Comparative / Homology Modelling

Topics
- rotamer optimisation
- loop prediction
- reliability of sequence similarity

Summary
- one protein sequence (protein 1)
- some related protein with structure (protein 2)
- put sequence 1 onto structure 2

Andrew Torda, Wintersemester 2017/ 2018, GST

The mission

`..AADEFGHIKHFEDA..` your sequence

No structure
- will not crystallise, too big for NMR, in a hurry, no money

You want to
- replace a residue for binding to a surface
- guess which residues in your sequence are involved in chemistry
- ...

...`AADEFGHIKH-GED`...   your sequence

do a blast search ... find

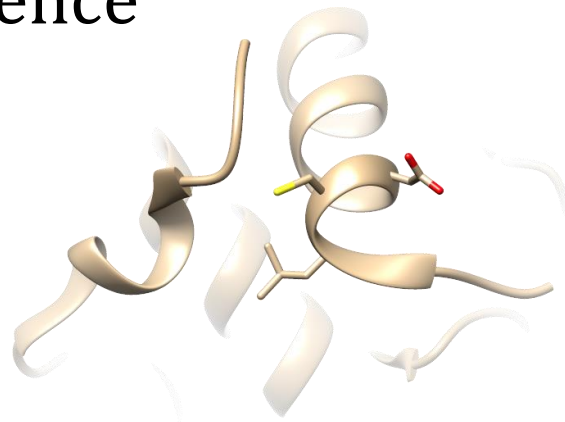...`AQDEF-HIKKGFED`   structure 4b49 in PDB

replace original
`..AQDEF_H..`
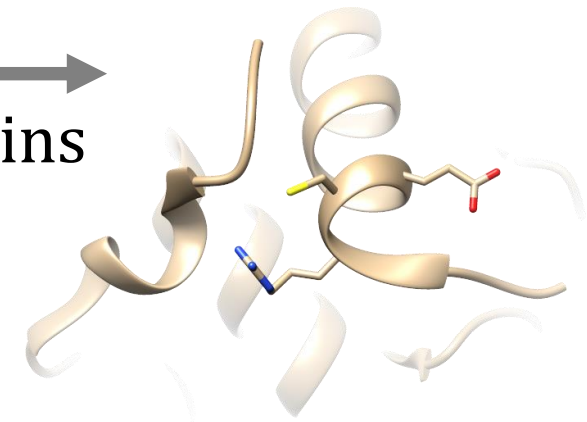
with your sequence
`..AADEFGH..`

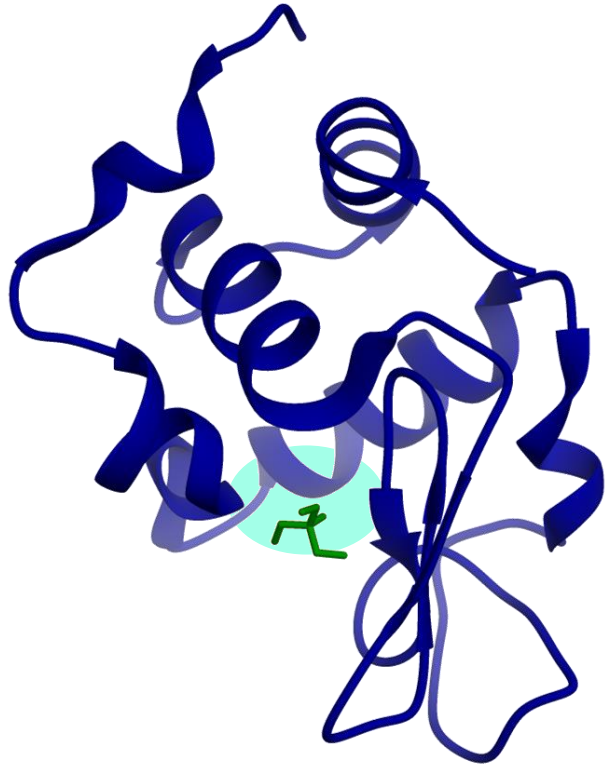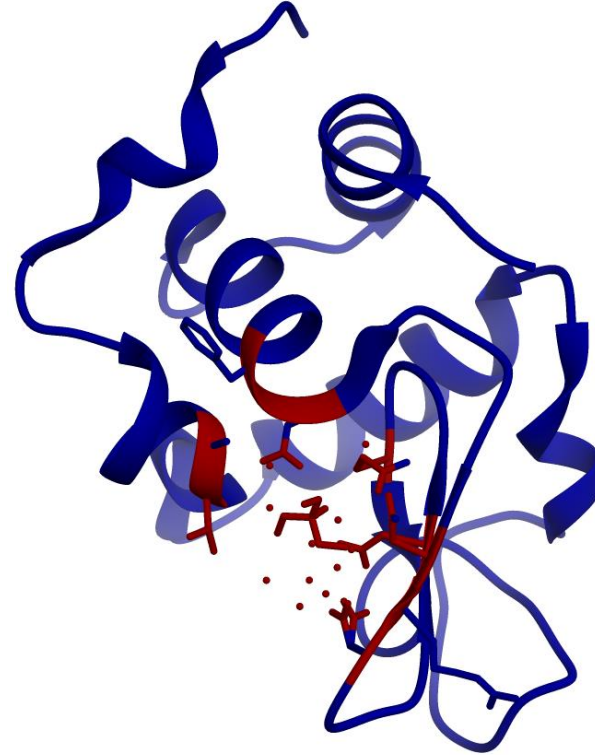original from PDB   replace sidechains   backbone with your sidechains
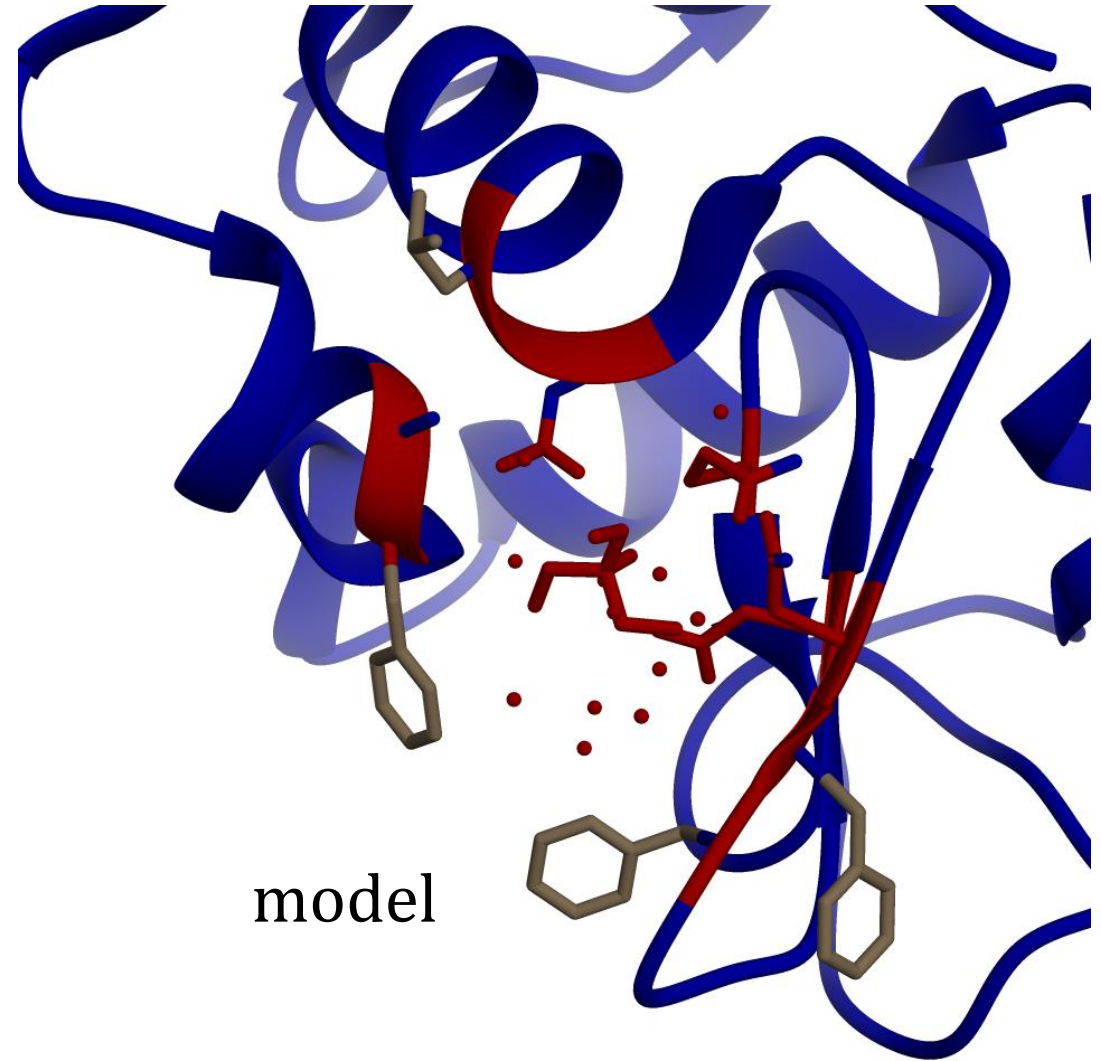
Using model

