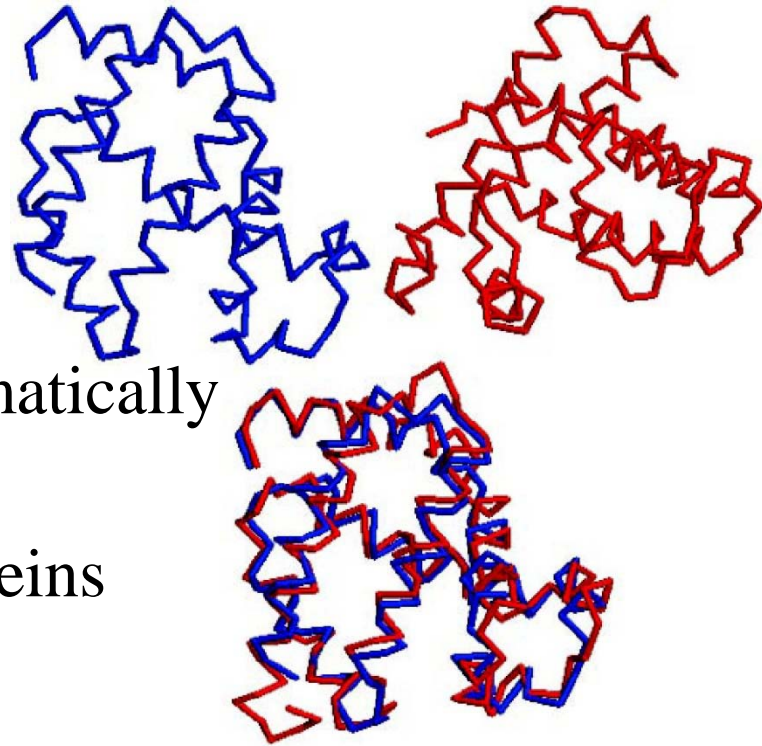
Protein Structure Comparison / Numerical

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

1 Å

- 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
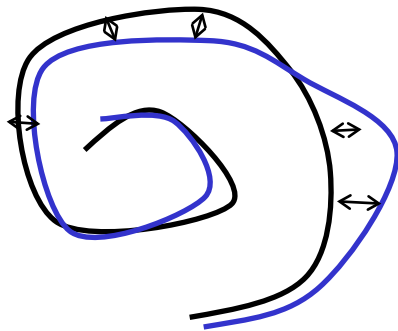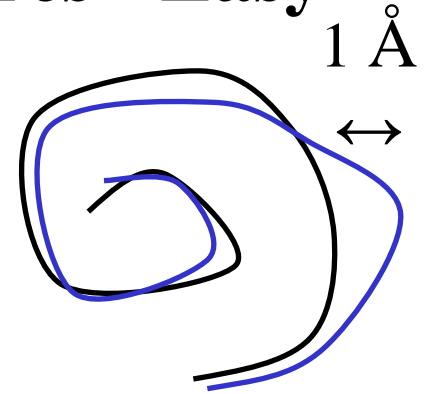
From Standard Deviation to RMSD

Analogy with comparing a set of numbers

- get average (mean)

$$\bar{x} = N^{-1} \sum_{i=1}^{N} x_i$$

$$\sigma^2 = N^{-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

- variance and standard deviation, σ
- apply this to coordinates of
  r and r'

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

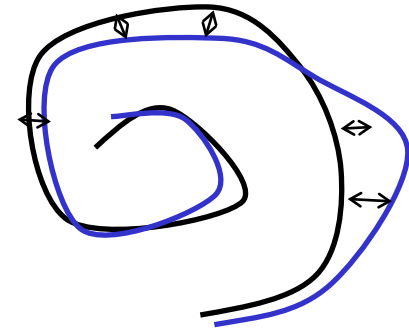$$RMSD = \left(N^{-1} \sum_{i=1}^{N} |\vec{r}_i - \vec{r}_i'|^2 \right)^{1/2}$$

