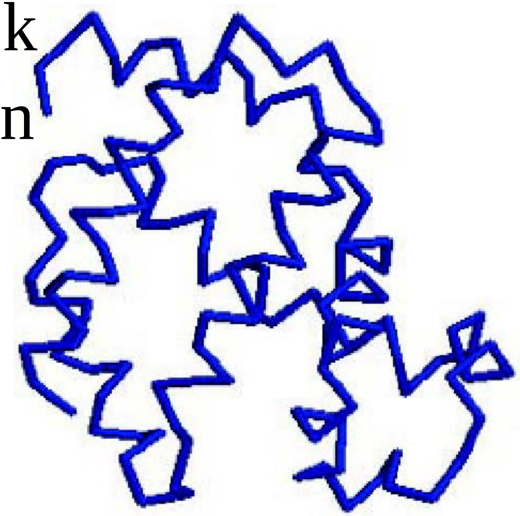
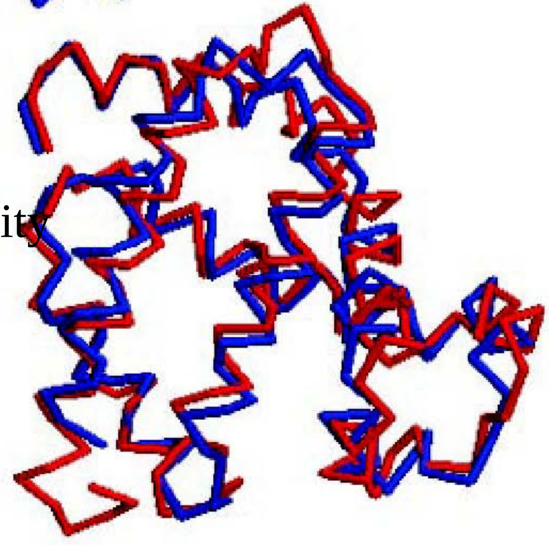
Comparing protein structures

- Fun problem – no textbook
- NOT sequence comparison

- why does it matter ?



1ecd, 1mbd
no significant
sequence identity



Blankenbecler, Ohlsson, Peterson, Ringne PNAS (2003)100, 11936–11940

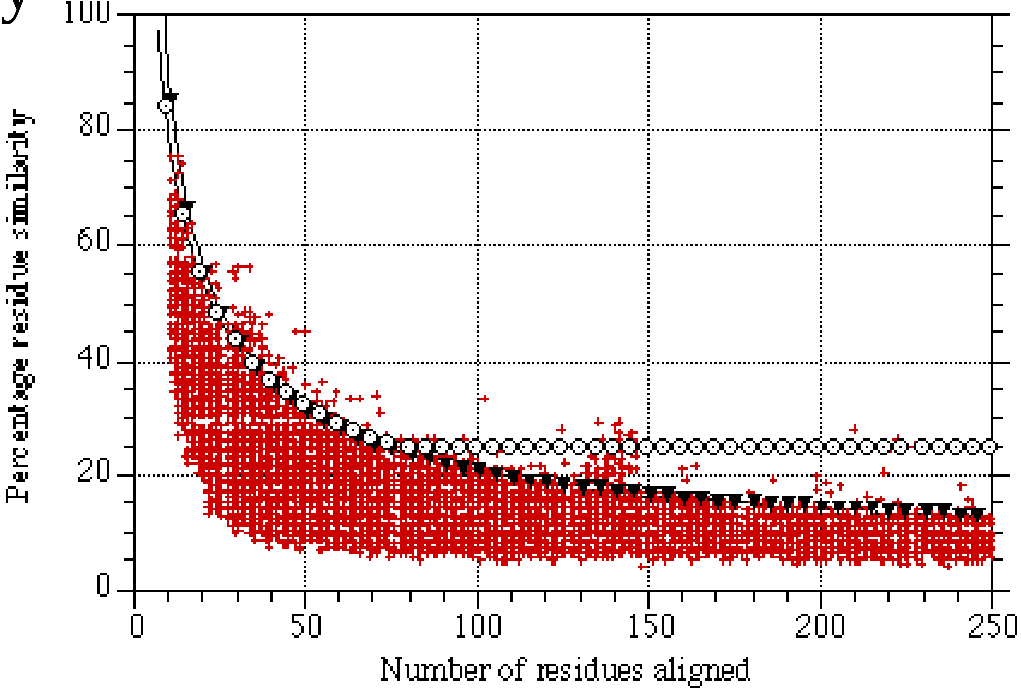
Structure versus sequence comparisons

- Protein Databank ≈ 6.8×10^4
- 90 % sequence similarity ≈ 1.7×10^4
- different shapes 2 to 5×10^3
- 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

Thresholds of sequence similarity

Take a set of pairs of proteins

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



- 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"

