Protein structures and comparisons

Andrew Torda 67.937 Bioinformatik, Mai 2008

Ultimate aim

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

On the way..

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

- Later
  - classifications of proteins

Sequence and structure similarity

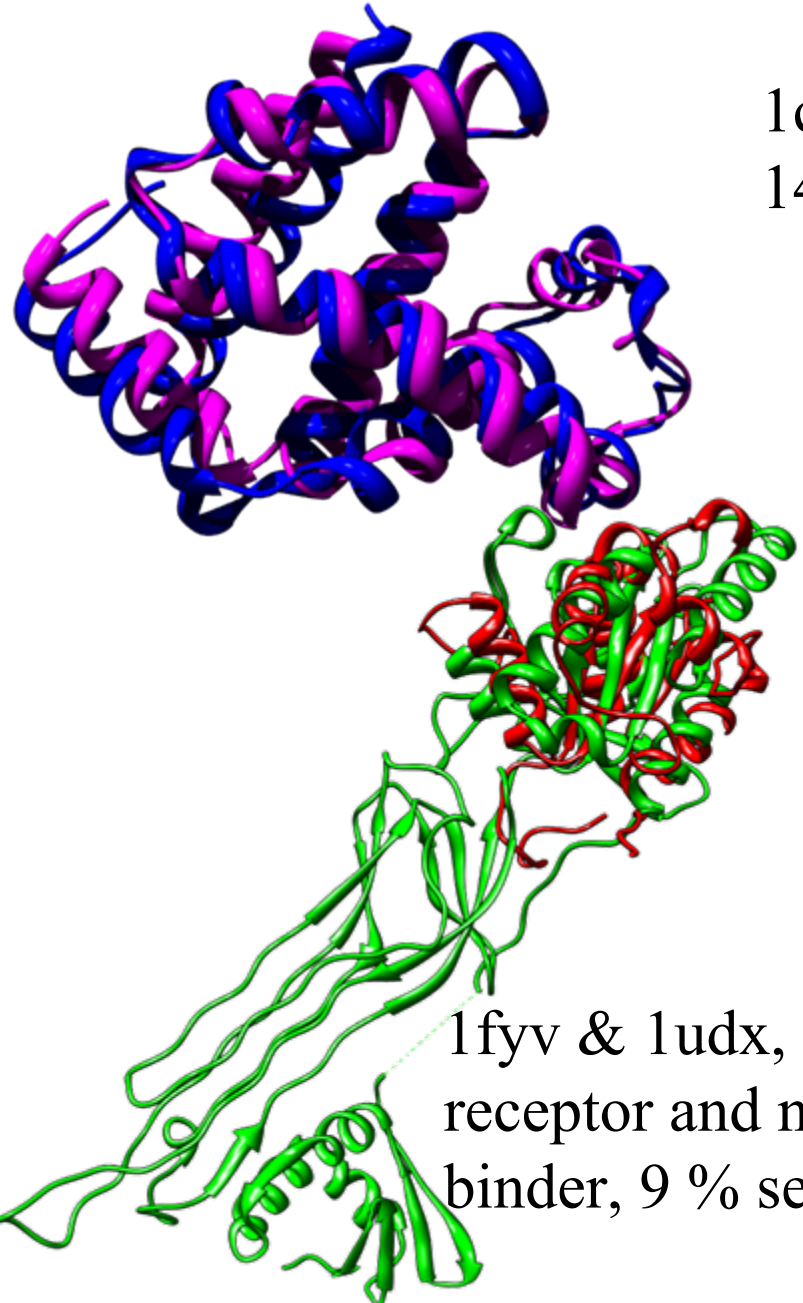
Claim from before

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

Question

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

Remote similarities

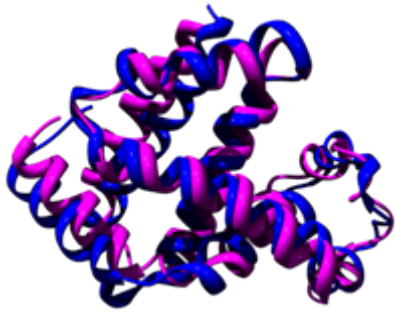


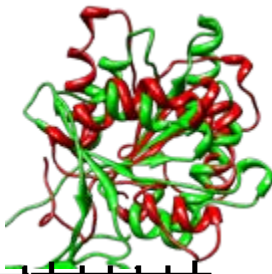
1cbl & 1eca (haemoglobin & erythrocruorin)
14 % sequence id

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

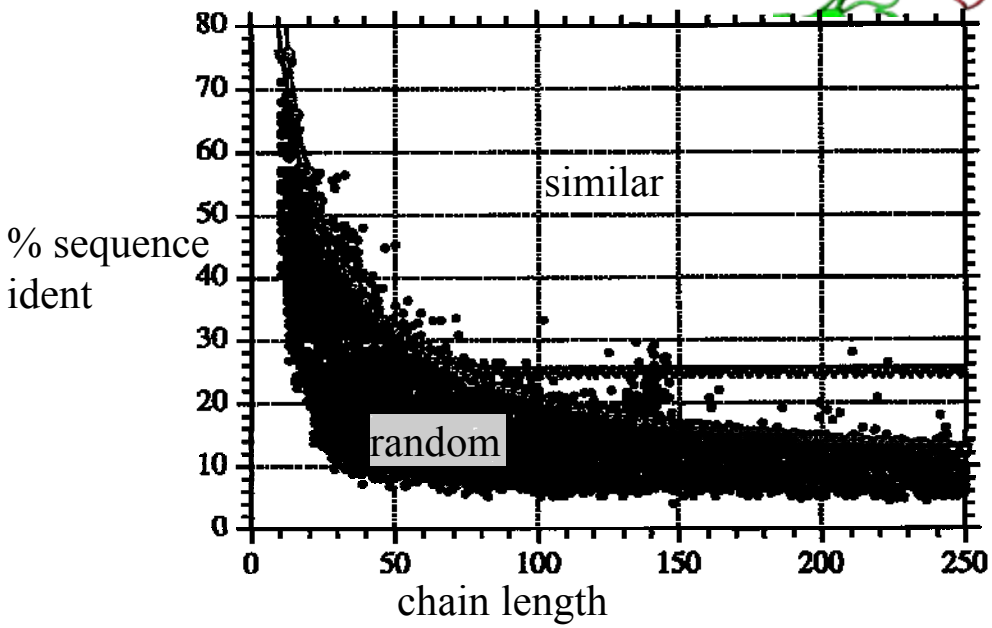
No sequence similarity – similar structures

