Comparative / Homology Modelling

You have a sequence

`..AADEFGHIKHFEDA..` but no structure

- no crystals, cannot phase, too big for NMR, in a hurry

You have your sequence and want to

- find residues that are far from active site and in a loop
- guess which residues in your sequence are involved in chemistry
- … say what certain residues do
  - are they in active site ? Surface ? Buried ?

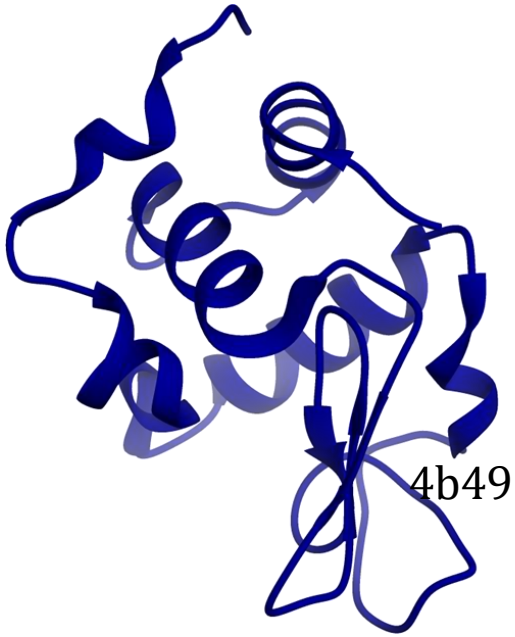
To do ?

Modelling

...AADEFGHIKH-GED... your sequence

- do a blast search
- find a related sequence in PDB - has a structure

...AQDEF-HIKKGFED found in PDB



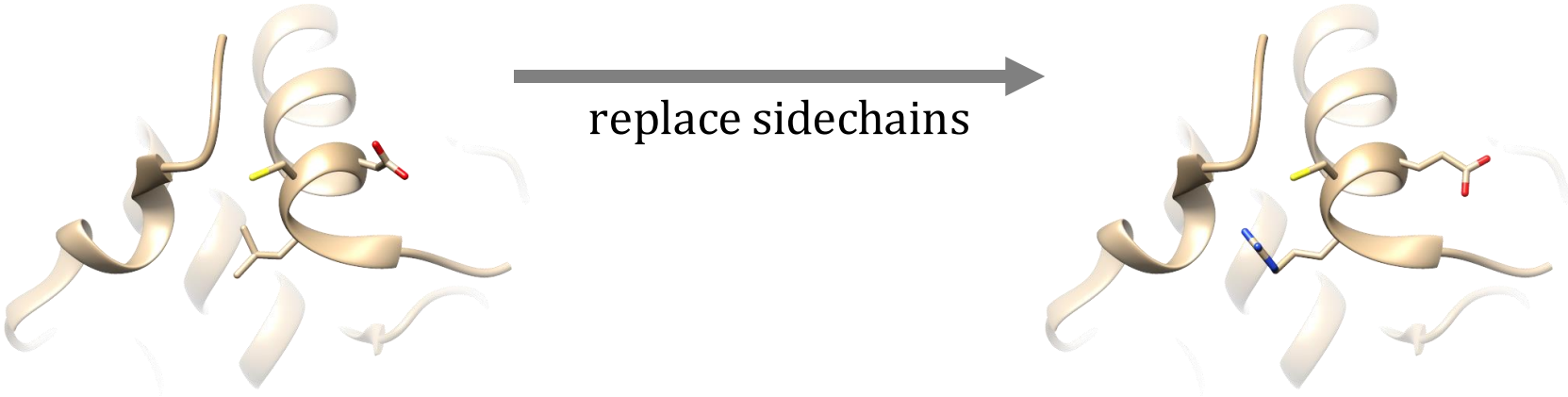
- put your sequence on to this structure
  - literally …

Modelling

. . L C D . . original residues

just replace with residues from your sequence

. . F C E . .