Example family

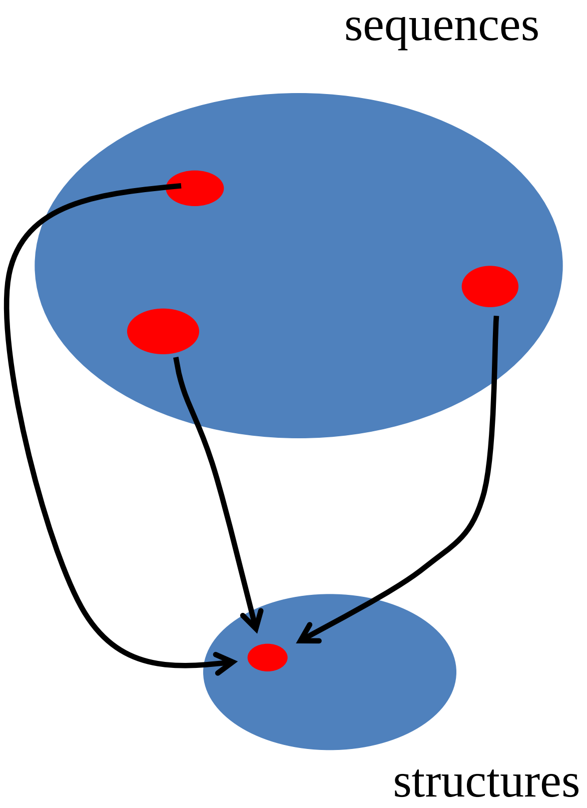
- example, neighbours of 1cun chain A
  - look at sequence identity (%id)
  - alignment length (lali = number of residues)
  - root mean square diff in Å

```
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
```

http://ekhidna.biocenter.helsinki.fi/dali/start

Sequence vs structure space

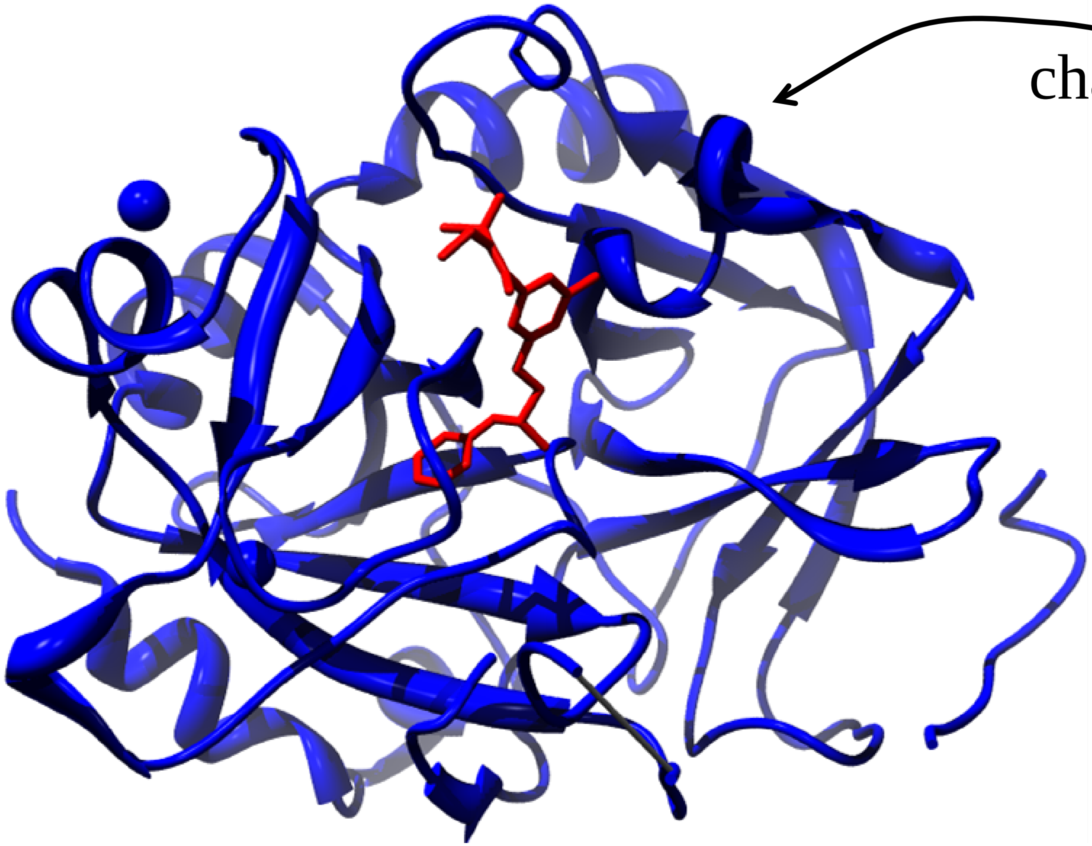
- there are 1000's of such families
- summarise
  - similar sequences
    - similar structures
  - very different sequences
    - similar or different structures
- why ?



Why ?

- typical – low sequence identity, similar structures
- physical reasons
  - compactness, stability
  - advantages of H-bonded conformations
- history / evolution
  - evidence
    - theoretical – geometric constructions
    - chemical – construction of artificial protein(s)
  - imagine all proteins evolve from some original molecule …

why can sequence change ?



change here

residue changes ? OK

structure changes ?