with substrate

...**AADEF****G****HIK****H****-GED**...
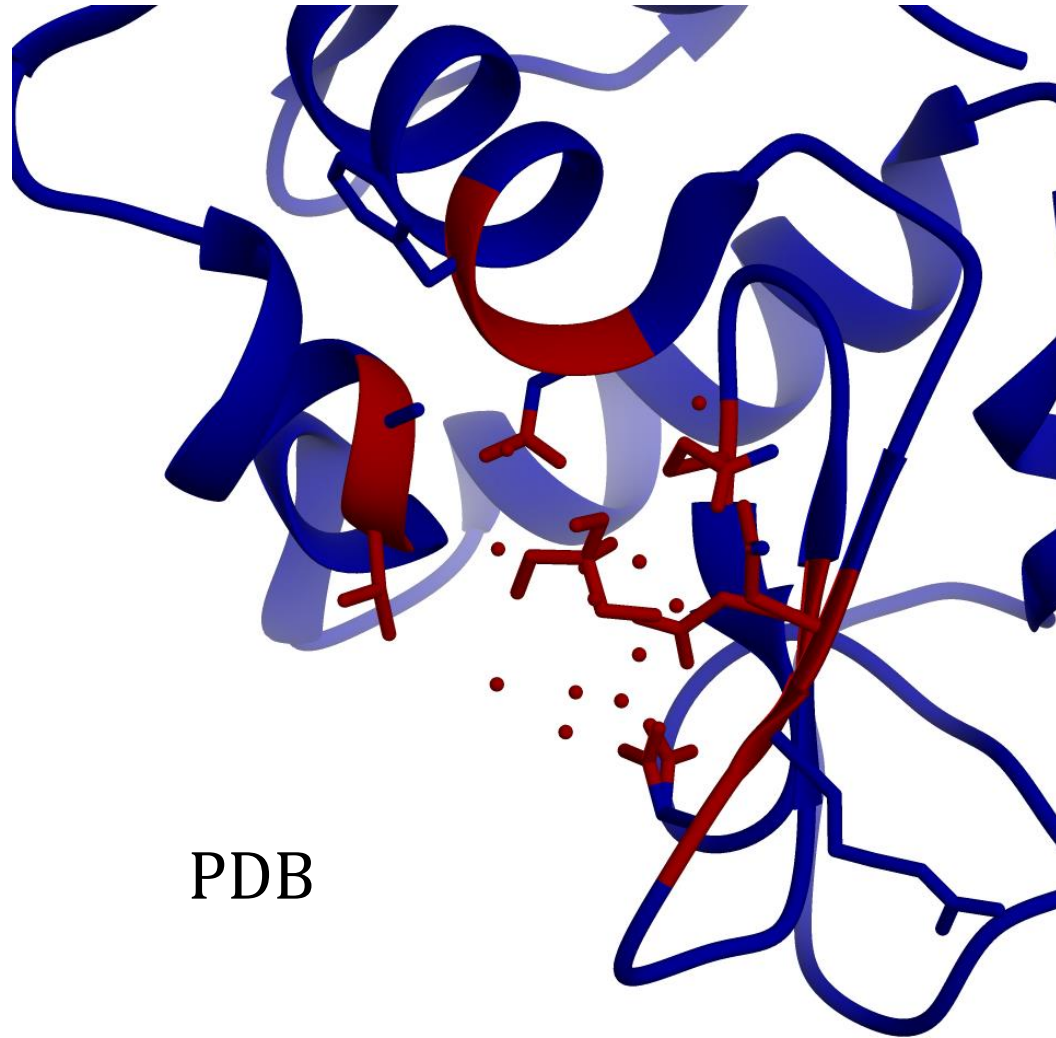
who is near substrate ?