original from PDB

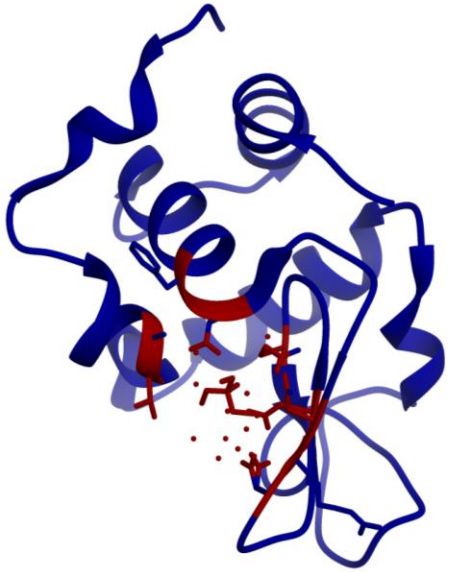
backbone with your sidechains

Using model

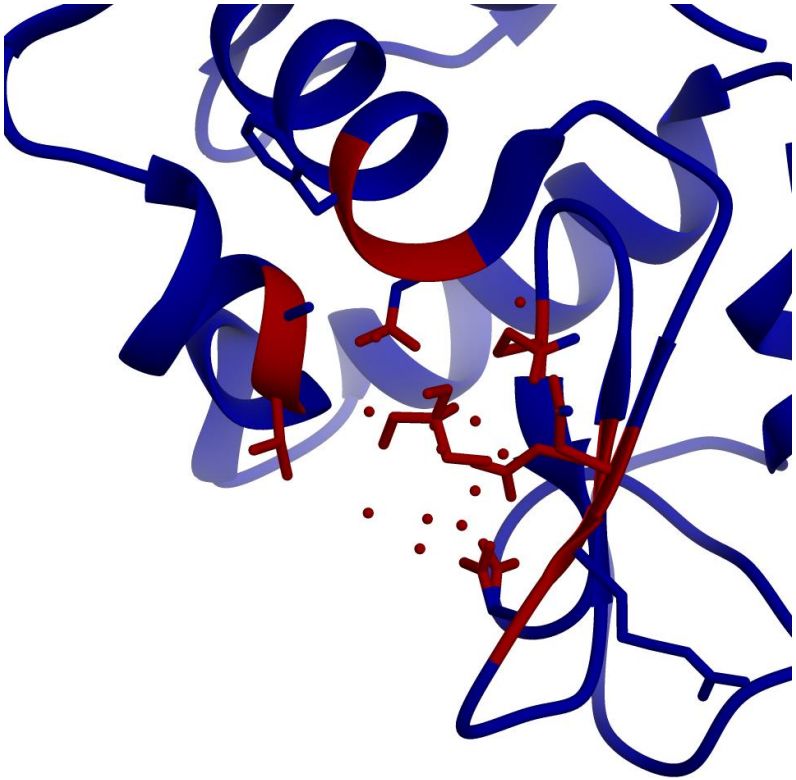
with substrate

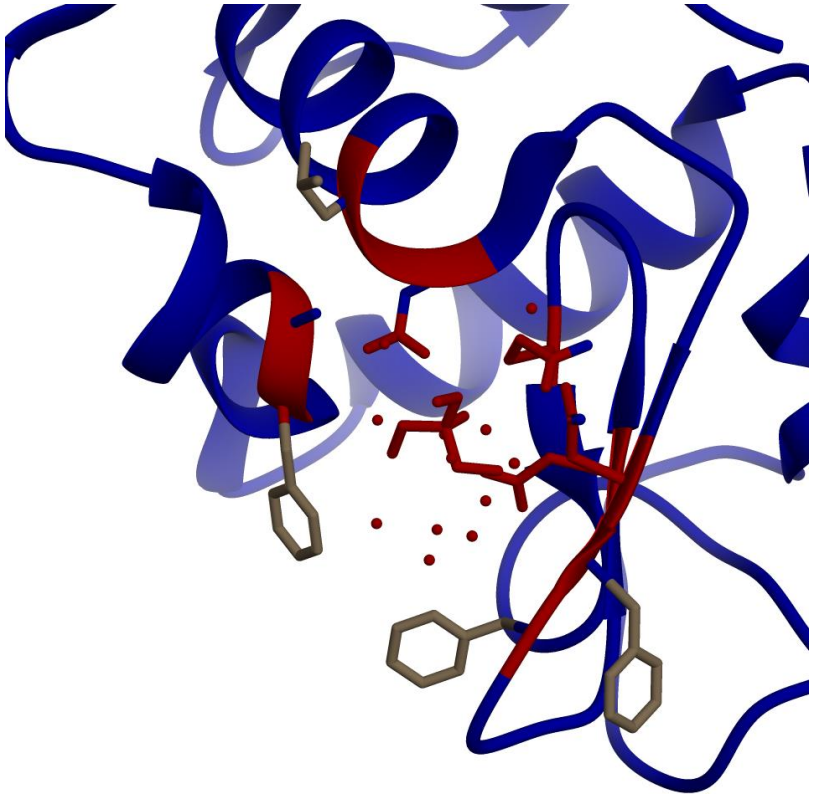
…**AADEFGHIKH-GED**…

who is near substrate ?



predictions as to active site





Accuracy

You now have coordinates for your sequence

- how accurate ?
- does it matter ?

May not need to be accurate

- phasing (X-ray crystallography)
- guiding mutagenesis

May or may not be good enough

- docking

Most basic rule

Guiding belief

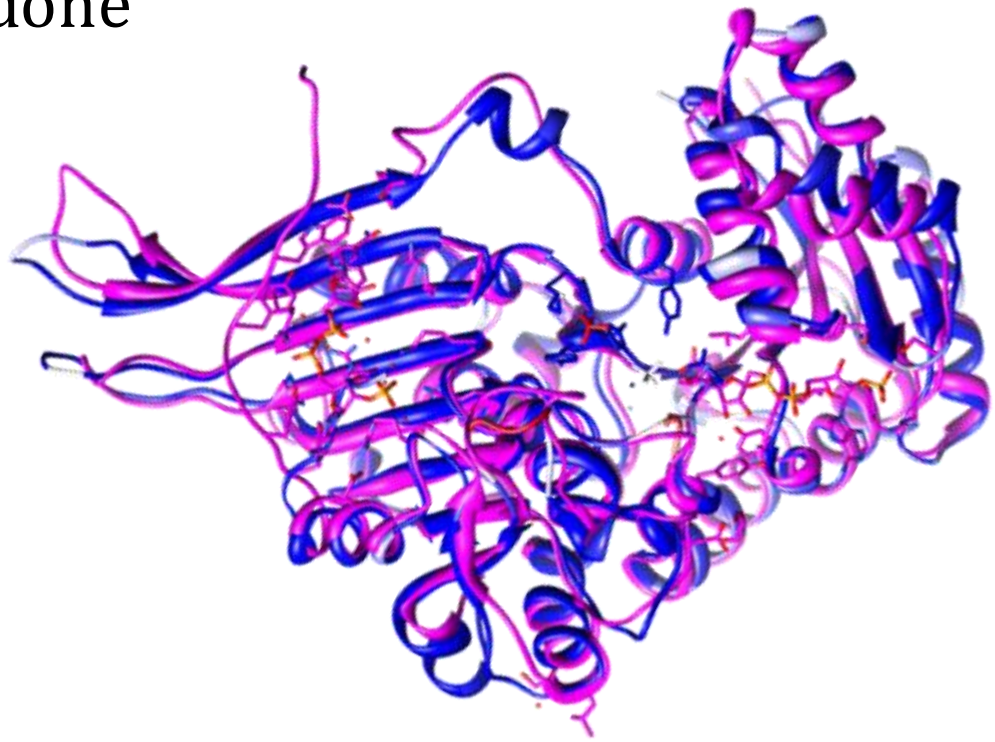
- similar sequence gives similar structure
  - evolution
  - chemistry