Bad

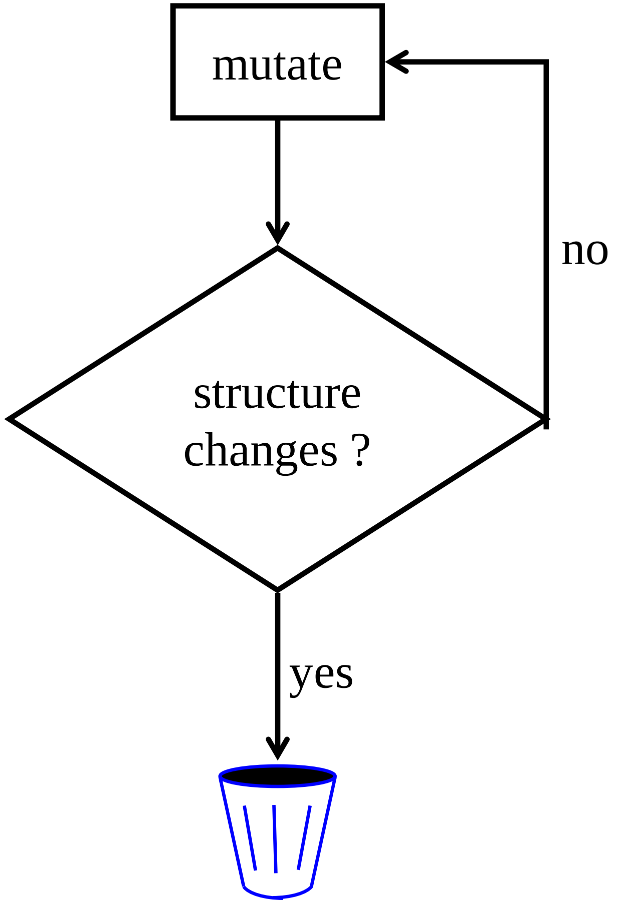
- a view of molecular evolution…

2j9m, 2cdk + aminopyridine

Evolution

mutate continuously

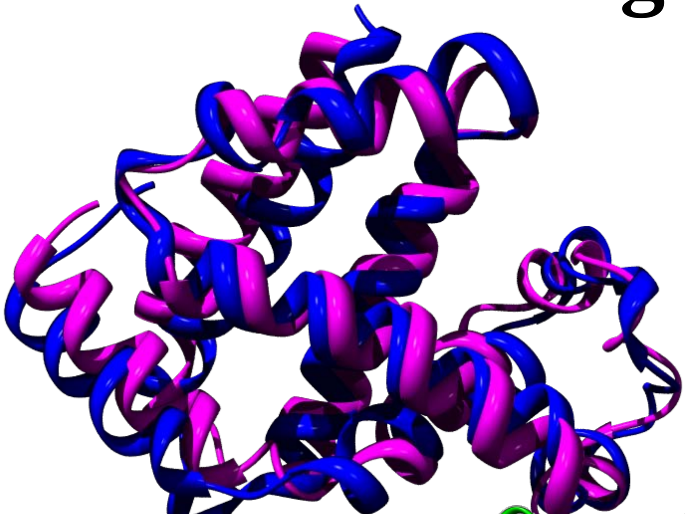
- mutations which are not lethal
  - may be passed on (fixed)
- if structure changes
  - protein probably will not function
  - not passed on