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





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



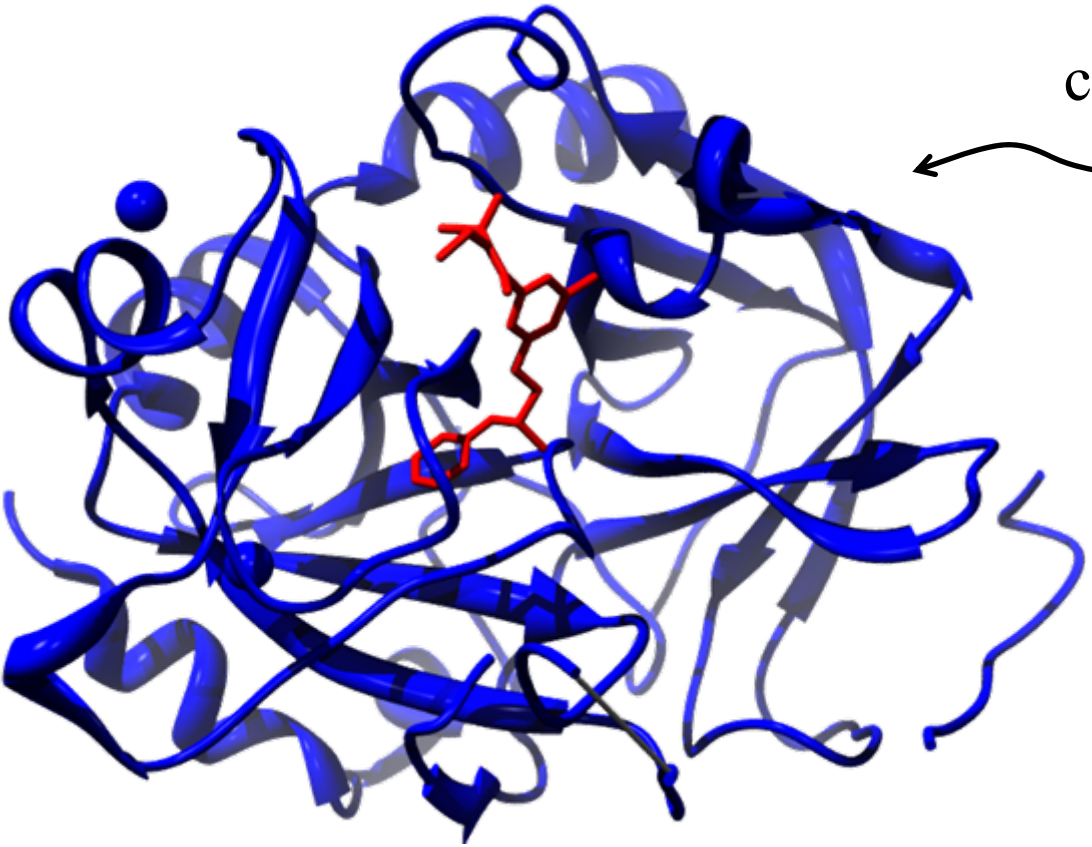
Rost, B.Prot Eng, 12,85–94, 1999

Structure versus sequence similarity

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

- protein function depends on having groups in orientation in space

Why can sequence change



change here

residue changes ? OK

structure changes ? Bad

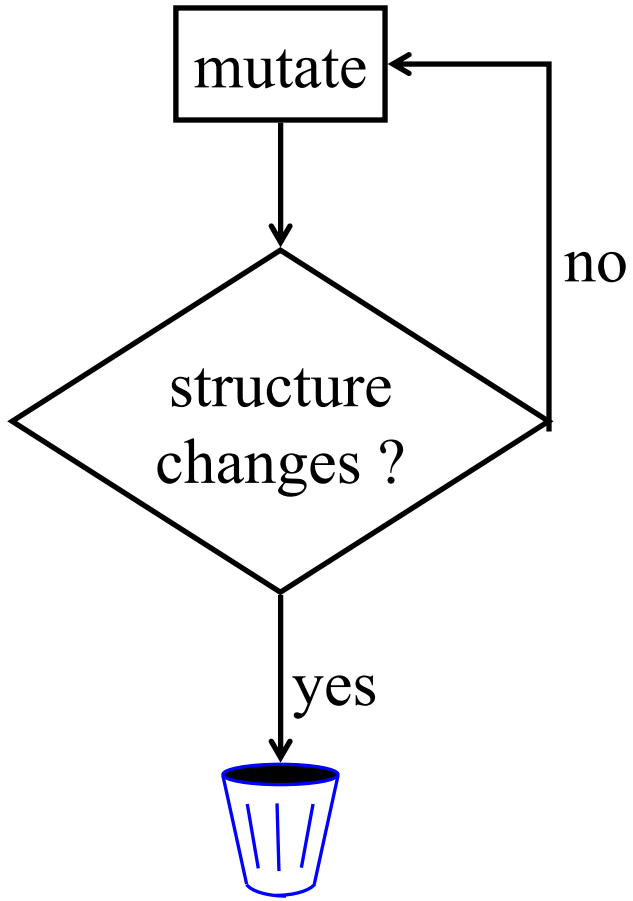
- a view of molecular evolution…

2j9m, 2cdk + aminopyridine

Simple view of molecular evolution

mutate continuously

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



Result

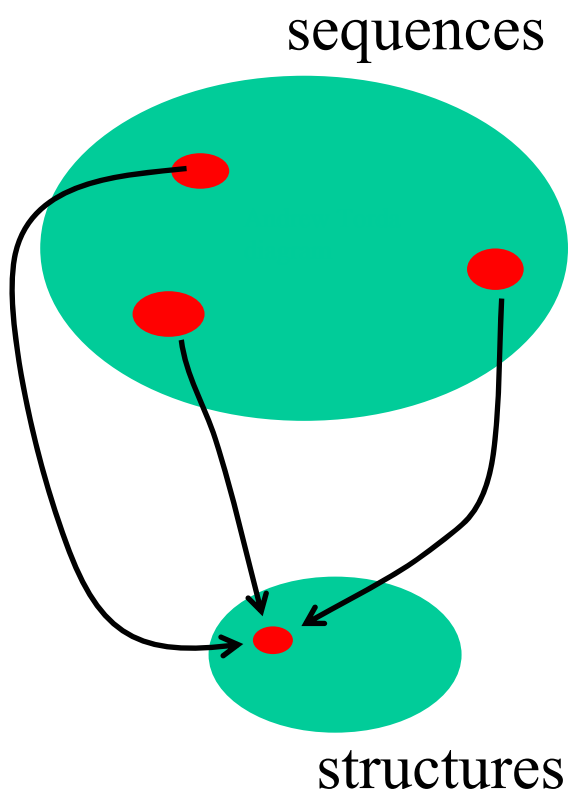
- evolution will find many sequences
  - compatible with structure
  - compatible with function
- how else would we see this ?

Sequence vs structure evolution

Sayings..