predictions as to active site

PDB

model

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 (sequence terms) – 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
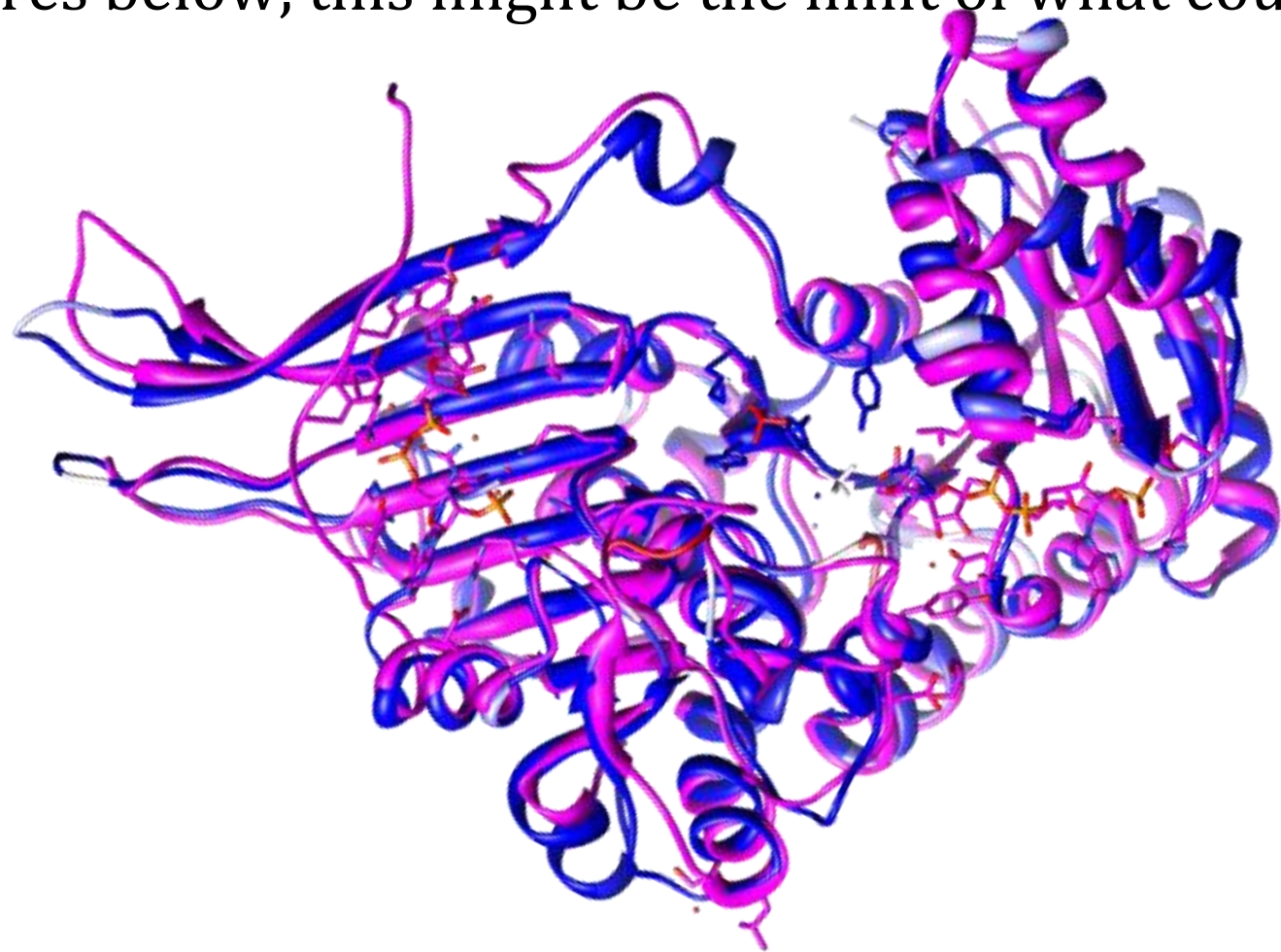
Sequence and structure similarity

Two proteins with similar sequence
- how likely is similar structure ? very
  - question of degree (how similar ?)

Reasons ?
- Intuitive – chemically obvious
- evolution

More on this next semester

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 ?

- given human haemoglobin, you would find horse, pig, and 10^3 globin structures
- given a strange enzyme from an exotic virus, it may have no obvious homologues – it has evolved too much
- blast / psi-blast / fasta / HMMs

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

Length and degree of similarity

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
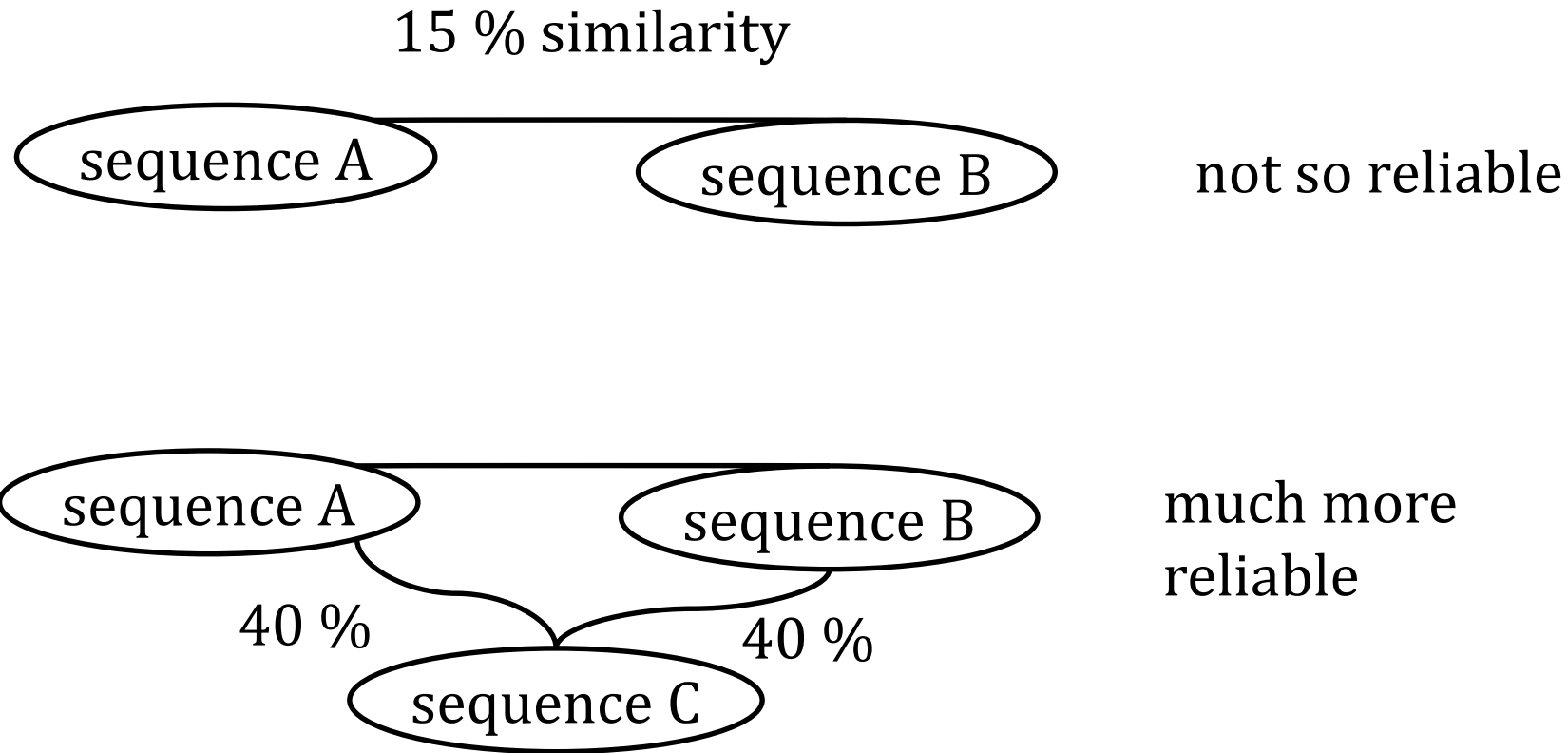 - more convincing