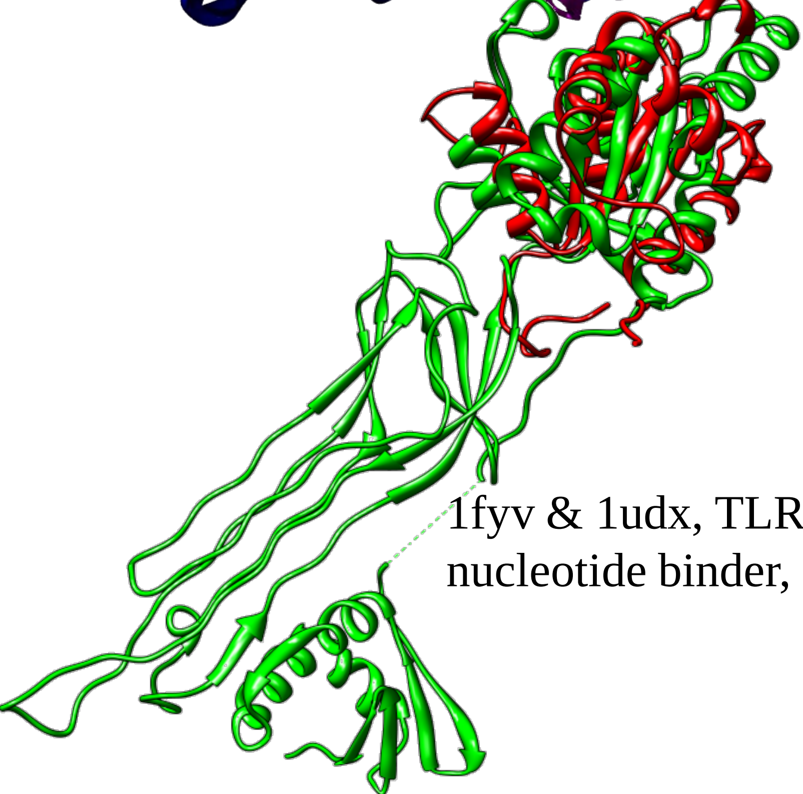
Result

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

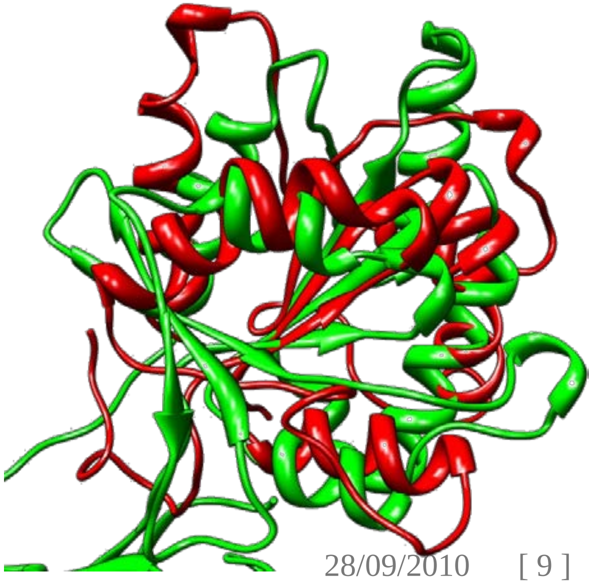
Meaning of structural similarity



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

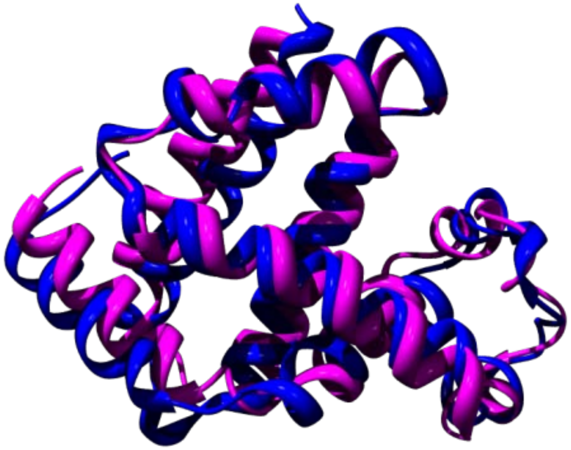


1fyv & 1udx, TLR receptor and
nucleotide binder, 9 % sequence id

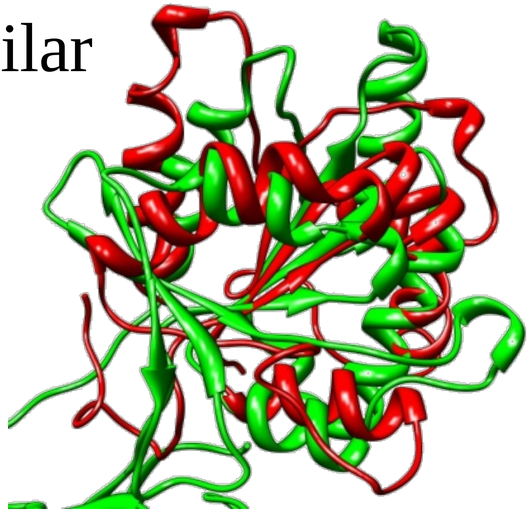


Quantifying similarity

very
similar



similar



- quantifying this ?

- assume we have an alignment of residues (later)
  - for each C^{α} in protein 1, corresponding C^{α} protein 2
- simplest / most common measure is *rmsd* (root mean square deviation) of C^{α} coordinates..

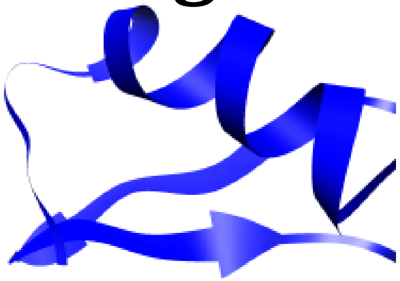
rmsd

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

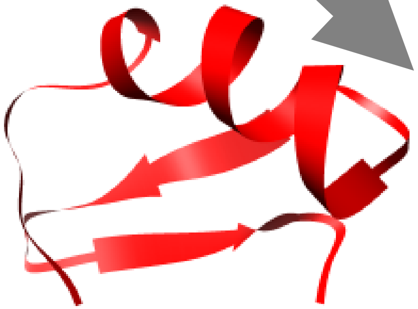
$$r_{rmsd} = \left(\frac{1}{N_{res}} \sum_{i=1}^{N_{res}} \left| \vec{r}_i^{\,a} - \vec{r}_i^{\,b} \right|^2 \right)^{1/2}$$

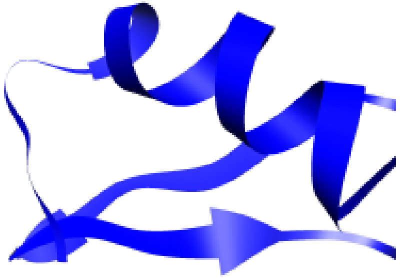
- everyone remembers Å (0.1 nm)
- many alternatives
    - *rmsd* of internal distance matrices
    - "gdt" fraction of atoms superimposible below thresholds
    - …

finding matching atoms

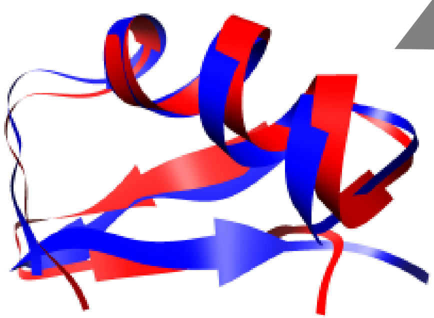


- if we had same number of atoms - easy
  - two conformations of one protein



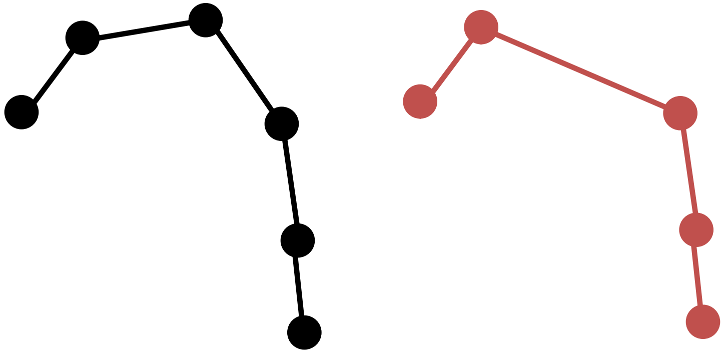


rotation and translation

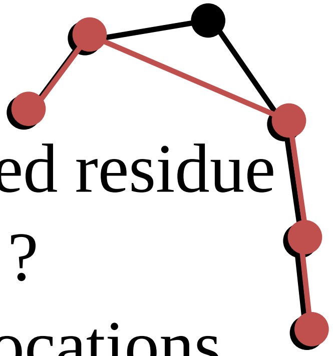


- analytical method

finding aligned atoms



- NP complete !
- how difficult ?