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

- Truths…



Practical Consequences

Sequences of proteins are nearly always known

- similar sequence
  - usually similar structure, similar function
- sequences not (obviously) related
  - maybe similar structure
  - maybe similar function

- What if structures are known ?

Sequence and structure similarity

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

- summarise from a different point of view

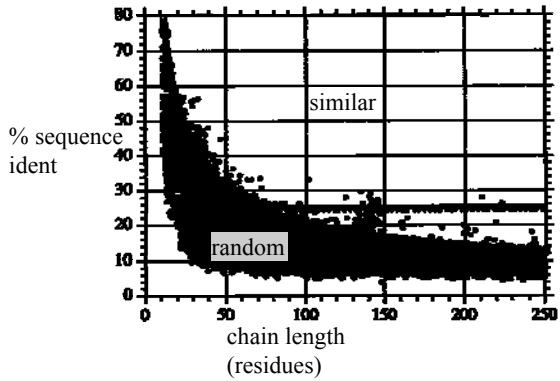
Sequence vs structure similarity

When comparing proteins

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

Similar sequences

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



Similar structures, different sequences

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

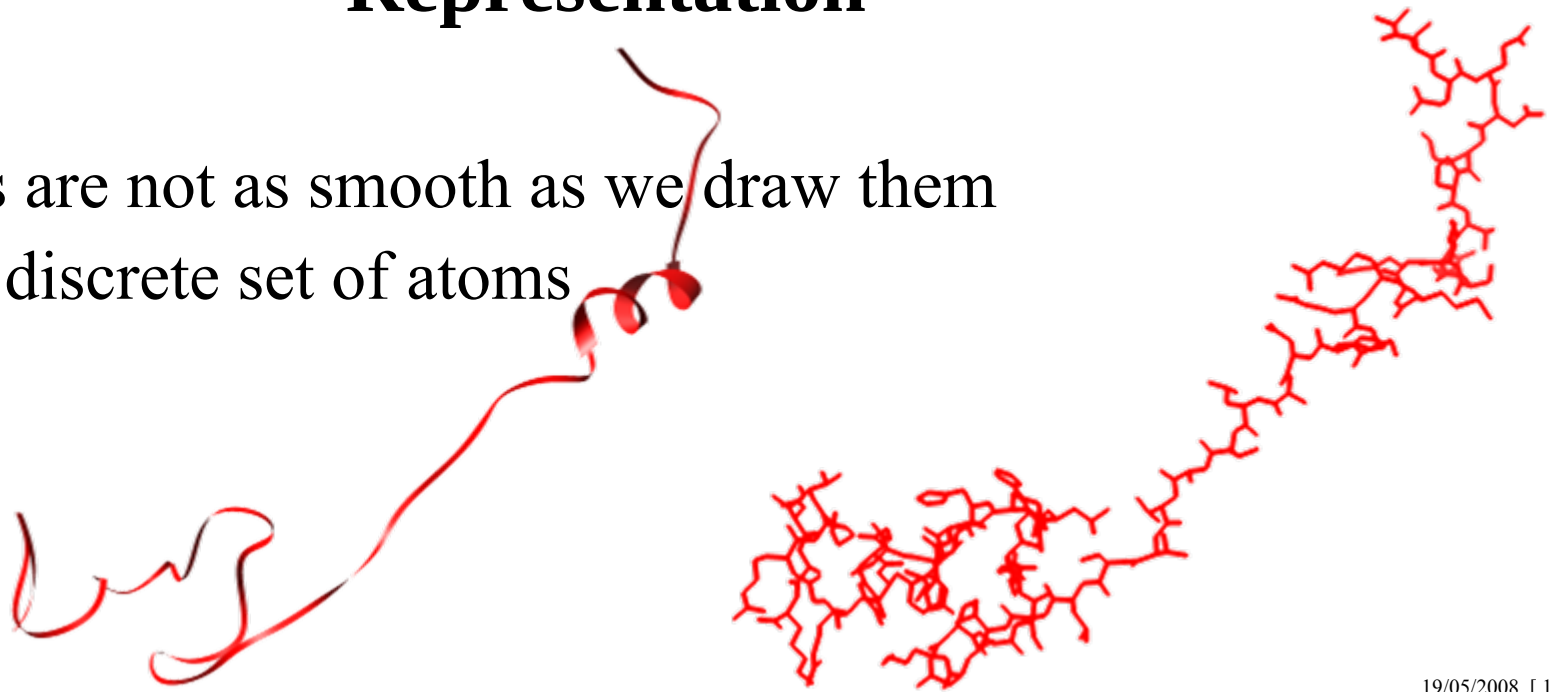
- what do we mean by similar structures ?