Most important

- closer the sequence is to template – better the model

Reasonable expectations

- two enzymes (G6Pdh) easy to find homology
- could one have been modelled, knowing the other ?
- knowing the structures below, this might be the limit of what could be done



Overall modelling protocol

1. decide on template
2. align sequence (unknown structure) to known structure / template / parent
3. replace sidechains of parent with new ones
4. fix
   - gaps
   - insertions
   - loops
5. overall structure

Finding a template / parent

How unique is my sequence ?

- human haemoglobin,
  - you would find horse, pig, and 10^3 globin structures
- enzyme from a virus
  - it may have no obvious homologues – has evolved too far

high sequence identity (> ~20-25 %)	low sequence identity (< ~20-25 %)	very low
blast, fasta, anything	psi-blast, HMMs	psi-blast, optimism

Why so vague ?

Template reliability

Old rule

- < 20 % not similar
- > 25 % similar
- otherwise (twilight zone)

Not a good rule

Template reliability

Why is this not enough ?

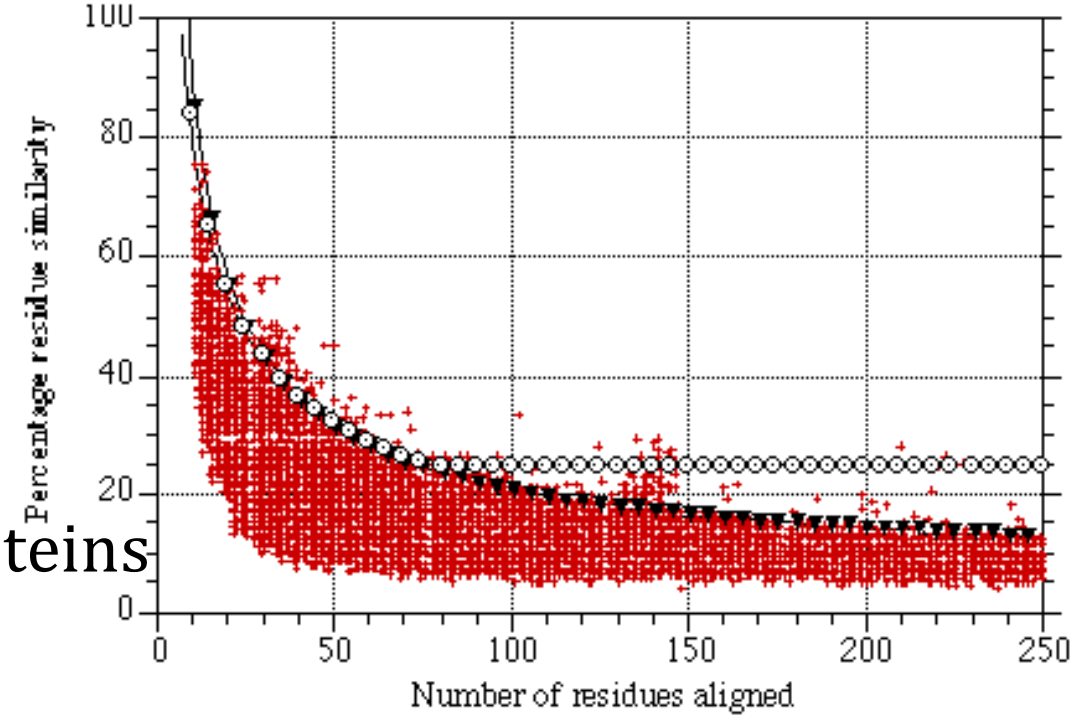
Consider random mixture of amino acids

- add bias of composition (some amino acids are rare)
- compare a lot of proteins and say
  - pairs have 15 % similarity (average)
- we see a pair of 20 % similarity for 50 residues
  - is it significant ?
- we see a pair of 20 % similarity for 600 residues
  - more convincing

Quantifying importance of similarity length

Reminder..