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

- strategies

how to align protein structures

- NP complete, dozens of approaches
  - all can be made to fail

- seeded methods – yuk not today
- cheat and use sequence
- overall superposition
- fragment based

cheating and using sequence

- example implementation in "chimera"

- Assumption
  - sequence identity is weak, but
  - I believe two proteins are related

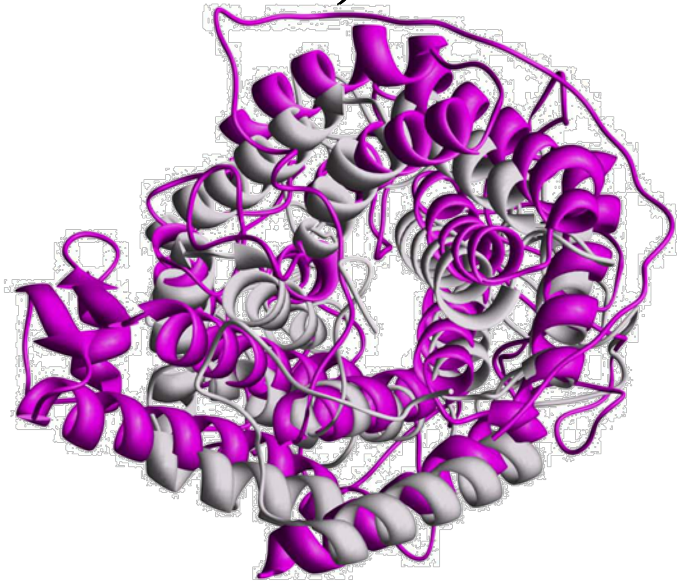
  - maybe the sequence alignment will be roughly correct
- can be used to get corresponding atoms
- from N residues, get N_{common} shared
- problem
  - we know identity is weak alignment will be bad
  - especially around loops, insertions, deletions

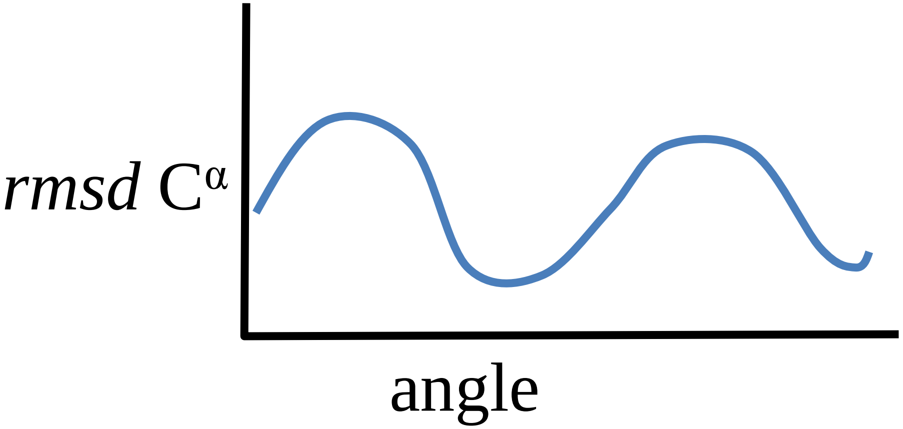
overall superposition

- philosophy
  - centre of mass is easy (average of C^{α} coordinates)
  - translation – seems easy

  - rotation ? – bit harder

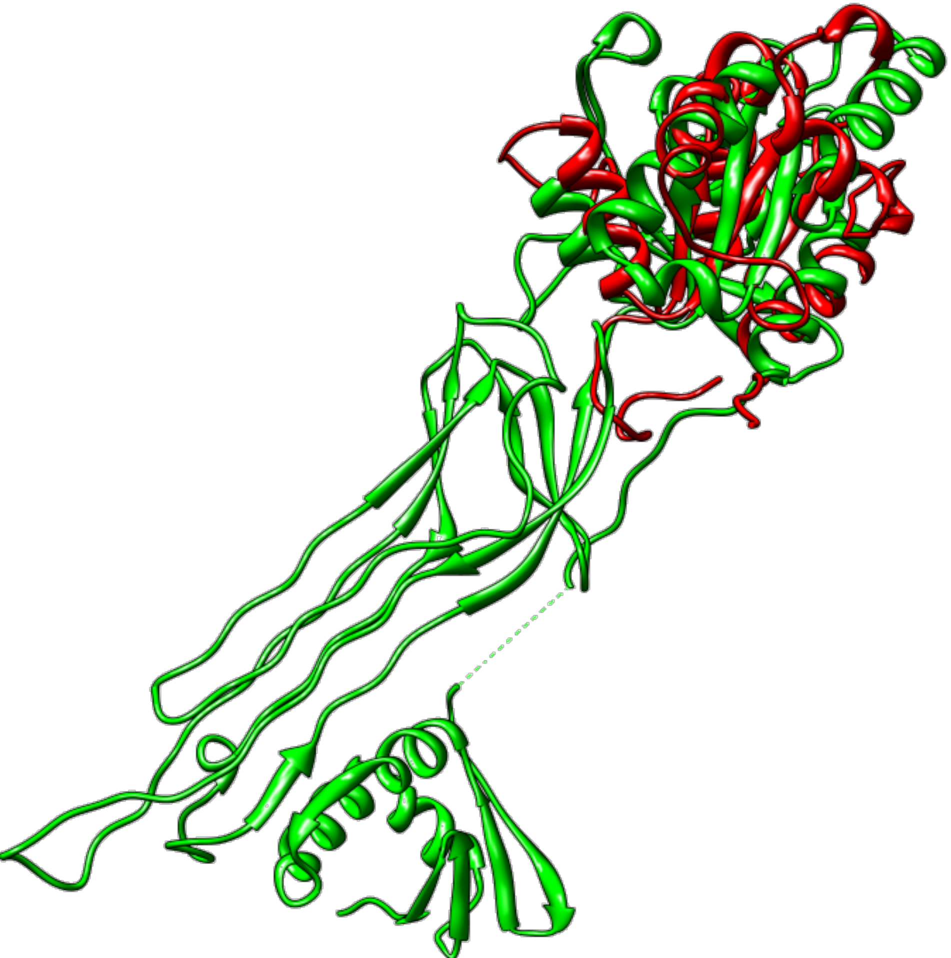
- friendly function to optimise ?
  - 6 degrees of freedom