Comparing proteins

- Representation of proteins
- comparison
- classification (later)

Representation

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



2vhm chain 0

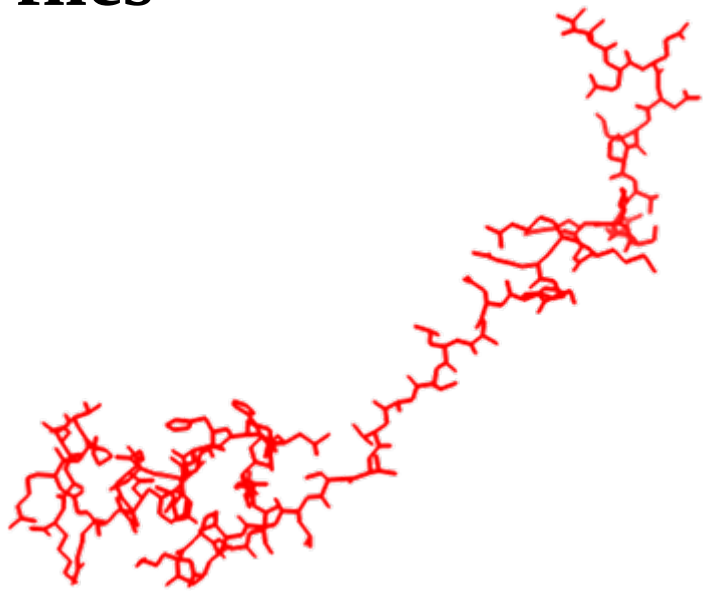
Protein coordinate files

Important detour - Protein data bank (www.rcsb.org)

- only significant database of protein coordinates
- deposition of coordinates – often requirement of publication
- > 50×10^3 structures
  - huge redundancy (> 500 T4 lysozyme)
- X-ray crystallography ≈ 85 %
- NMR ≈ 14 % (more in smaller proteins)

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

Protein coordinate files



What would you expect ?

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

PDB File

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

x y z

- Note coordinates
  - three decimal places – often 5 significant digits
- come back to this in a few Folien

PDB File

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

residue

mobility

How accurate are coordinates ? X-ray

- 3 decimal places ? (10^{-3} Å) 5 significant digits ?
- typical resolution in a protein file ? 1 to 2.5 Å

- what is "*x* Å resolution" ?
  - two points < *x* Å separated, they seem to be one point

- mobility ?
  - at room temperature atoms some atoms move by some Å

- what does 0.001 Å mean ?
- where do the numbers come from ? Later
- NMR ? ..

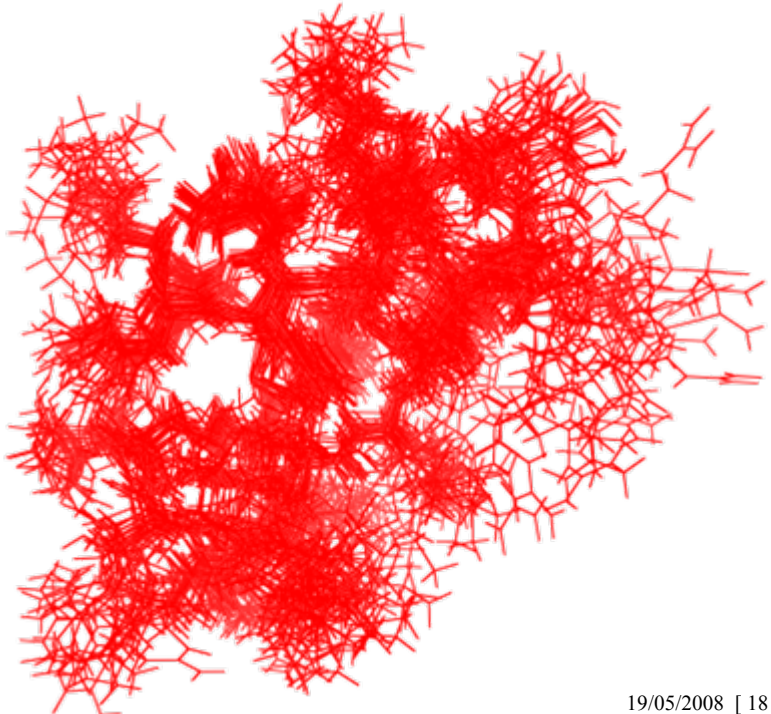
How accurate are coordinates: NMR

NMR

- precision / accuracy not well defined

Most common method

- generate 100's structures based on NMR data
- keep best 20 or 50
- submit these to protein data bank