Quantifying importance of similarity length

We know the size of an alignment
how often are the two proteins
not structurally related ?
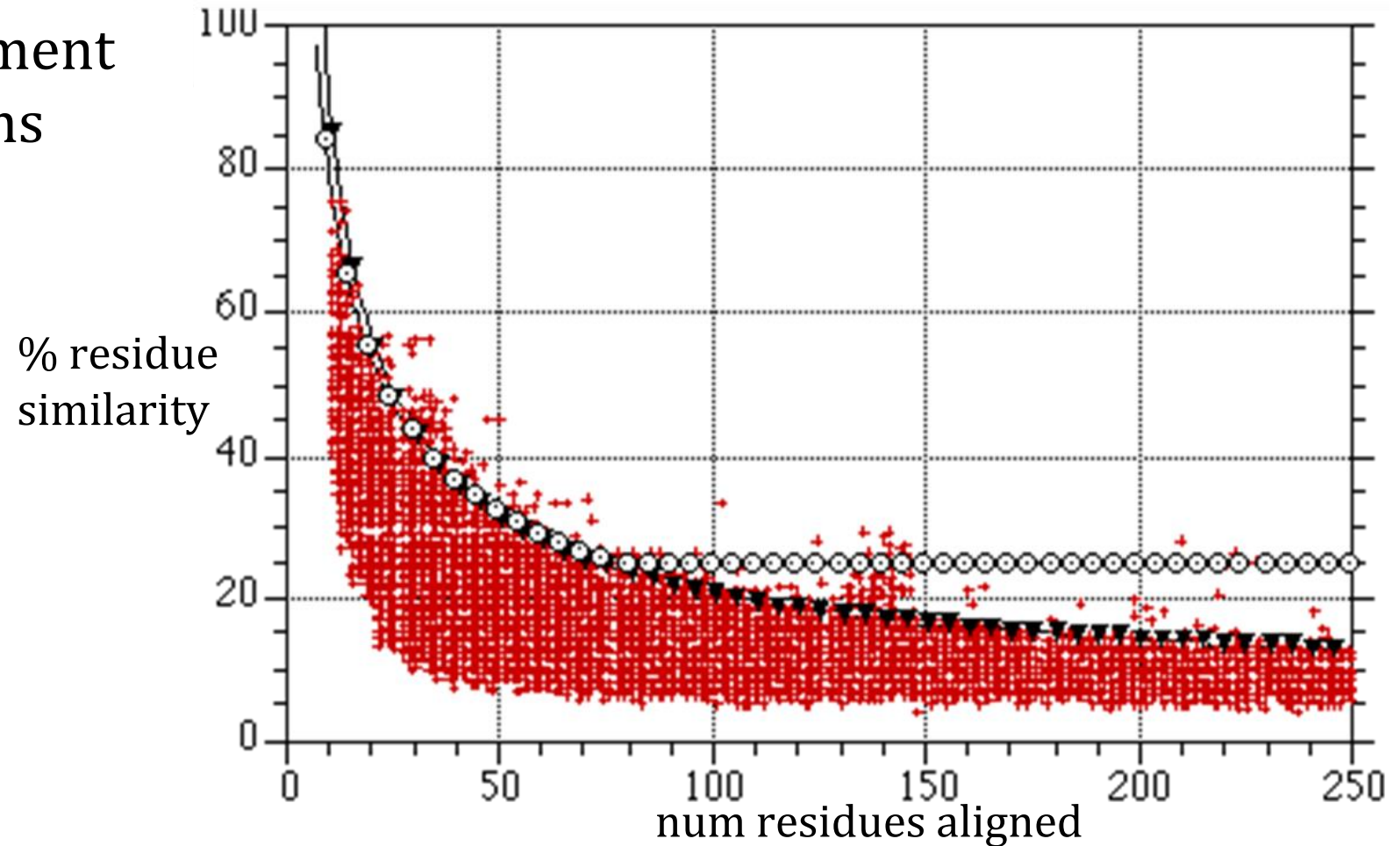
but there is more to deciding whether or not similarity is significant

More to reliability

15 % similarity

sequence A ⎯⎯⎯⎯ sequence B   not so reliable

sequence A ⎯⎯⎯⎯ sequence B   much more
reliable

40 %   40 %

sequence C

How significant is the similarity between two proteins ?
- does not only depend on the two proteins
- psi-blast in sequence lectures

Summarise

Sequence identity is most important

It is not enough to say 20 – 25 % similarity

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

Need an alignment

- difference compared to database searches ?
  - not scanning a database (10^7 sequences)
    - we can do best possible alignment

Careful alignments

Computer time not a problem - use
- most expensive alignment algorithm, could be one of
  - Needleman and Wunsch
  - Gotoh
  - Smith and Waterman

Maybe
- fancy substitution matrix
- play with gap penalties

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

Difficult alignment example

- sequence with unknown structure `ANDREW`
- sequence of structure `ANDRWQANDRKWSANDRWWC`
- reasonable alignments

```
ANDR-WQANDRKWSANDRRWC
ANDREW---------------  guess 1          [ includes gap
-------ANDREW--------  guess 2
--------------ANDREW-  guess 3
```

- Is one correct ? More likely to be correct ?
- guess 1 - a residue has disappeared (difficult to model)
- guess 2 - `K->E`;   guess 3   `R->E`
- very dependent on
  alignment quality / scoring / substitution matrix

Sidechains – should we worry ?