We know the size of an alignment



How often are the two proteins not structurally related ?

but

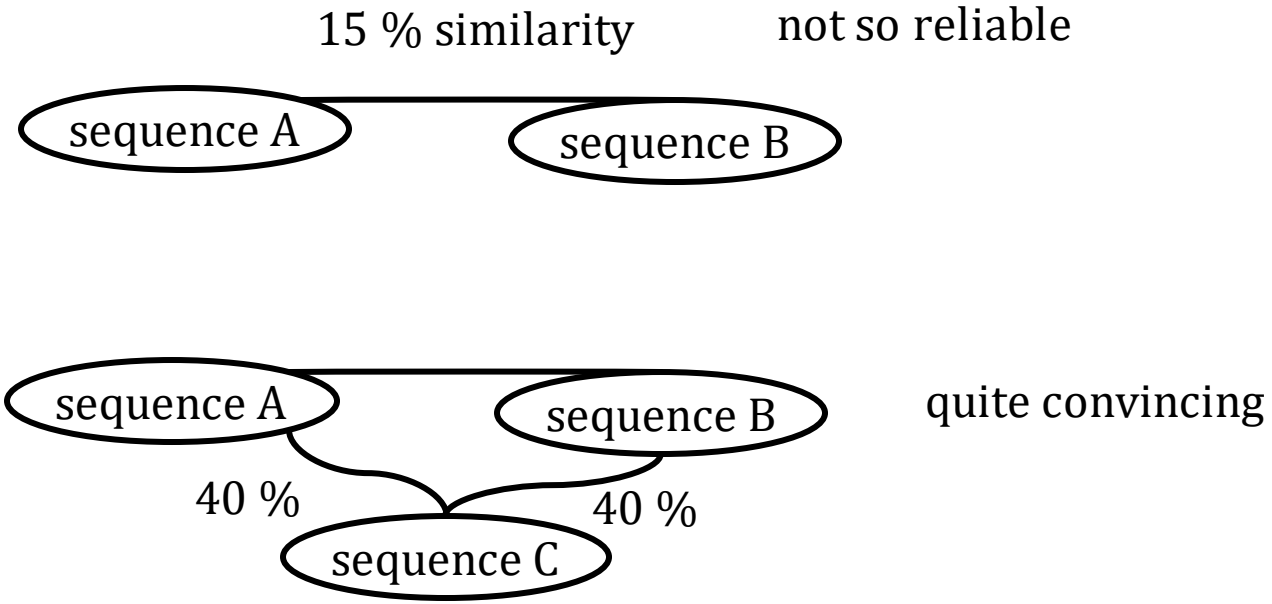
more to deciding if similarity is significant

Rost, B. Prot. Eng. 12, 85-94 (1999)

Transitive relations

How significant is the similarity between two proteins ?

- does not only depend on the two proteins



sequence C – called transitive relation

Summarise

- Sequence identity (sequence to template) is most important
- It is not enough to say 20 – 25 % similarity
  - depends on length of alignment
  - depends on common relations (transitive)

Sequence alignment

We have picked a template for our sequence now...

1. decide on template
2. **align sequence (unknown structure) to known structure / template / parent**
3. replace sidechains of parent with new ones
4. fix
   - gaps
   - insertions
   - loops
5. overall structure

- we need an alignment
- difference compared to database searches ?
  - not scanning a database (10^7 sequences)
    - we can do best possible alignment – time is not important

Careful alignments

Computer time not a problem - use

- most expensive alignment algorithm, could be one of
  - Needleman-Wunsch/Smith-Waterman family
  - multiple sequence alignment with related sequences for template and query sequence

How important ?

Alignment errors

```
ANDREW
ANQEW
```

two reasonable alignments

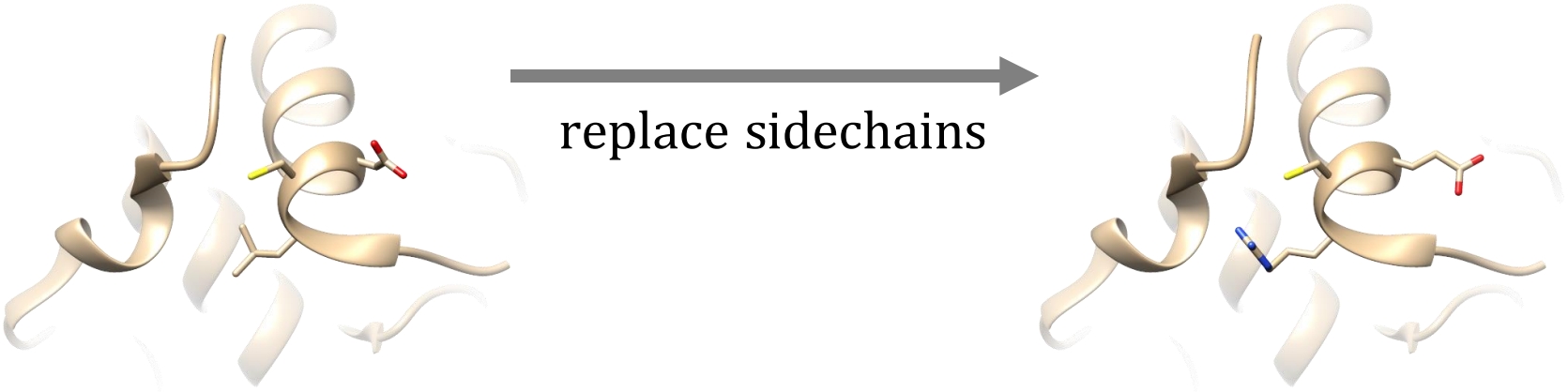
```
ANDREW          or ANDREW
ANQ-EW          or AN-QEW
```

difference ?