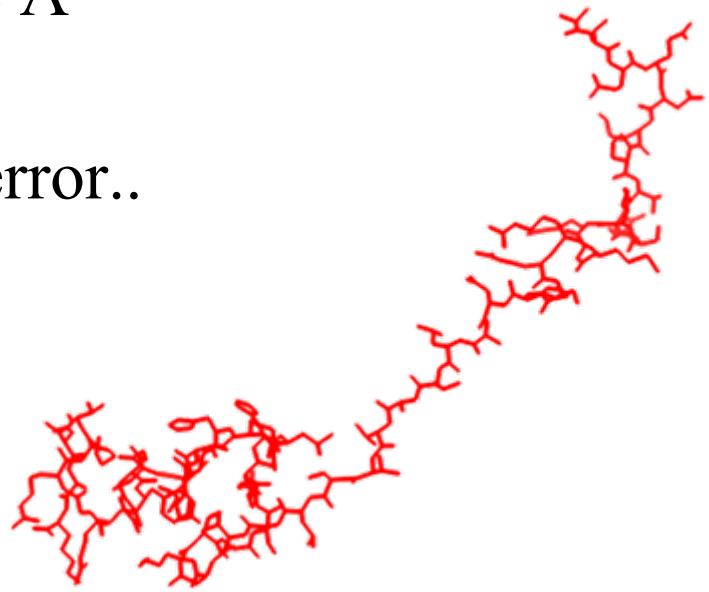
How accurate are they ?

- clearly the individual Å mean little

Where do the digits come from ?

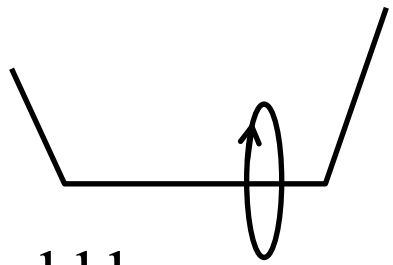
We cannot resolve atoms to Å; where do the digits come from ?

- If every atom had an error of 1 or more Å
  - structures would not look so regular
  - but the measurements do have this error..



How to reconcile this

- start with an idea of covalent geometry
  - CC bonds 1.1 Å, CN bonds 1.3 Å..
  - standard values for angles
- treat protein as a chain with these restrictions
  - think of all the rotatable angles
- fit every atom to where the experiment says it should be
  - result ?



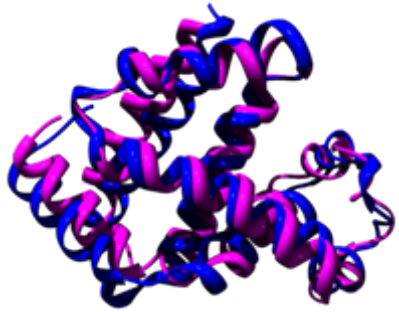
Resulting coordinates

- coordinates look perfect – not as accurate as they seem
- coordinates in PDB are mixture of
  - experiment (where NMR / crystallography puts them)
  - what we expect – all bond lengths, angles

- Given some coordinates, how can we compare them ?

Comparing coordinates

- These are very similar



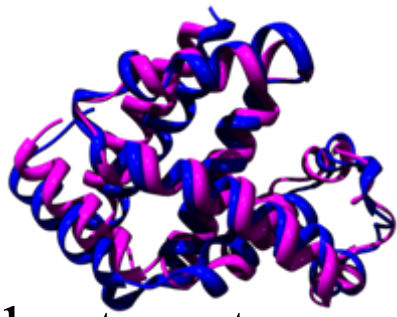
- These are clearly related, less similar

- We want to put numbers on this property

First some notation

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

Comparing two proteins



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

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

root mean square

- 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

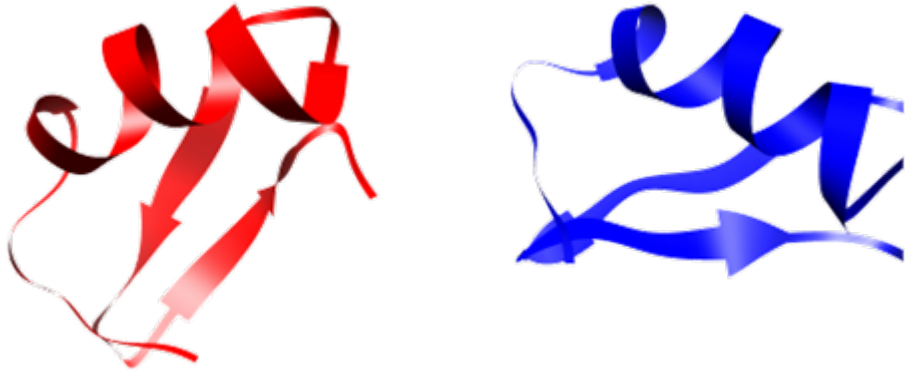
$$r_{rmsd} = \left(\frac{1}{N_{res}} \sum_{i=1}^{N_{res}} \left| \vec{r}_i^{\,a} - \vec{r}_i^{\,b} \right|^2 \right)^{1/2}$$

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

- complications

Before calculating rmsd

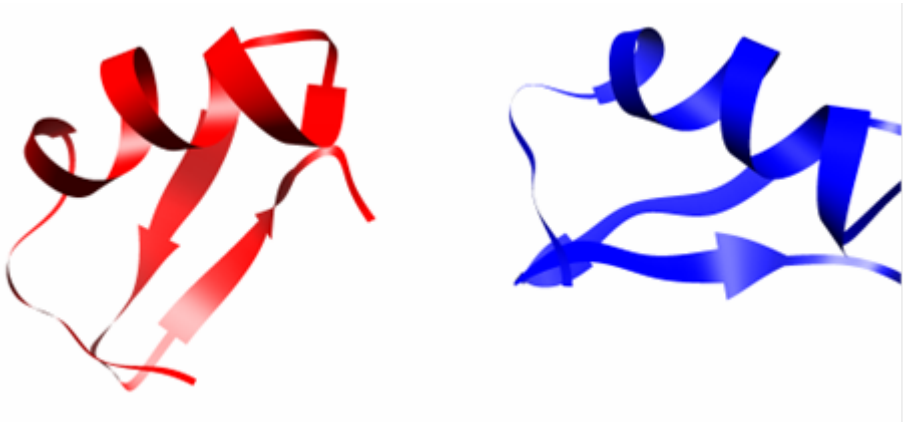
- two very similar proteins
  - coordinates are in different orientations
  - not on top of each other