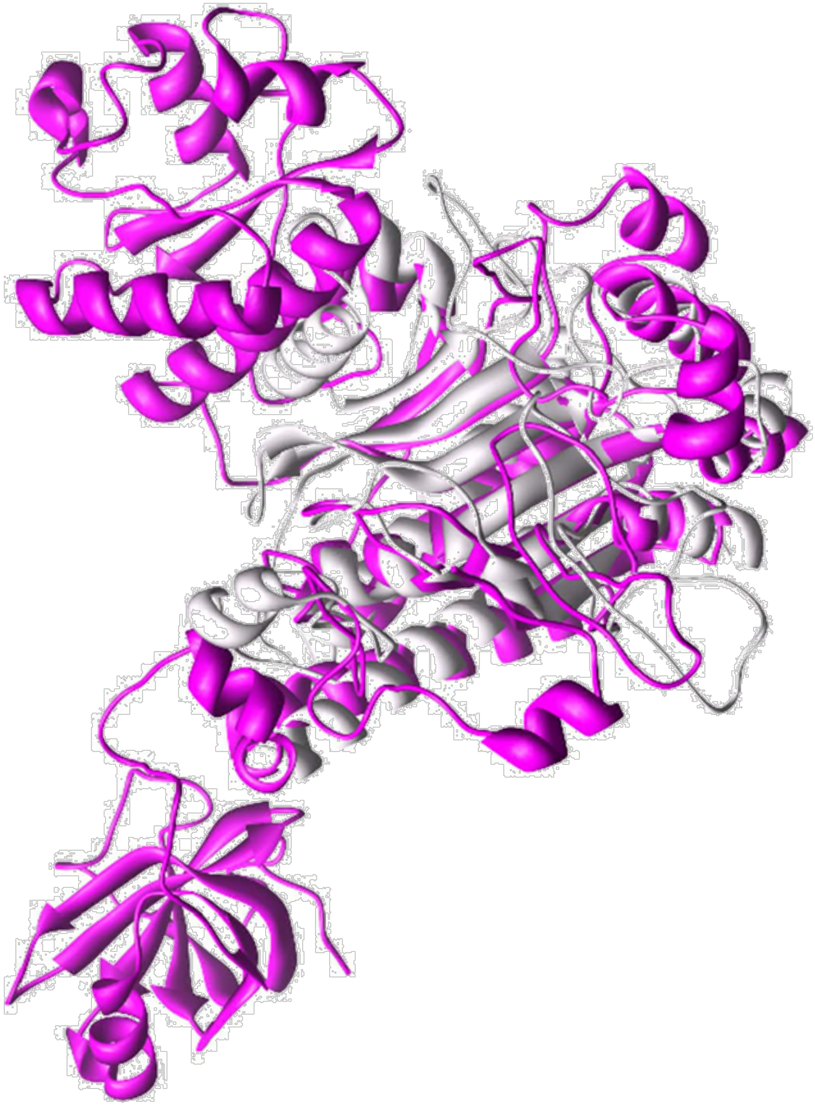




Overall superposition (broken)