When do we not care ?
- for some residues, not meaningful (ala/gly)
- some residues entirely on surface of protein
  - interact with solvent
  - barriers to rotation ?   smaller than kT
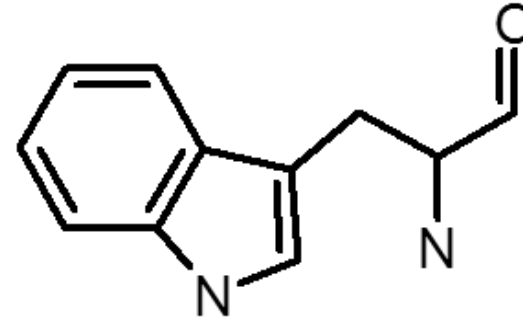 - all conformations accessible
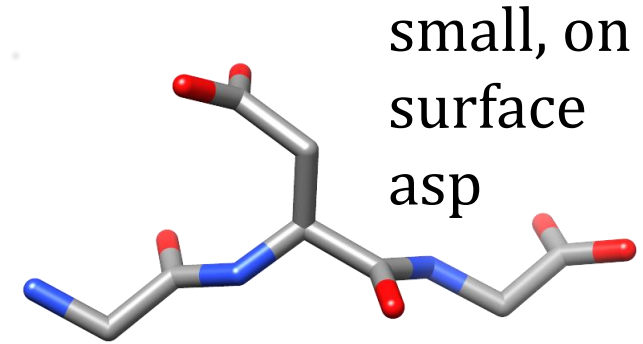
When is it sensible to worry ?
- sidechain is big and buried
- sidechain is charged and buried (salt bridge ?)
- example – trp usually
  - big
  - buried
  - hydrophobic
  - not very mobile

small, on surface asp

big buried trp
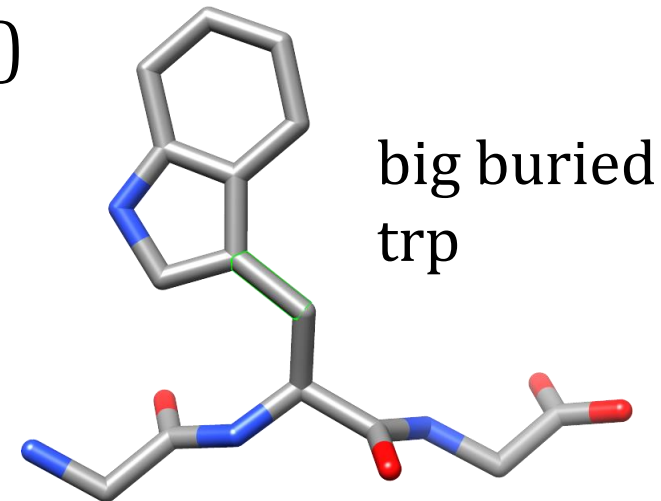
Sidechain placement

How to place sidechains
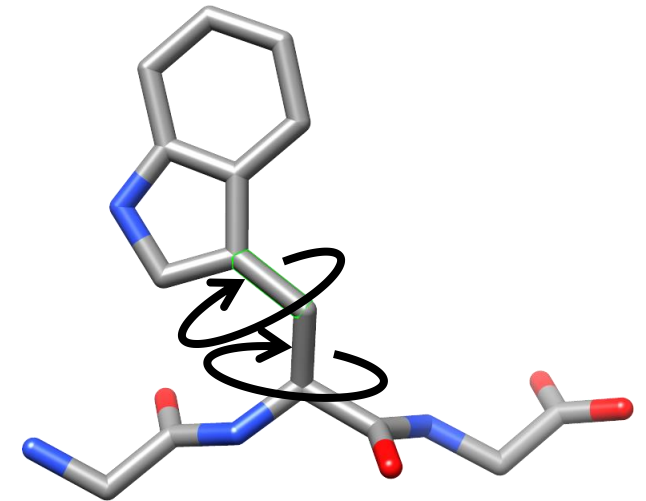- if identical to parent
    - re-use parent coordinates

Question
- what angle should I have at each rotatable bond ?

Reasonable strategies
- initial placement
    - random
    - probabilities from protein data bank ?
- fix !..

Fixing sidechains

Considerations
- atoms do not lie on top of each other
- residues like to pack (few holes in proteins – energy arguments)
- hydrophobic residues like each other
- charged and polar residues usually talk to solvent
- buried charges in salt bridges / no free charges in protein core

Can we write this down as a formula ?
- almost
  - an energy function should contain this (next Semester)

Optimising sidechains

Basic philosophy

- write down some function for energy +
  - energy minimisation
  - molecular dynamics
  - Monte Carlo / simulated annealing
  - self-consistent mean field methods
  - clique method – our example
- so as to rotate side-chains / make conformations more likely

Rotamers and cliques

Many ways to optimise side chains

- annealing, simulations, self-consistent mean field optimization
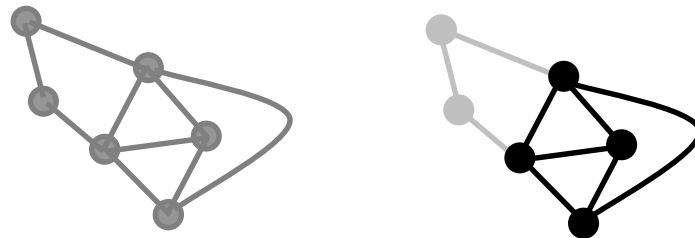
Clique detection

- just one example (not best, fastest, ...)