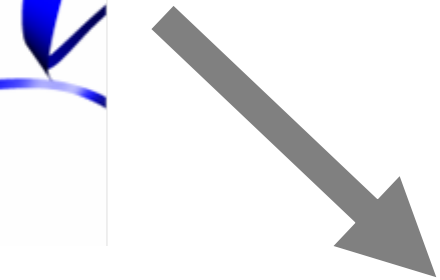


- what are the orientations of files in PDB ?
  - totally arbitrary

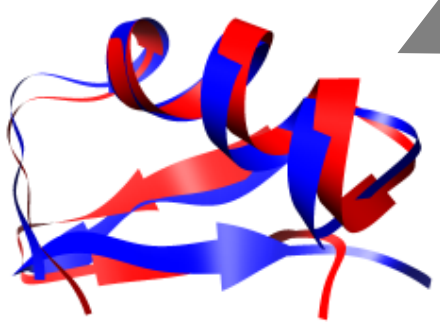
- first some other steps

Superposition of coordinates





rotation and translation



now use formula for *rmsd*

first problems with *rmsd*

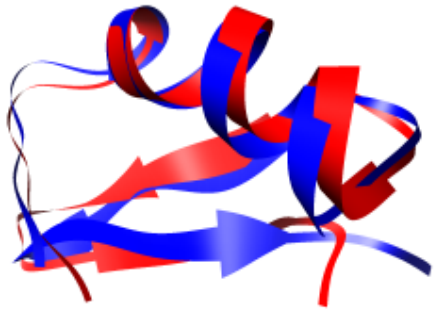
- Before calculating *rmsd*
  - coordinates must be “superimposed” (translation + rotation)

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

meaning of *rmsd*

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

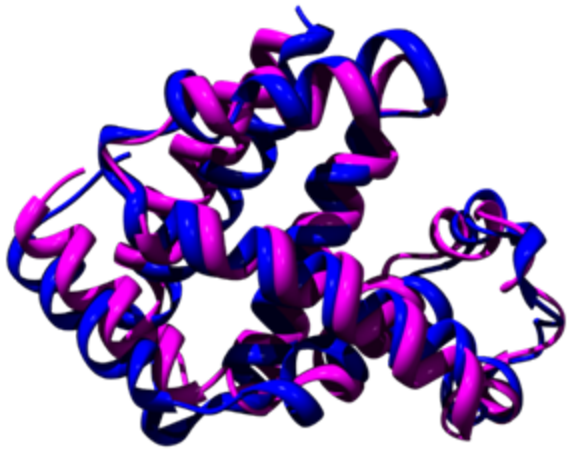
Difficulty with *rmsd*



- these two proteins have the same number of residues

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

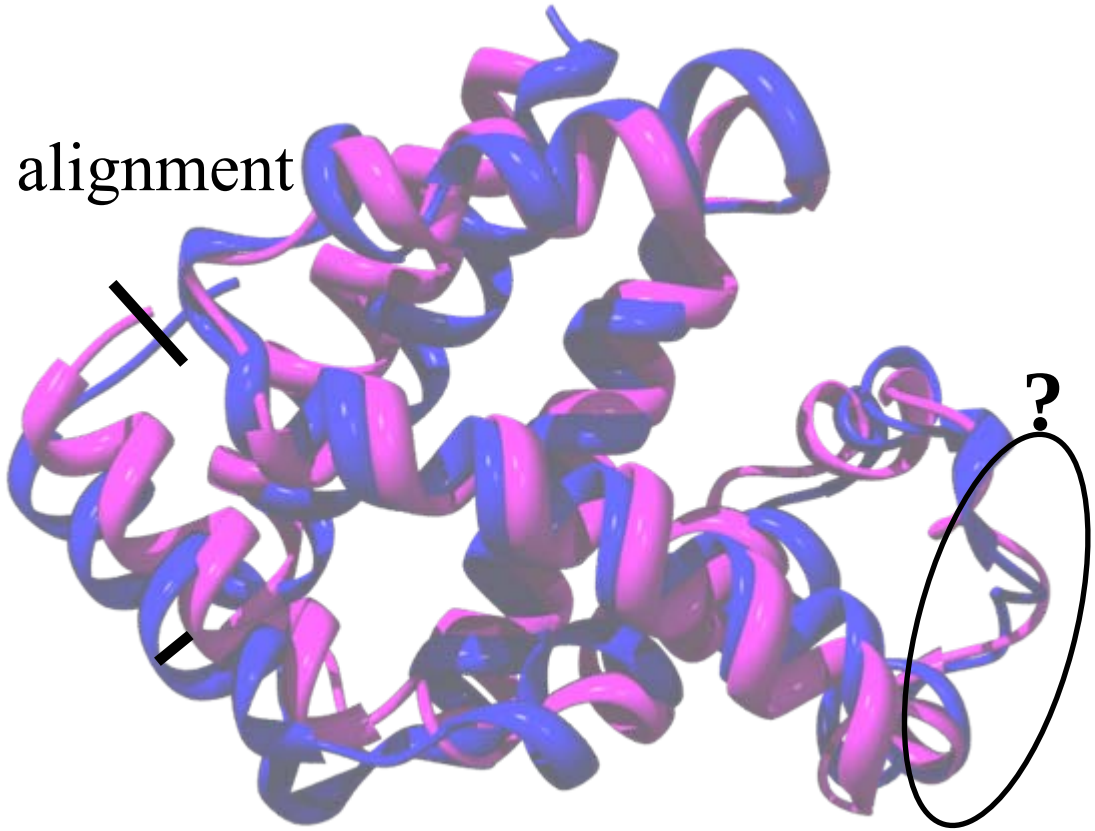
- if $i = 1, 2, 3, ..$ we use residue 1, 2, 3 in both proteins



- these two proteins have slightly different numbers of residues
- we cannot compare residue 1 to 1, 2 to 2..