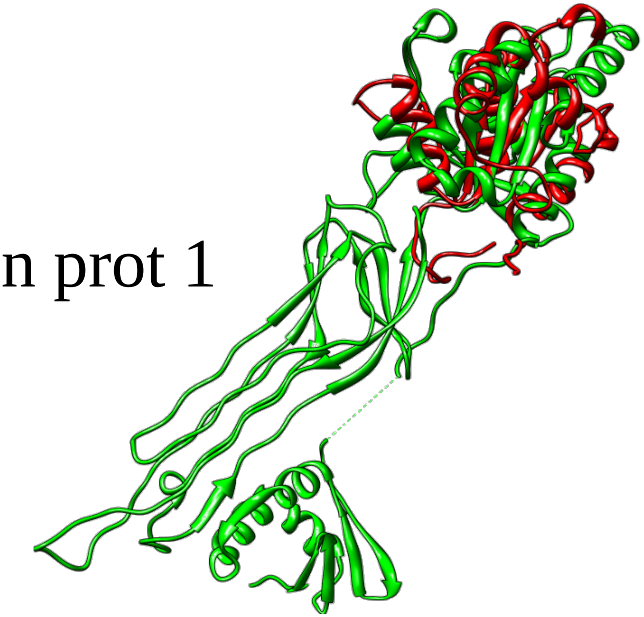
- the centre of mass is not relevant





a fragment based strategy

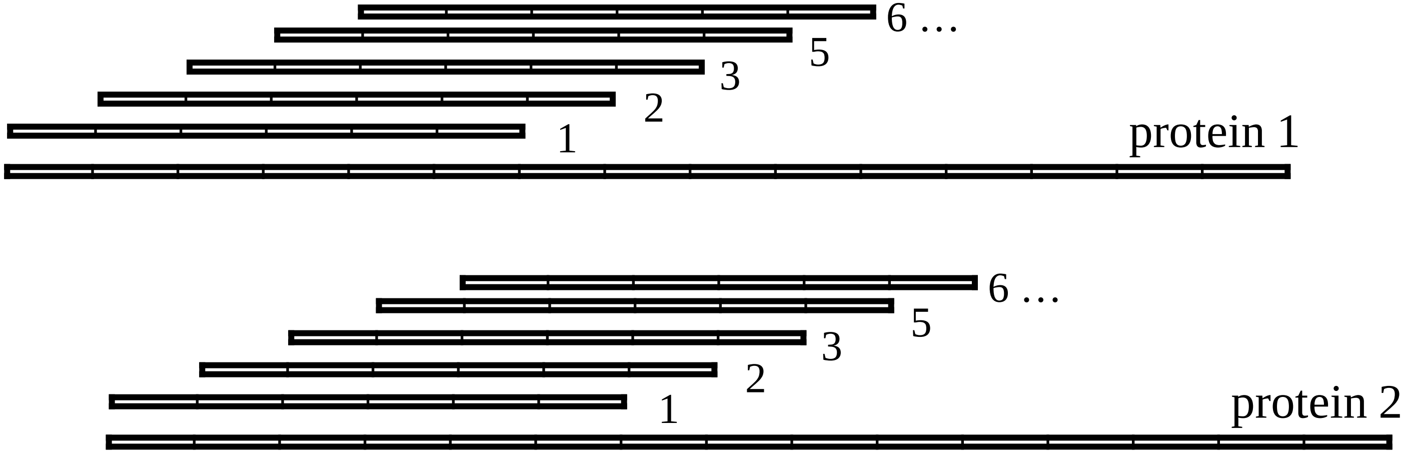
- want..
  - find common regions
  - some way of comparing each C^{α} in prot 1 with C^{α} in prot 2
    - some kind of structural label



protein 1

		1	2	3	..
	1				
	2				
protein 2	3				
	4				
	..				

fragments to similarity matrix



protein 1

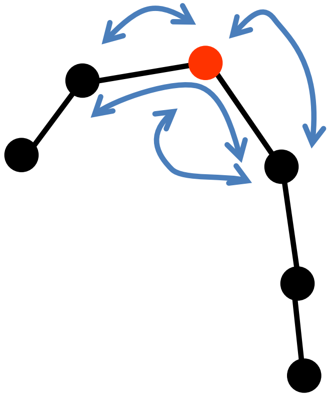
		1	2	3	..
	1				
	2				
protein 2	3				
	4				
	..				

characterising the local environment distance matrices

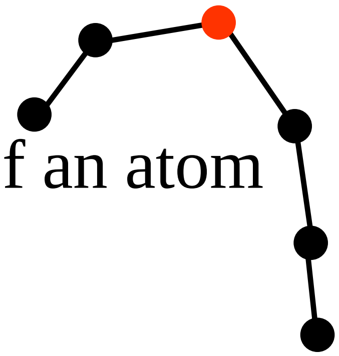
- secondary structure ?
    - not very specific, thresholds
- distance matrices..
    - given a C^{α} what is the pattern of neighbours ?

Å

neighbour	1	2	3	4…
1	0	4.2	5.5	5.0
2	4.2	0	6.0	5.5
3	5.5	6.0	0	4.4
4…	5.0	5.5	4.4	0



distance matrix comparison



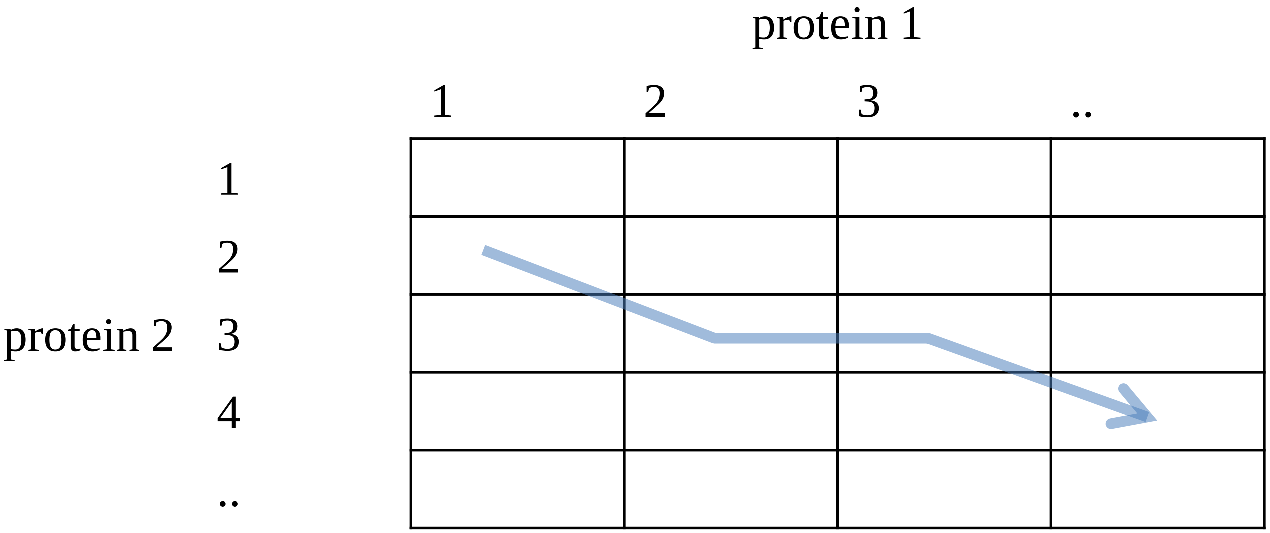
- given two matrices
  - each characterises the environment of an atom
  - compare with *rmsd* – like measure