Ingredients

- side-chain rotamers (discretisation)
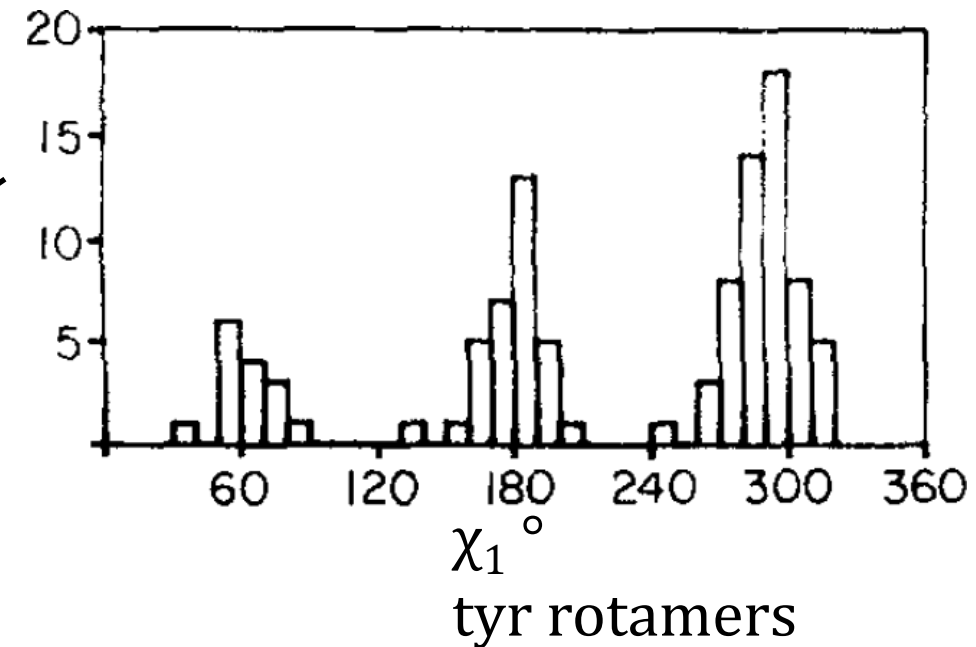- score for energies / clashes

Definition

- clique – subgraph where each point is connected to all others

Rotamers

Most sidechains have rotatable angles (more than 1)

- for each angle – usually 2 or 3 angles are more likely
- approximate:
  - pretend each side chain may only exist in one of the preferred positions "rotamers"
  - per sidechain
    - maybe 3, 9, .. rotamers
- crude ? yes
- useful ?
  - transform problem into a smaller search

χ_2

χ_1

χ_1 °

tyr rotamers

McGregor, M.J., Islam, S.A., Sternberg, M.J.E. J. Mol. Biol, 198, 295-310 (1987)
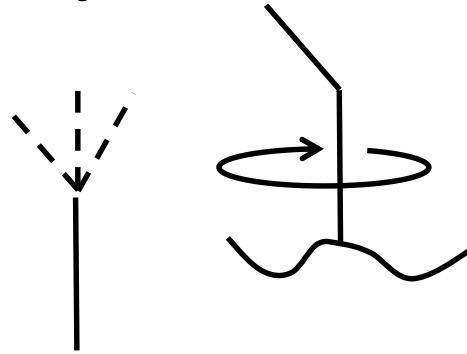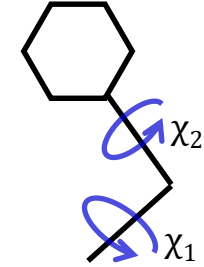
Rotamers

Fitting rotamers in a protein

Simple quasi-energy function
- atoms may not clash
- imagine 0 is fixed
- 0 does not fit with 1
  - OK with 2 or 3
- 1 is not OK with 0, 2, 3
  - OK with 4, 5, ...9
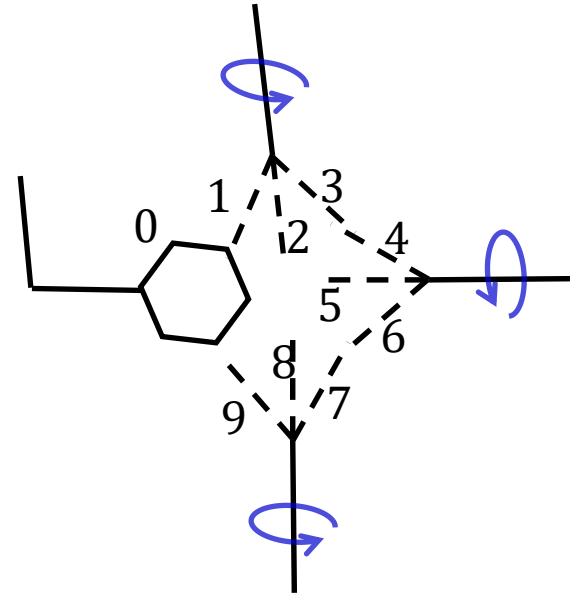
What we want – lists of who is compatible with who

Rotamers

Draw as a graph

- lines connect who is compatible with who
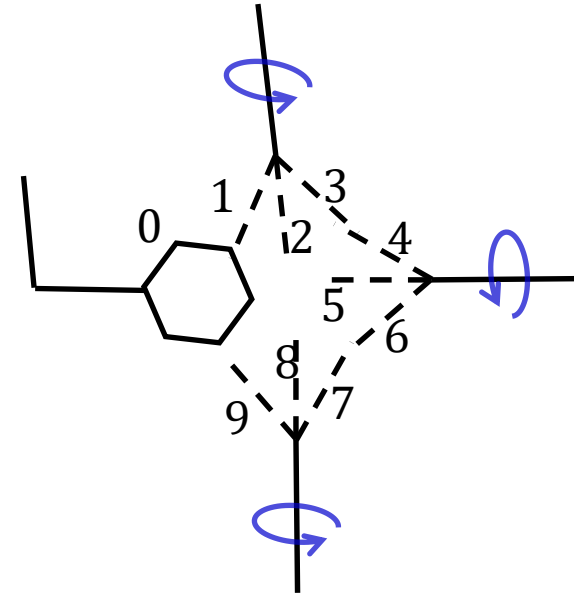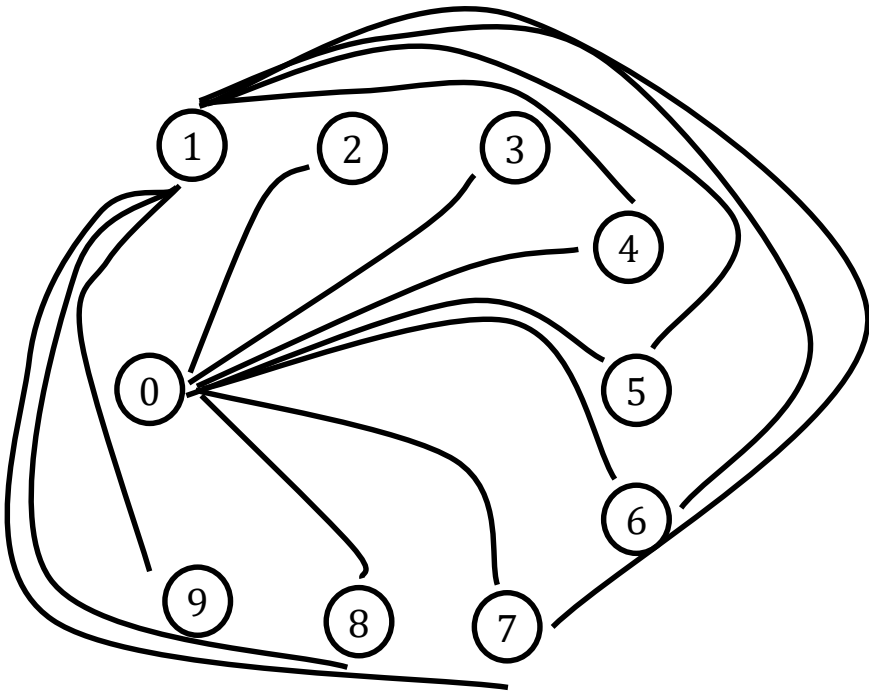
- connections for 0 and 1 drawn
- do for all other nodes (rotamers)
- no edges between nodes for 1 residue