Proteins of different sizes – first version

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



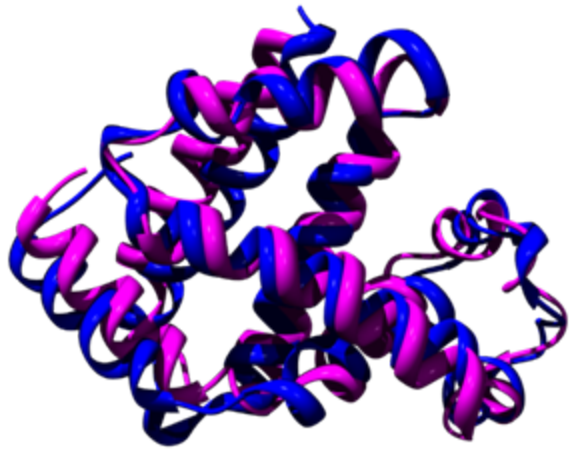
Selecting residues for alignment

- take the sequence of each protein, calculate alignment

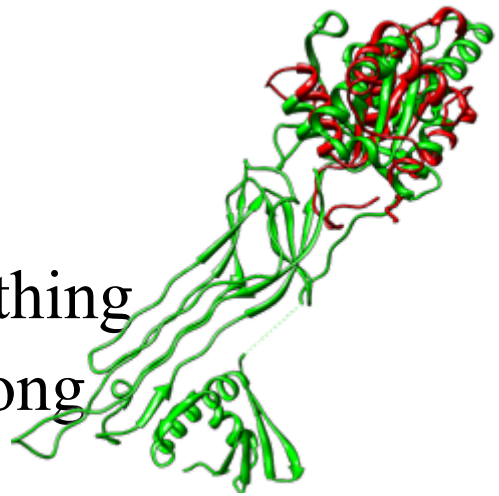
```
ACDEFG-IK-MNP..
A-DEGGHIKLMNP..
```

use these residues

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



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



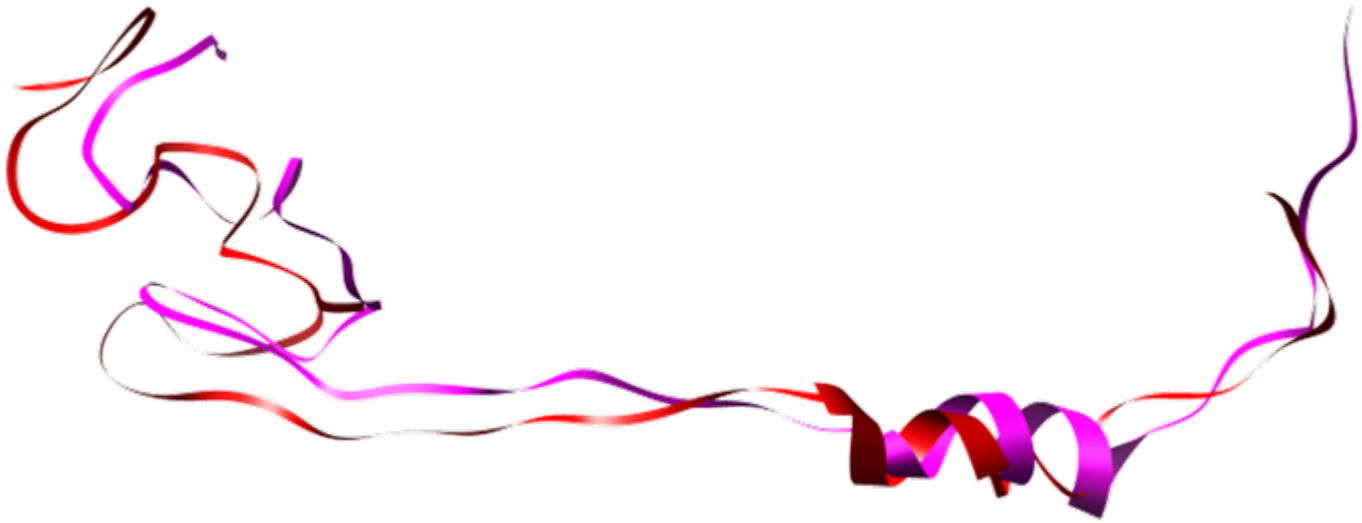
Selecting residues for alignment - better

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

- Answer : yes but..
  - no guaranteed correct solution
  - many different methods
- naïve approach – how difficult ?