- from C_i^{α} to C_{i+1}^{α} almost 4 Å

Sidechains – where to put them ?

- new sidechains ? need coordinates
- should you worry ?
  - No - surface residues – maybe not – they rotate
  - Yes - residues with contacts / interactions



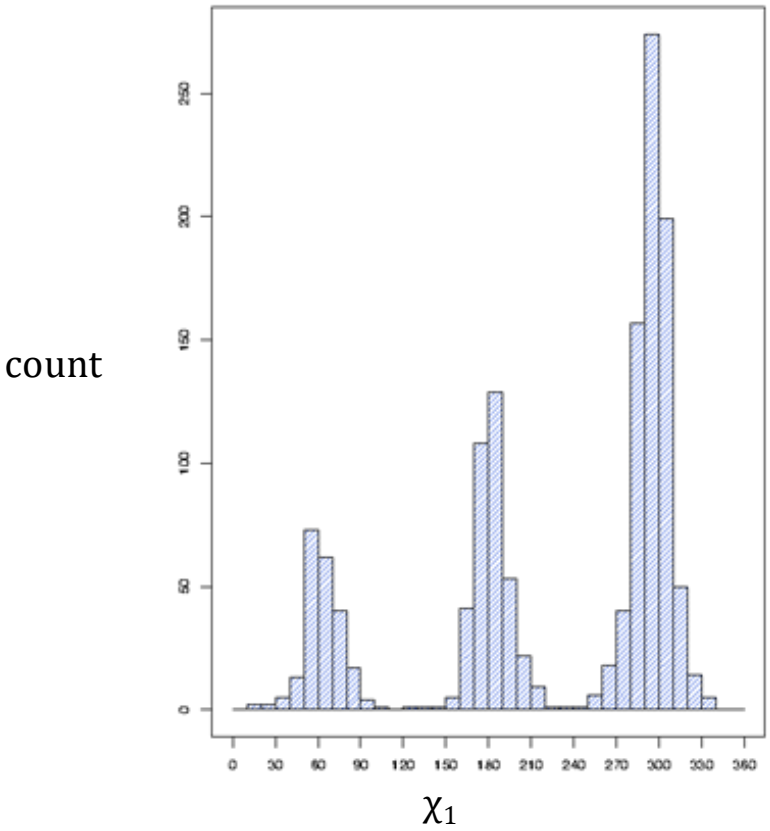
original from PDB

with your sidechains

Rotamers for sidechains

Approximation / simplification

- sidechain coordinates are taken from likely rotamers



Dunbrack Jr, R.L., Curr. Opin. Struct. Biol. 12, 431-440 (2002)

Example – replace ala with trp

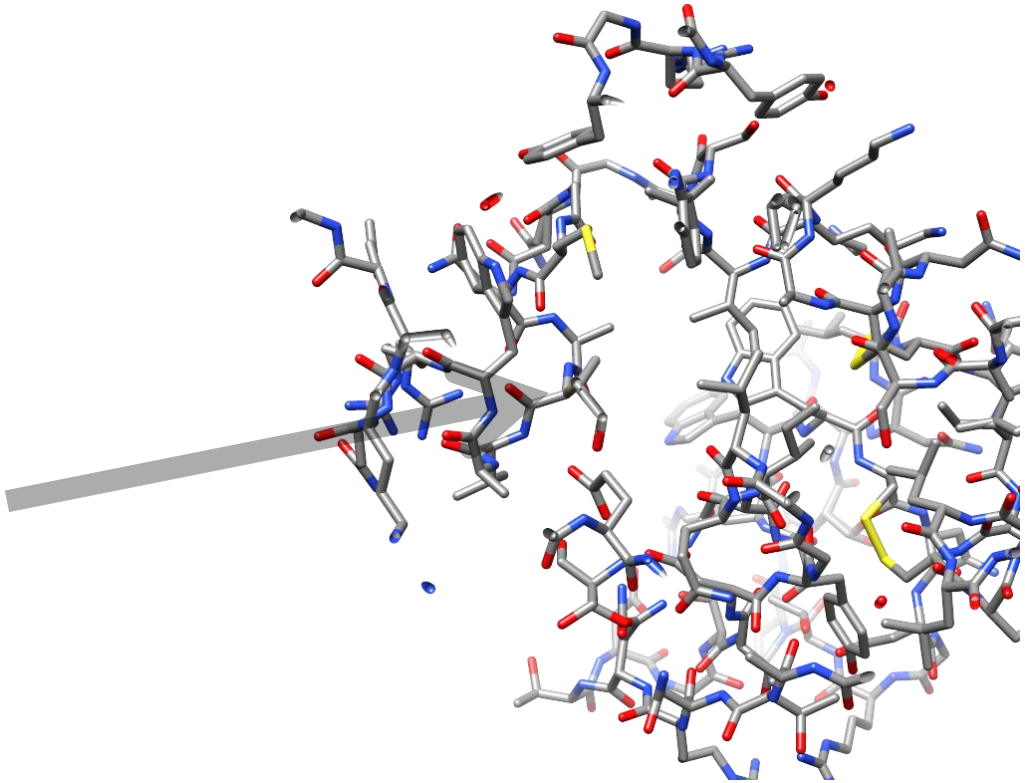
Rotamers

- concede that you are happy with discretization

Trp rotamers

- 3 rotamers at χ_1
- 3 rotamers at χ_2

What do they look like ?