Rotamers

Imagine there is only one possible set of rotamers

- every node (rotamer) will be connected to every other
    - is a clique

Imagine there are two solutions

- there will be two cliques

Application

- take protein
- build graph
- find all cliques
- write out lists of sidechain conformations
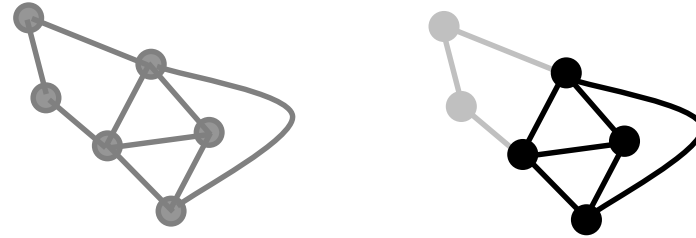
What was a very difficult problem seems to be tractable but...

Rotamers – problems with cliques

Killer problem

- finding maximal cliques is very very difficult

Rotamer concept

- side chains do not exist at only 0, 120, 240°

Better energy functions are more complicated

- not compatible/incompatible
- requires thresholds

Broken main chain

Typical situation

```
ANDR-WQANDRKWSANDRWWC  parent
ANDREW---DRKWS--DRWWC  model
```

our model…
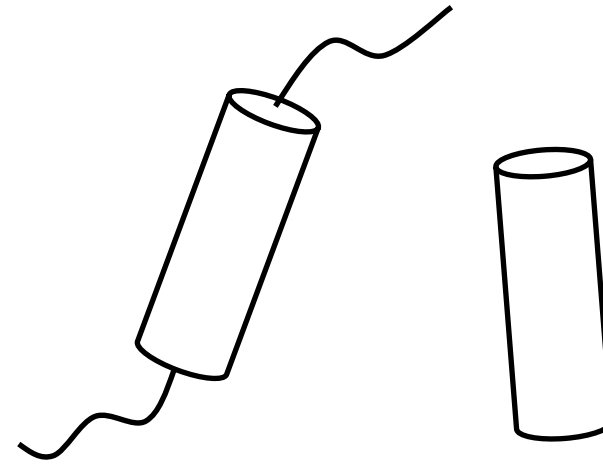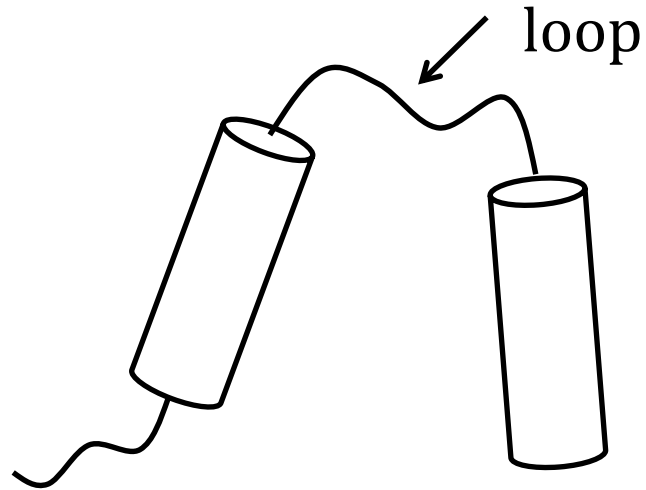
loop

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 (Modeller Übung)
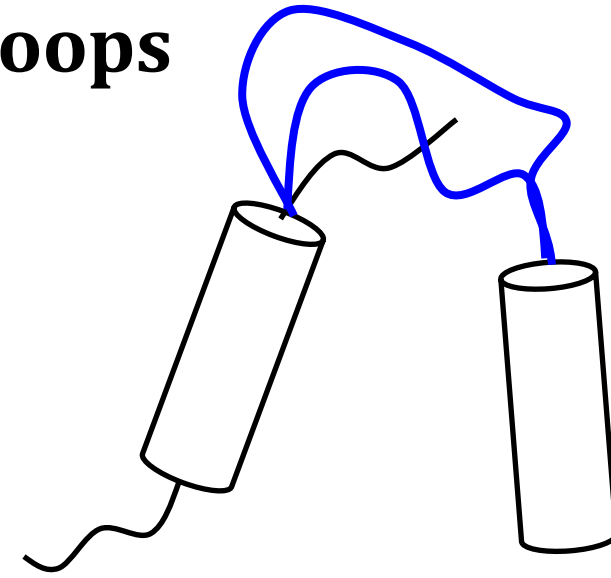
Results ?

- 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
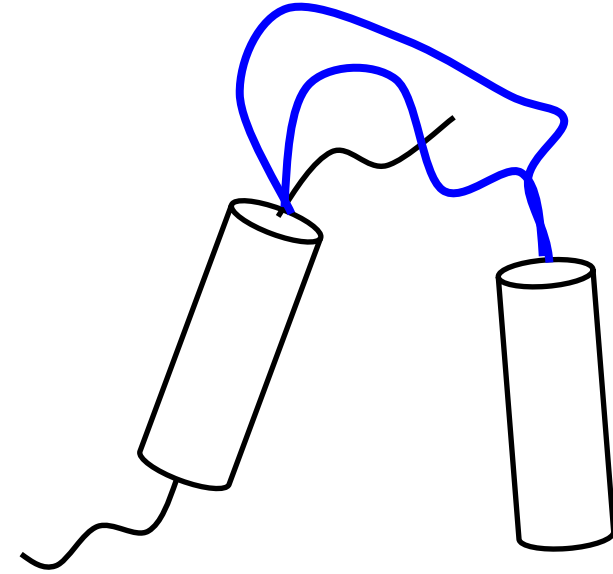
Loops – brute force

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 = $144^{N_{res}}$
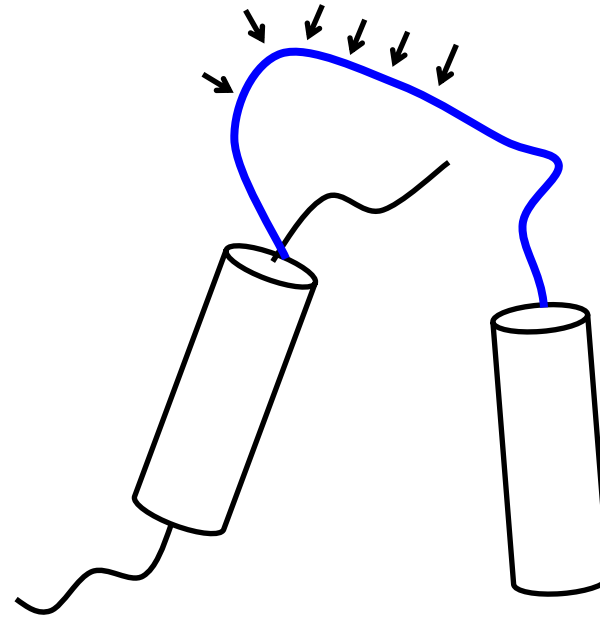
General repairs

What do we have now ?