Vital

- formula above, names below
- rms = rmsd = RMSD = root mean square difference
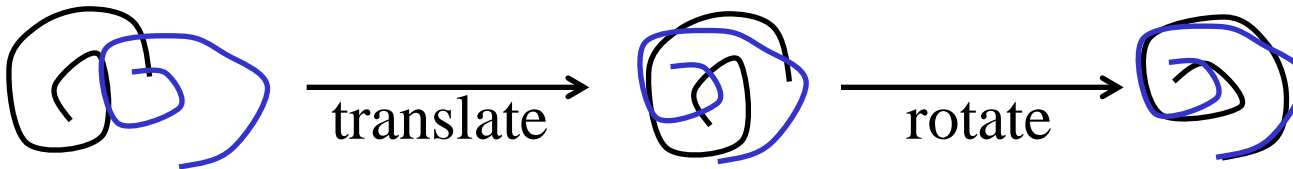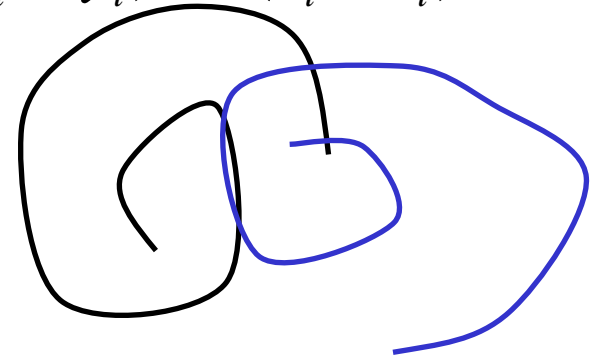
Applying this…

Calculating rmsd

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

- start at one end
- difference between pairs of atoms

$$\left| \vec{r}_i - \vec{r}_i' \right|^2 = (x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2$$

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

translate ⟶ rotate ⟶

Translation and Rotation

translation

- c.o.m. = centre of mass
- subtract difference vector

$$\vec{r}^{c.o.m} = \left(\sum_{i=1}^{N} m_i \right)^{-1} \sum_{i=1}^{N} \vec{r}_i m_i$$

$$\vec{r}_{diff} = \vec{r}^{c.o.m.} - \vec{r}'^{c.o.m.}$$

- rotation
  - messier..
  - find rotation matrix to minimise $RMSD = \left(N^{-1} \sum_{i=1}^{N} \left| \vec{r}_i - \vec{r}_i' \right|^2 \right)^{1/2}$

- summary
  - translate
  - rotate
  - apply formula
- still not finished

Which Atoms ?

What tells me the shape of a protein ?

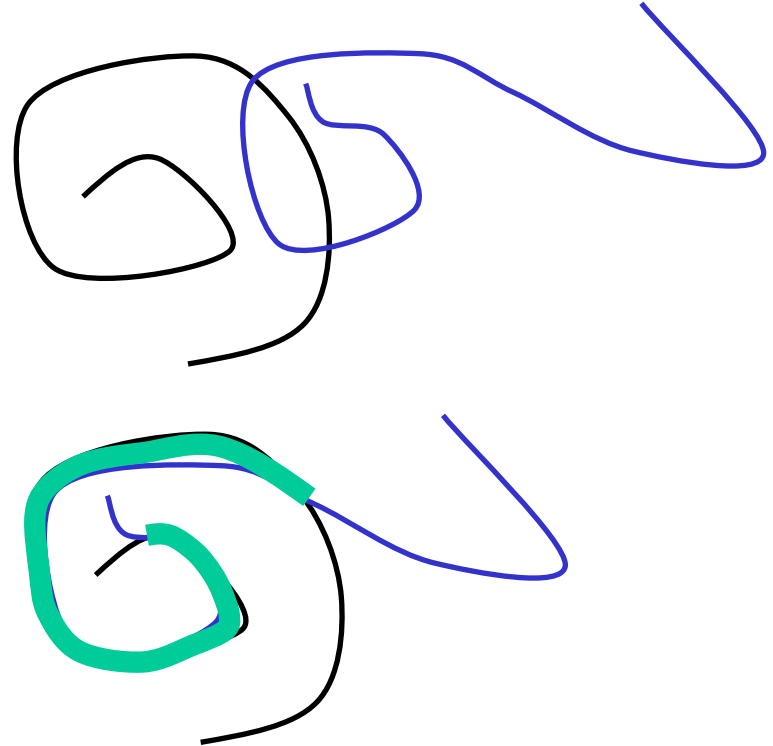- backbone trace

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

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}={|rᵢ-r'ᵢ|} i=1..N
sort {d}