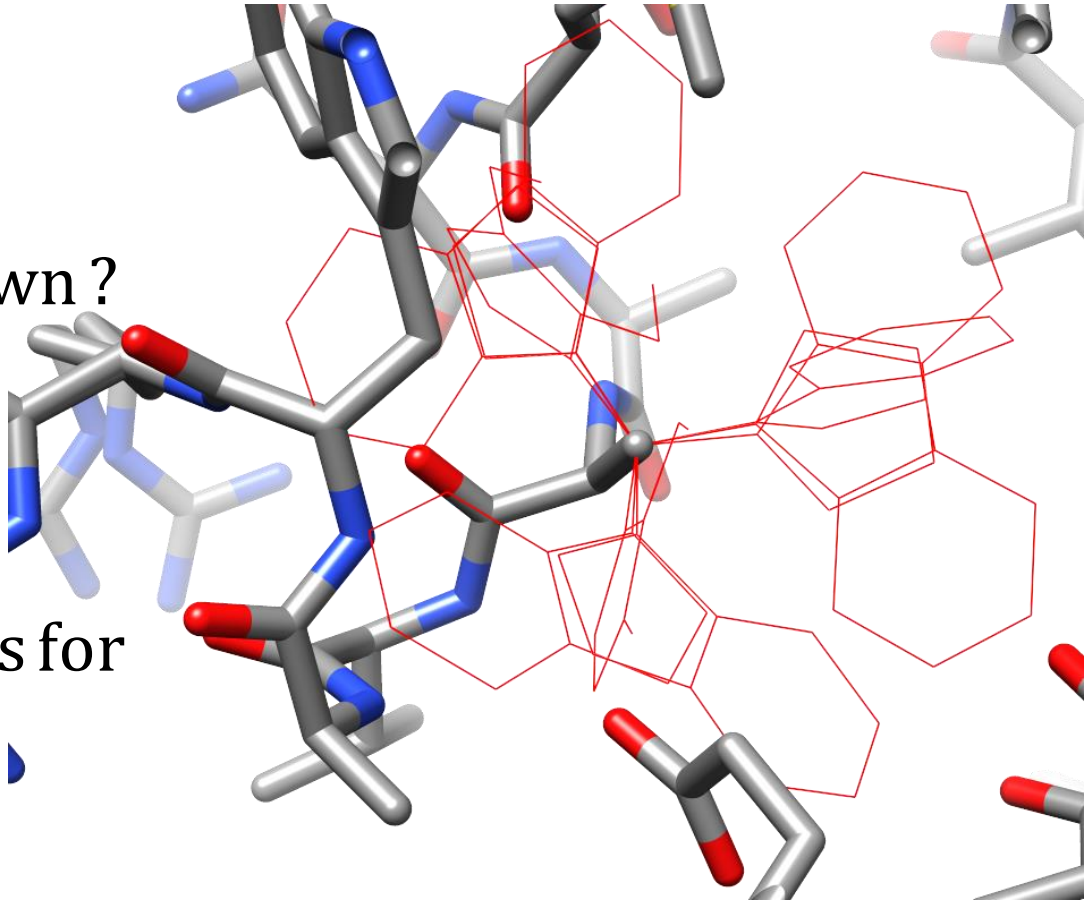


9 possibilities

- many are silly
- have to be checked

- how difficult ?
- are the neighbours known ?

- if we have 9 possibilities for a neighbour
  - already 9 × 9



Sidechain placement

If sidechain in your sequence is the same as template

- use template coordinates

New sidechains

- say m_i possibilities at each site i
- make lists of possibilities at each site
- try to find biggest network of rotamers which compatible with each other
- use simple scoring scheme (clashes)

How bad is the calculation ?

rotamer search

- at each site i we have m_i possibilities
- could say $\Pi_i\, m_i$ possibilities ($m_1 \cdot m_2 \cdot$) or just m^n

- most sidechains have only a half a dozen neighbours
- usually minutes of cpu time (not days)

Are you finished ?