$$d_{rmsd} = \left(\frac{2}{N_{res}\left(N_{res}-1\right)} \sum_{j>i}^{N_{res}} \sum_{i=1}^{N_{res}-1} \left(d_{ij}^{a} - d_{ij}^{b}\right)^{2} \right)^{1/2}$$

- now make a similarity matrix
  - elements are matrix similarities

		protein 1			
		1	2	3	..
protein 2	1				
	2				
	3				
	4				
	..				

aligning protein structures



protein 1

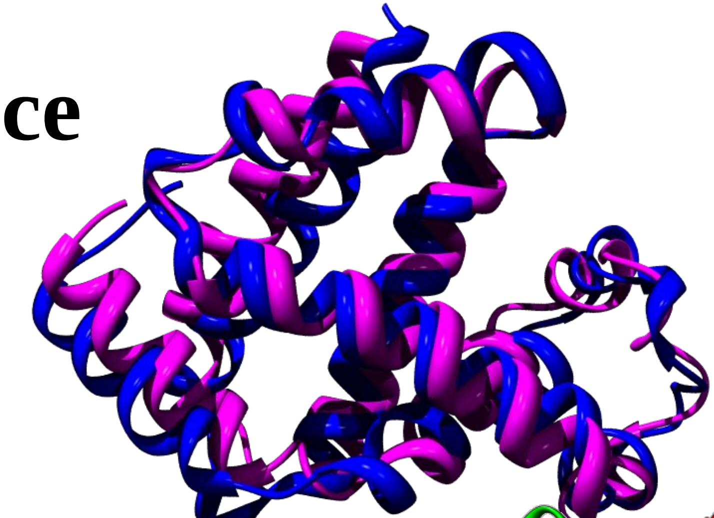
1 2 3 ..

protein 2

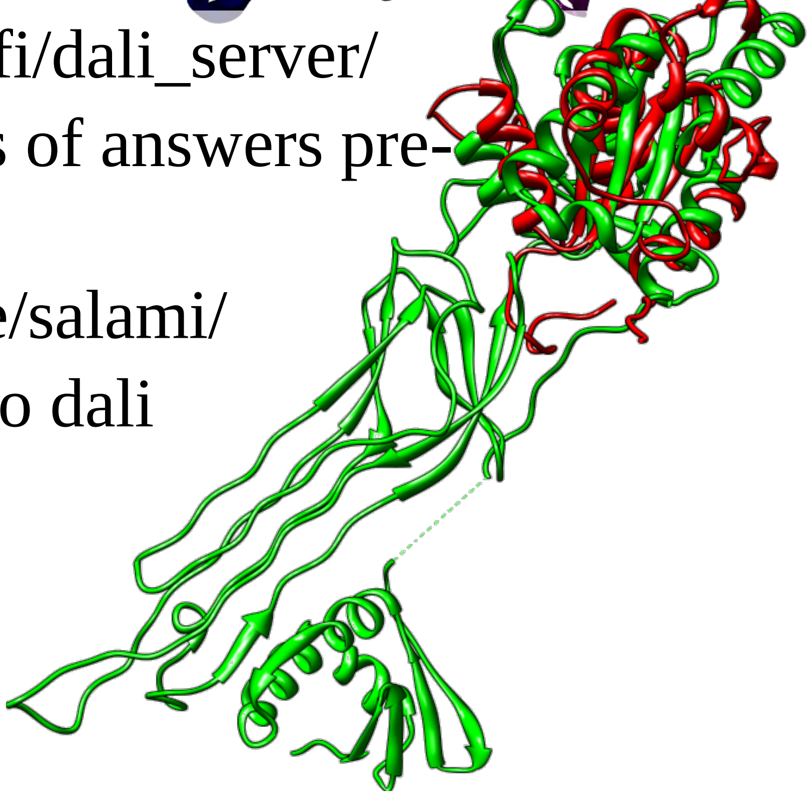
1 2 3 4 ..

- optimal path through matrix is pure structure based alignment

In practice



- dali   ekhidna.biocenter.helsinki.fi/dali_server/
  - very good results, not fast, lots of answers pre-calculated
- wurst public.zbh.uni-hamburg.de/salami/
  - very fast, very similar results to dali
  - used for most of pictures here



In practice

- only relevant when structures are known
  - 6.8×10^4 versus 10^7 sequences
- will detect more remote similarities
- structural genomics / function prediction

- applications
  - searching PDB for similarities
  - phylogeny based on structure …

- Coffee