diff= rmsd ({rᵢ},{rᵢ'})
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
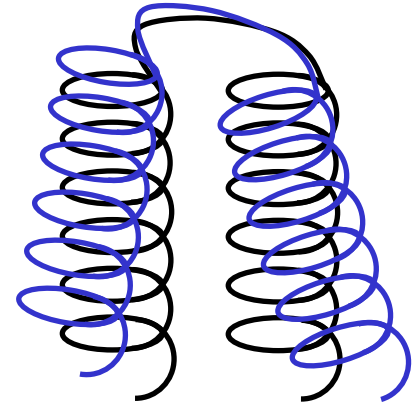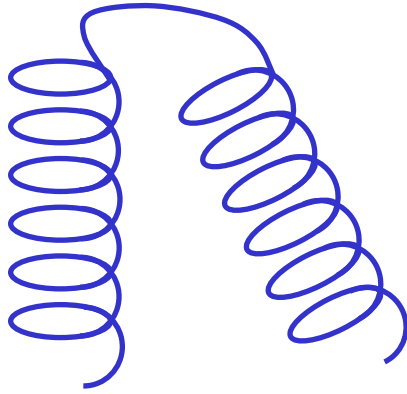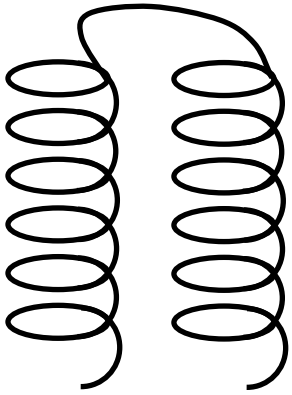 - as number of atoms ↓ calculated *rmsd* ↓

Where we were up to

- Superimposing structures
  - selecting atoms
- Structures of the same size

Where to

- distance matrices for comparison
- finish comparison of structures
  - structures of different sizes
- introduction to modelling
  - interesting bits
- time for some revision next week

Why Not Use *rmsd*

- helices identical, fold identical *rmsd* ?

- superposition requires rotation, affects all atoms

- big *rmsd*, but structure has hardly changed
- do not see that helices are identical
- solutions
  - use angles (other problems)
  - distance matrices

Distance Matrices With Numbers
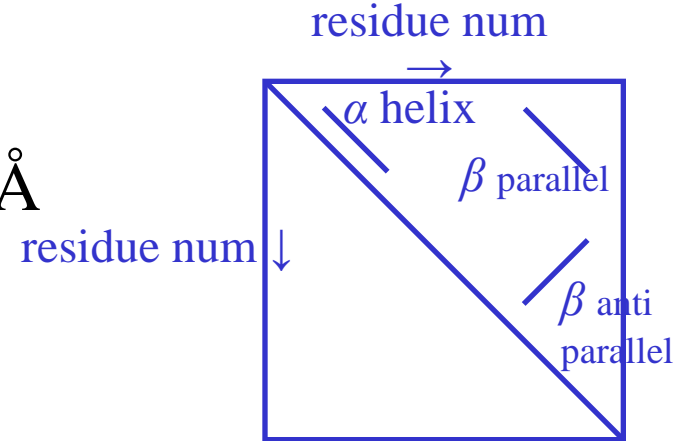
Another characteristic of structures