- maybe
- can do a energy calculation to make coordinates nicer

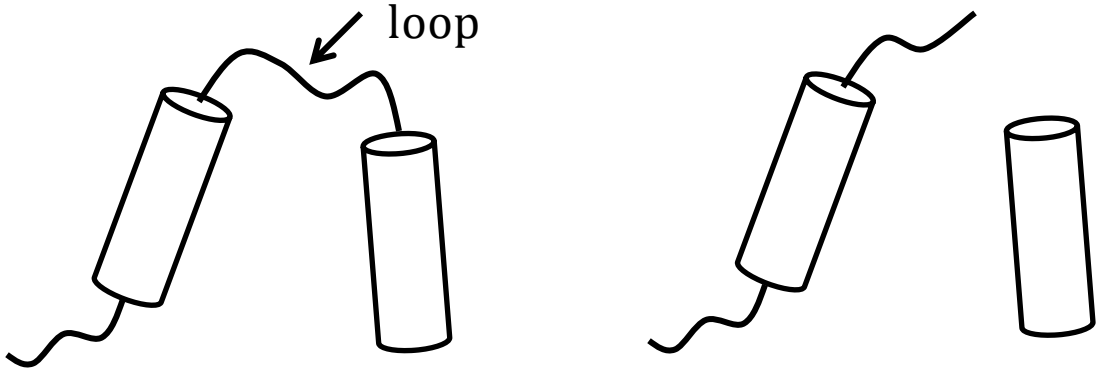
Broken main chain

Typical situation

`ANDR-WQANDRKWSANDRWWC` parent

`ANDREW---DRKWS--DRWWC` model

our model…



Basic problem…

- pieces of unknown structure
- endpoints relatively fixed
- should be joined

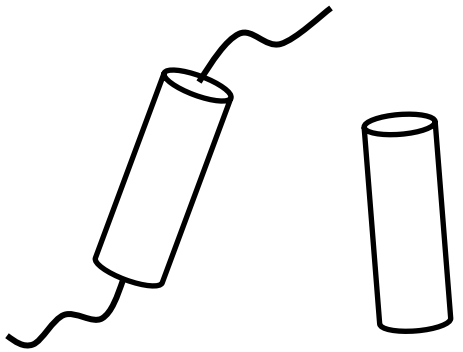
Loop modelling

Loop problem

- do not want to disturb regular secondary structure
  - more likely to be correct
- ends of loop relatively well known
- composition (sequence) of loop

The problem specifically:

- find an arrangement of backbone and sidechains which
  - is geometrically possible
  - low energy



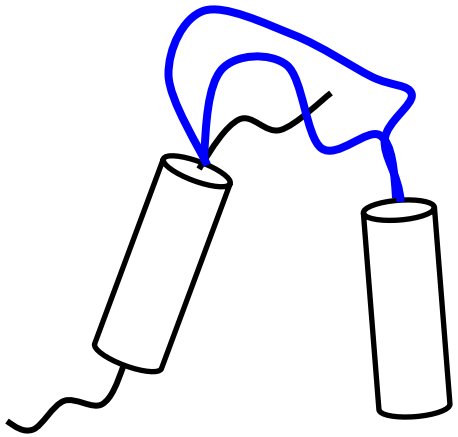
Possibilities

- distance geometry
- database search
- brute force

Methods for loops

Distance geometry

- we know
  - end points and distances
  - sequence of loop
    - all bond lengths and angles



- use distance geometry to generate plausible arrangements

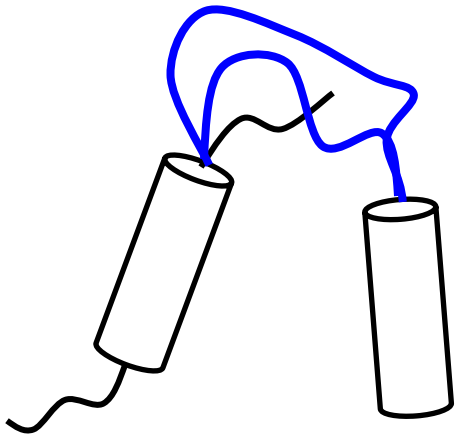
Results ?

- arrangement of atoms with
  - correct covalent geometry
  - no atoms on top of each other (set by minimum distances)
- little consideration of torsion angles

Loops Database searching

Database searching

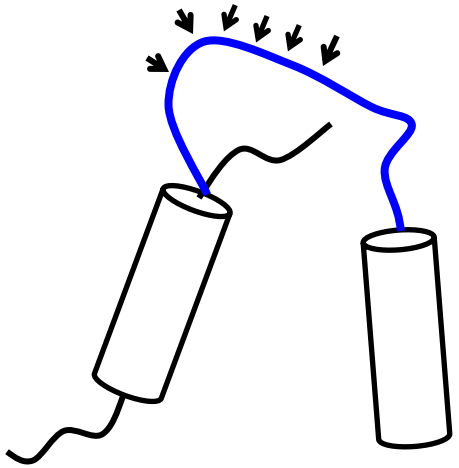
- imagine we have a 9 residue loop
- take protein data bank
- collect coordinates of all 9-residue loops
- insert those with correct end to end distance
- refinement…
  - insert those with almost correct distance &
  - similar sequence to loop residues



Loops – brute force

Desperation / brute force for small number of residues

- divide angles into pieces (maybe 30°), 360/30 = 12
- test every combination (joining ends, energy)
- called "grid search"



- How many angles ?
- per residue
  - fix ω
  - phi φ, psi ψ 12×12=144
- possibilities = $144^{N_{res}}$

Quality

- energies
- geometries
- statistics of backbones / sidechains

Remember energy/geometry/statistics are related

Real world

Recipe on these slides rather simple

- usually many models generated and checked
- multiple templates
- multiple templates simultaneously ?
- interaction with experiment (predictions tested)

- automatic methods are very good

What does one achieve ?

Folklore – history - testing

Very easy cases ?

- not much change from parent

Very difficult ?