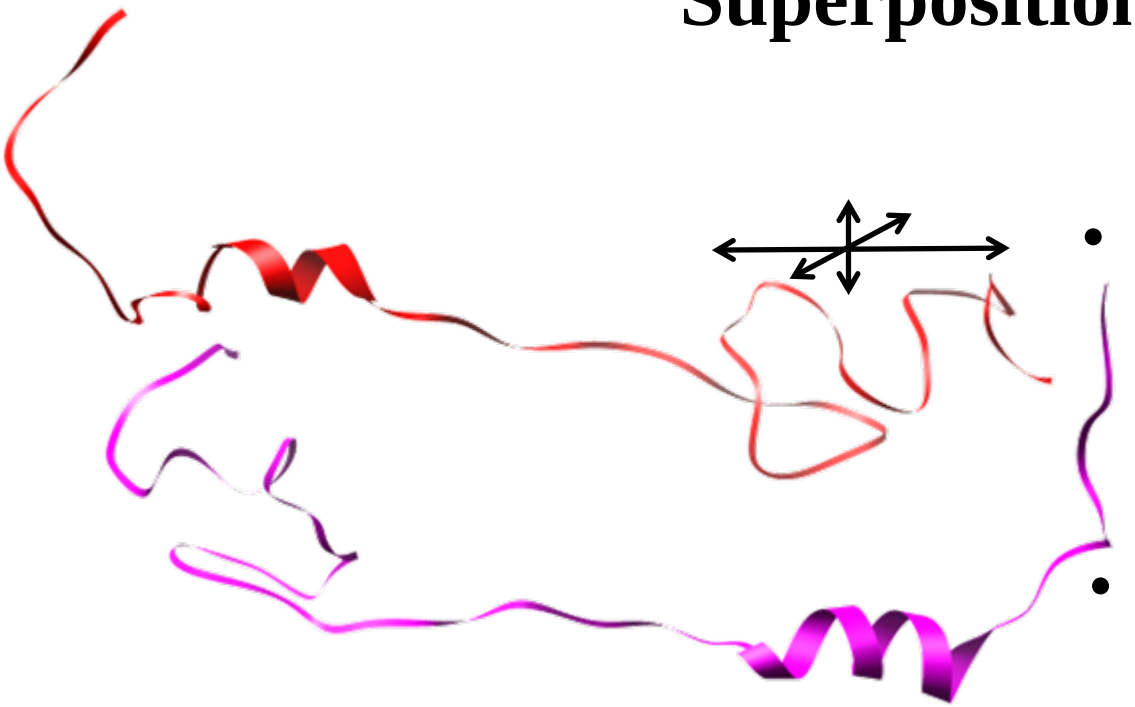
Naïve structural alignments

- how hard is it to superimpose two structures (different sizes)

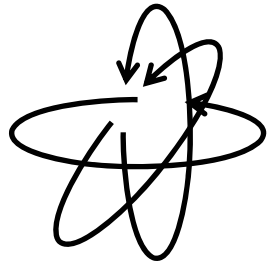


- looks easy when answer is known

Superposition



- 3 types of translation
- at each step in *x, y, z*
  - 3 rotations to fix

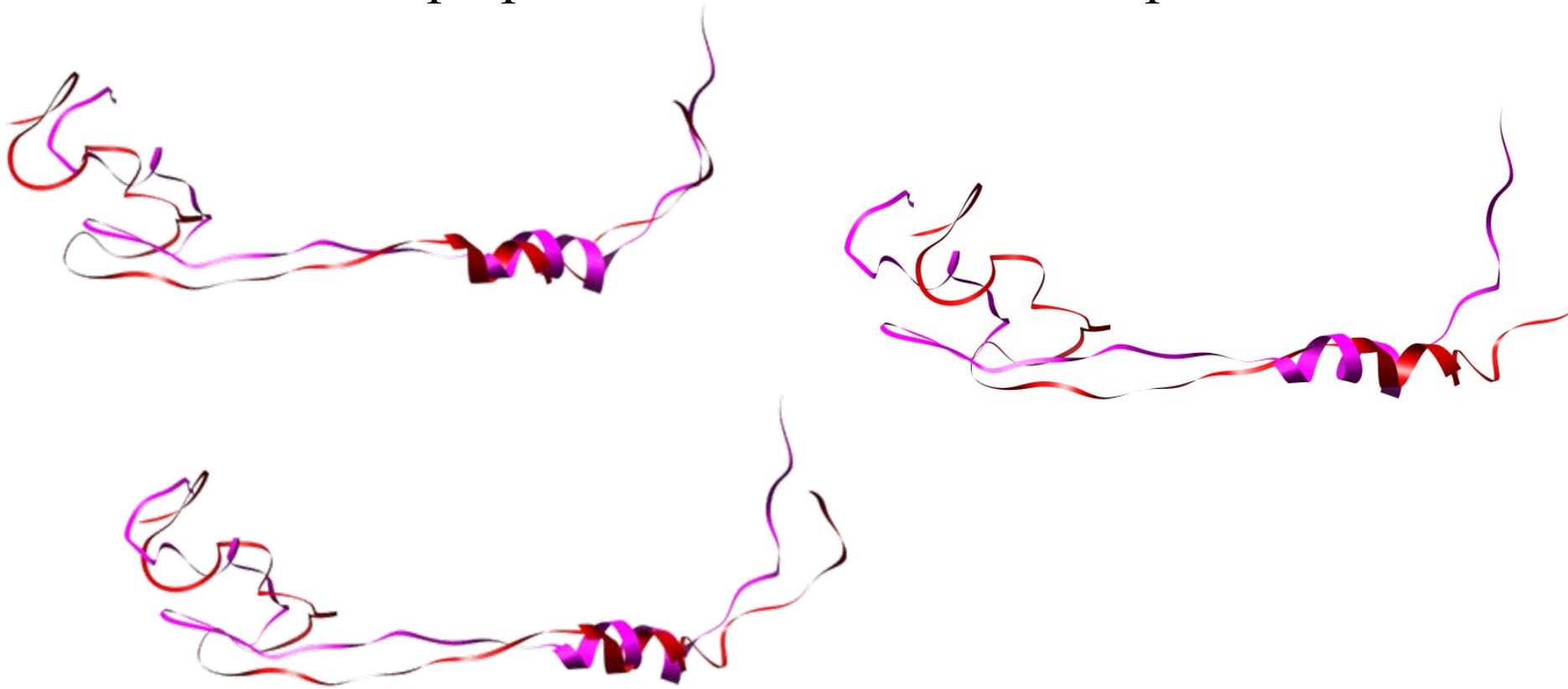


worse problem ..

- as you move one structure match gets better and worse

Local minima

- each of these superpositions matches different parts of the helix



- difficult to calculate quality of match
  - each superposition implies a different alignment

Summary of comparing two structures

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

- Next topic..
  - what can we do now that we can quantify similarity ?
  - classifications