- 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

residue num →

α helix

β parallel
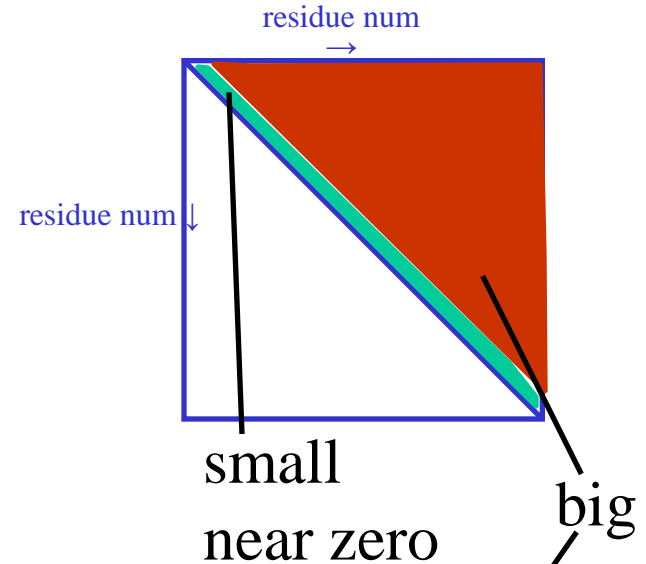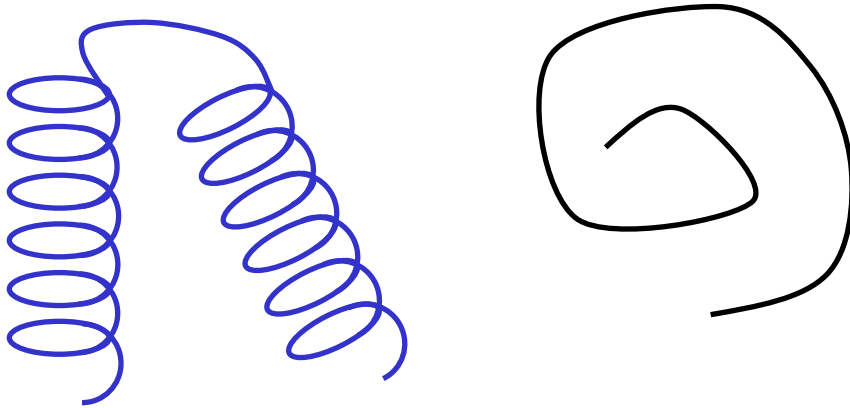
residue num ↓

β anti parallel

Distance matrix for comparing structures

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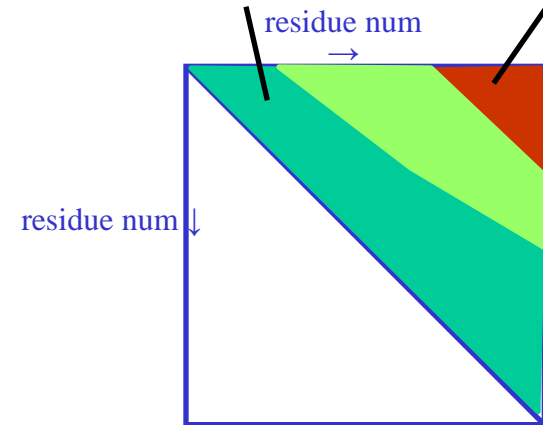 - look at the difference of distance matrices