- lots of errors

An Example

2mnr and 4enl

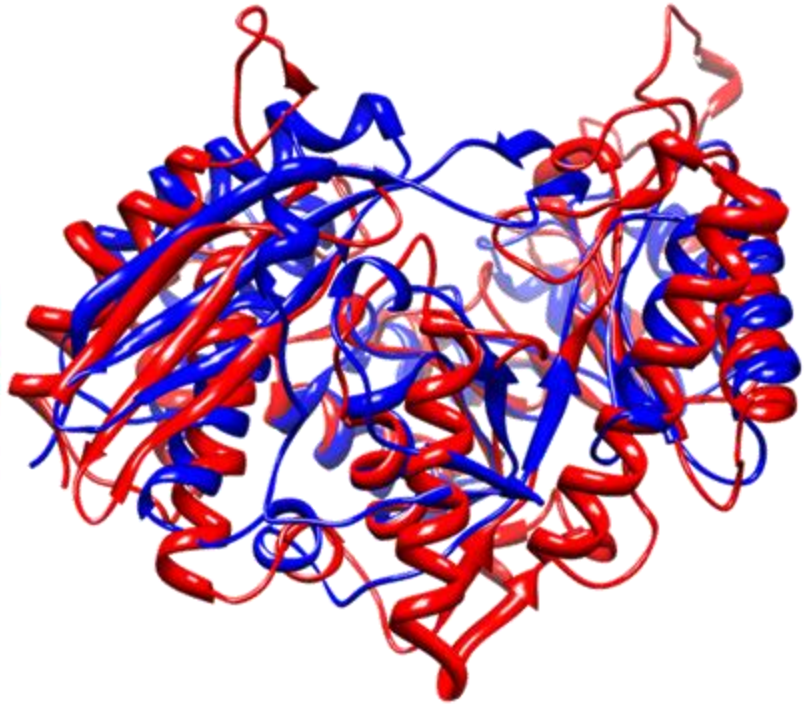
- would be a typical modelling target
- in real world
  - alignment would not be perfect
  - loops may be quite wrong

The sequence alignment

```
Seq ID 25.1 % (81 / 323) in 373 total including gaps
    :    1    :    2    :    3    :         4    :    5
    :    0    :    0    :    0    :         0    :    0
sktyavlglgnggha faaylalkgqsv--lawdidaqr------ikeiqdrgaiiaegpg
svehimrdv-nggwa-mryihangaslfflavyihifrglyygsykapreitwivgmviy
  0    :    0    :    1    :    1    :    1    :    1    :
  8    :    9    :    0    :    1    :    2    :    3    :
  0    :    0    :    0    :    0    :    0    :    0    :

    :    0    :    0    :    0    :    0    :    1         :
    :    6    :    7    :    8    :    9    :    0         :
    :    0    :    0    :    0    :    0    :    0         :
la--gtahpdlltsdiglavkdadvililvvpaihhasiaaniasyisegqli---ilnpg
llmmgtafmgyvlpwgqmsfwgatvitglfgaipg--igpsiqawllggpavdnatlnrff
    1    :    1    :    1    :    1         :    1    :    1
    4    :    5    :    6    :    7         :    8    :    9
    0    :    0    :    0    :    0         :    0    :    0

  1    :    1    :    1    :    1    :    1    :    1    :
  1    :    2    :    3    :    4    :    5    :    6    :
  0    :    0    :    0    :    0    :    0    :    0    :
atggalefrkilrengapevtigetssmlftcrserpgqvtvnaikgamdfaclpaakag
fslhyllpf-viaalvaihiwafhttgnnnptgvevrrtskadaekdtlpfwpyfvikdl
  :    2    :    2    :    2    :    2    :    2    :    2
  :    0    :    1    :    2    :    3    :    4    :    5
  :    0    :    0    :    0    :    0    :    0    :    0

  1    :    1    :    1    :    2    :    2    :         2
  7    :    8    :    9    :    0    :    1    :         2
  0    :    0    :    0    :    0    :    0    :         0
waleqigsvlpqyvavenvlhtsltnv-navm-hplptllnaarcesgtpf----qyyl-
fala-l--vllgffavvaympnylghpdnyvqanplstpahivpewyflpfyailrafaa
  :         2    :    2    :    2    :    2    :    3    :
  :         6    :    7    :    8    :    9    :    0    :
  :         0    :    0    :    0    :    0    :    0    :

      :    2    :    2    :    2    :    2    :    2    :
      :    3    :    4    :    5    :    6    :    7    :
      :    0    :    0    :    0    :    0    :    0    :
-egitpsv-gslaekvdaeriaiakafdlnvpsvcewypatiyeavqgnpayrgiagpin
dvwvvilvdgltfgivdakffgviamfga-i-avmalapw-ldtskvrsgayr----pkf
3         :    3    :    3    :    3    :    3    :         3
1         :    2    :    3    :    4    :    5    :         6
0         :    0    :    0    :    0    :    0    :         0

 2    :    2    :    3    :    3    :    3    :    3    :
 8    :    9    :    0    :    1    :    2    :    3    :
 0    :    0    :    0    :    0    :    0    :    0    :
lntryffedvstglvplselgravnvptplidavldlisslidtdfrkegrtleklglsg
---rmwfwflvldfvvltwvg-a--m--pt-eypydwis-liastywfay-flvilpllg
      :    3    :    3         :    3    :    4    :    4
      :    7    :    8         :    9    :    0    :    1
      :    0    :    0         :    0    :    0    :    0

 3    :
 4    :
 0    :
ltaag--irsave
atekpepipasie
   :    4
   :    2
   :    0
```

2mnr and 4enl example

- sequence alignment not the same as alignment from structures



Summarise für Klausur

Ideas of sequence similarity

Technical issues

- loops
- sidechain placement

Why not to build a model

Why do people like models ?

- Here is a picture of my protein

- Is it necessary ? Not always

`aacsdefgh...`	known structure
`aactde-gh...`	some related sequences
`aqctdewg...`	your sequence
`gacsdeggh...`	more related sequences

...

your question

- is your sequence the same kind of enzyme ?
- has the active site changed ?

if ser 4 is part of active site in known structure

- you can say thr 4 in your sequence is the corresponding residue
- coordinates are not necessary – information is in sequence