residue num →

residue num ↓

-

residue num →

residue num ↓

=

residue num →

residue num ↓

0
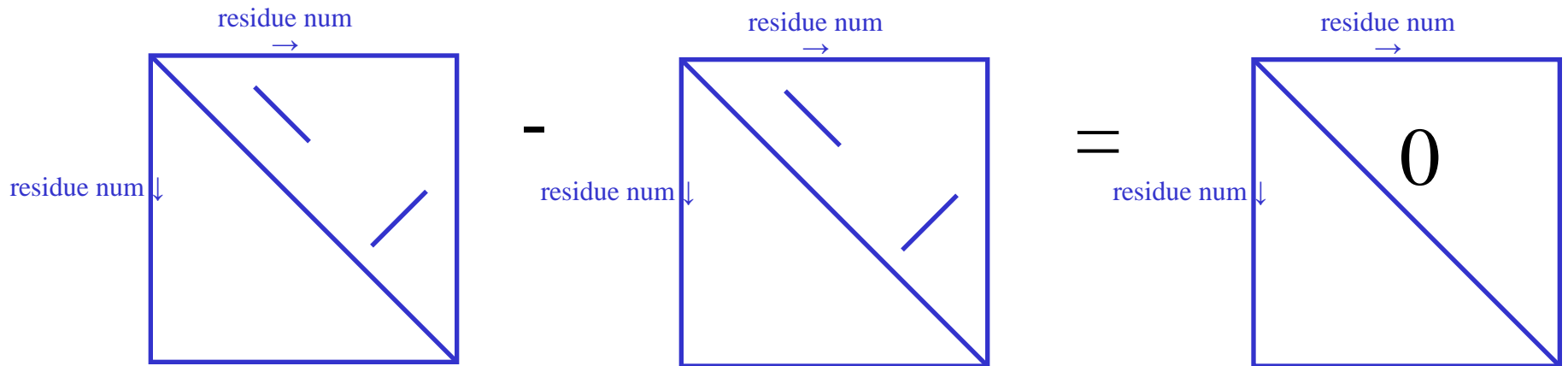
Comparing Distance Matrices

- consider two very different structures

residue num →

residue num ↓

small
near zero

big

- two related structures

residue num →

residue num ↓
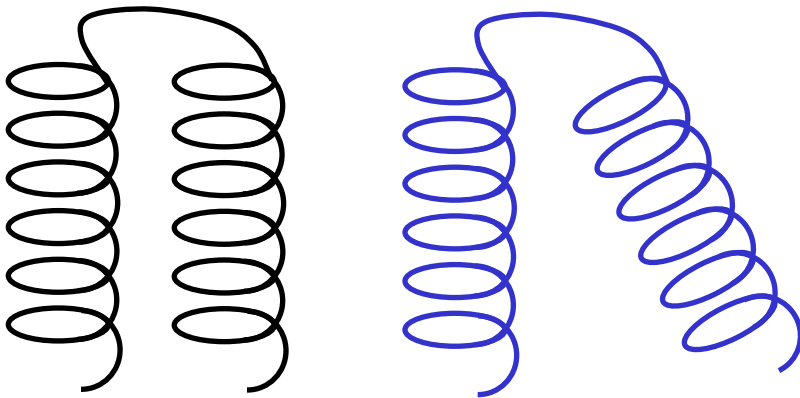
- pictures are better than any single measure, but…

From Distance Matrices to Single Number

For lots of comparisons, single number is more convenient
- root mean square (*rms*) difference of distance matrices
  - define distance between C^α atoms *i* and *j*

$$d_{ij} = \left| \vec{r}_i - \vec{r}_j \right|$$

- *rms* of distance matrices measure is

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

- 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
- 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 Different Proteins

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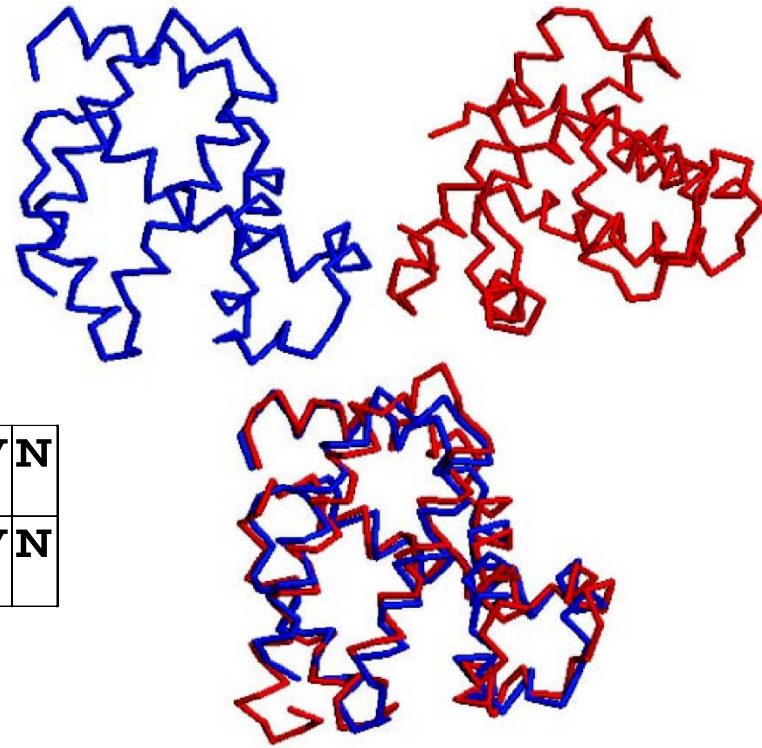
- green residues - backbone atoms
- pink residues – ignore
- is this useful for similar proteins ? very (rat vs human haemoglobin)
- for very different proteins ? no

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
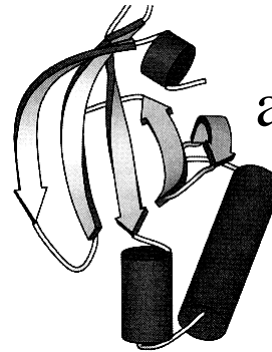   - combinatorial explosion
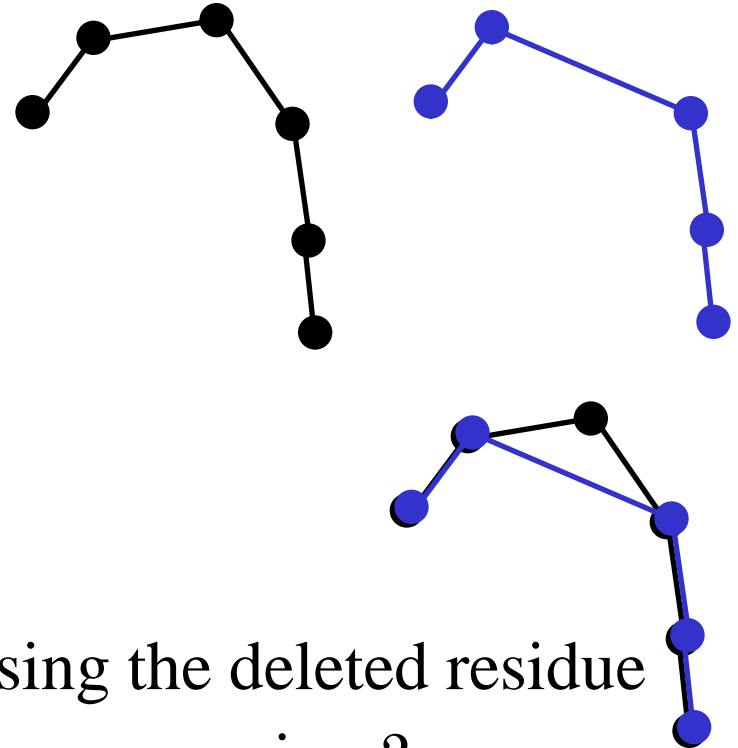
Strategies For Comparing Different Structures
1. use secondary structure

- Combinatorial explosion is the problem
  - reduce size of problem
 - use elements of secondary structure
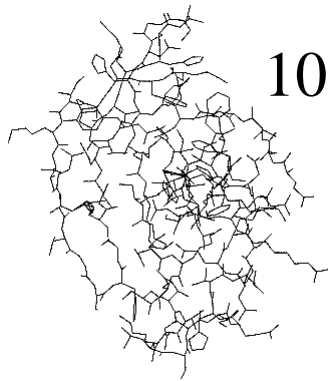
10^2-10^3 atoms

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)
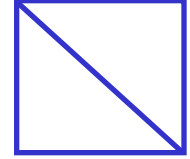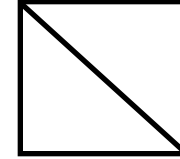 - protein is no longer a string of residues nor a whole structure

whole protein
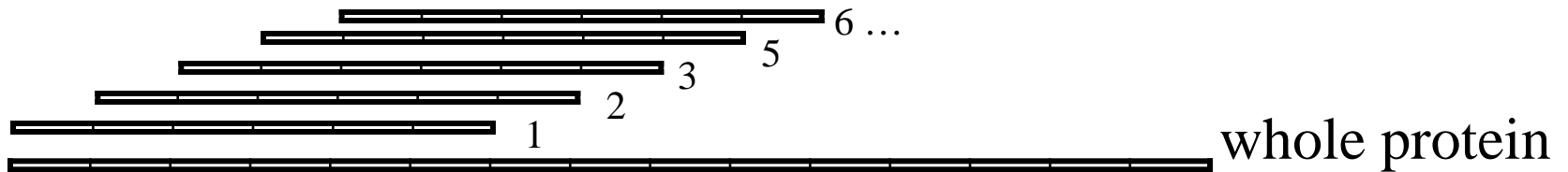
- each fragment is a little distance matrix

Fragment Based Comparison

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

protein 2
fragments ↓

protein 1
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

- 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 1000 proteins
- find structurally similar pairs
- plot sequence identity

may not be found by
sequence methods
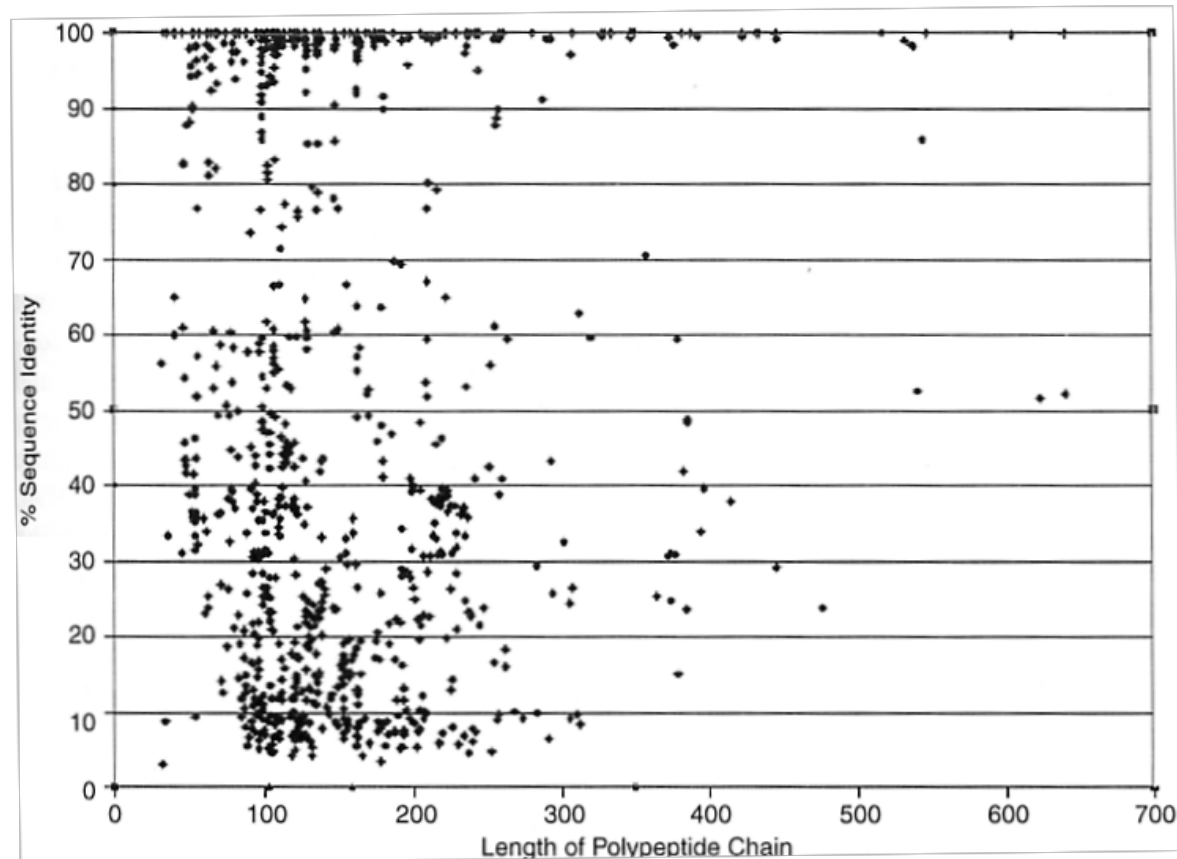
Summary of All Protein Comparisons

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