- sidechains placed and maybe optimised
- rough guess coordinates for all residues (including loops)

Broken ?

- sidechains and loops often wrong
- small changes in other parts of structure
- time for last refinement .. again
  - energy minimisation / molecular dynamics / ...

Quality

General vs specific

- general
  - energies / geometries (almost the same)

- specific properties of this protein (vague and not for exams)
  - expected residues in active site
  - known reactive residues on surface
  - ... any experimental data

Checking by energy

Use a classical energy function (details next semester)
- if physics were perfect, would include all ideas mentioned
- details good (atom overlap, angles, ..)

Statistical approach
- take features you believe in
  - hydrophobic residue on surface, buried residue in middle..
  - phi / psi distributions
  - count occurrence in databank
- count occurrence in your model
- see if model is statistically plausible

Real world

Recipe on these slides ?
- too simple
  - steps combined / repeated
  - usually many models generated and checked multiple templates
  - multiple templates simultaneously ?

- automatic methods are very good

What does one achieve ?

Very easy cases ?
* not much change from parent

Very difficult ?
  * lots of errors

Why bother ?
* good modellers are experts on their systems
* some proteins are so important (money) – no waiting on
  * experiment
  * competitors
* simple predictions
  * which residues may I modify (binding to sensor...)
* consider absolute limits

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
sktyavlglgngghafaaylalkgqsv--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--gtahpdlltsdiglavkdadvilivvpaihhasiaaniasyisegqli---ilnpg
llmmgtafmgyvlpwgqmsfwgatvitglfgaipg--igpsiqawllggpavdnatlnrf
    :    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
```
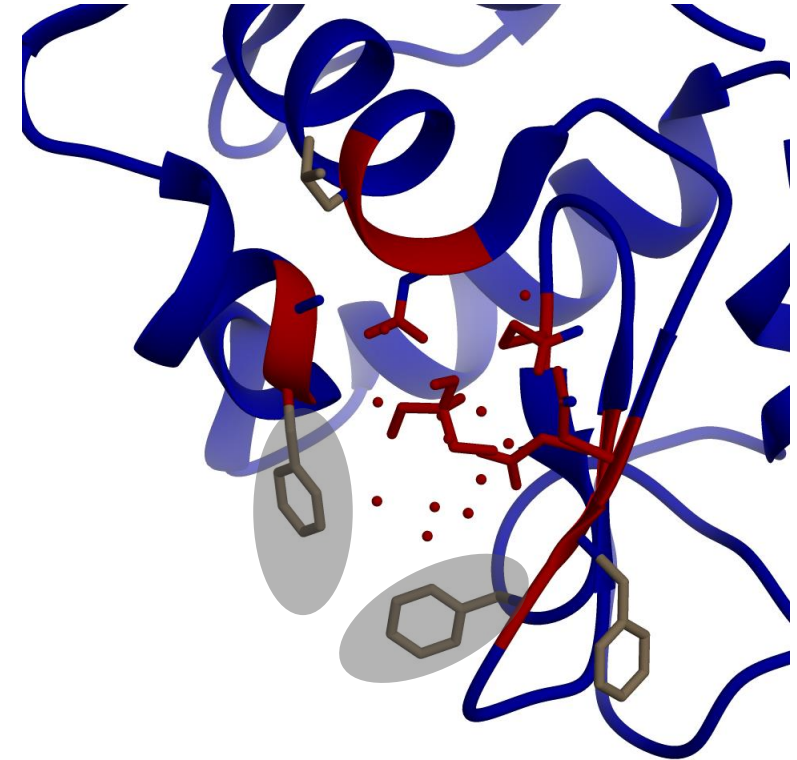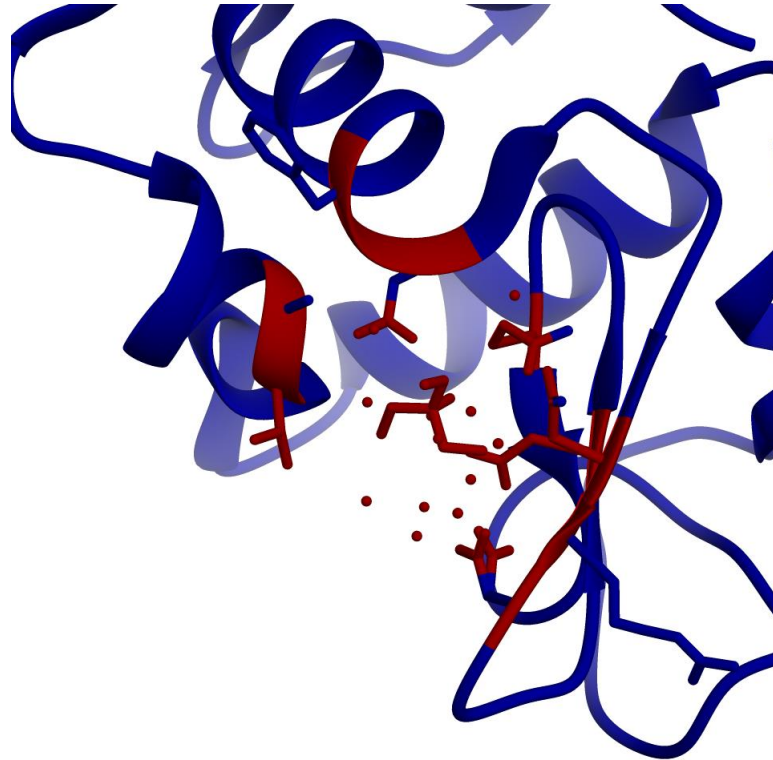
2mnr and 4enl example
- sequence alignment not the same as alignment from structures
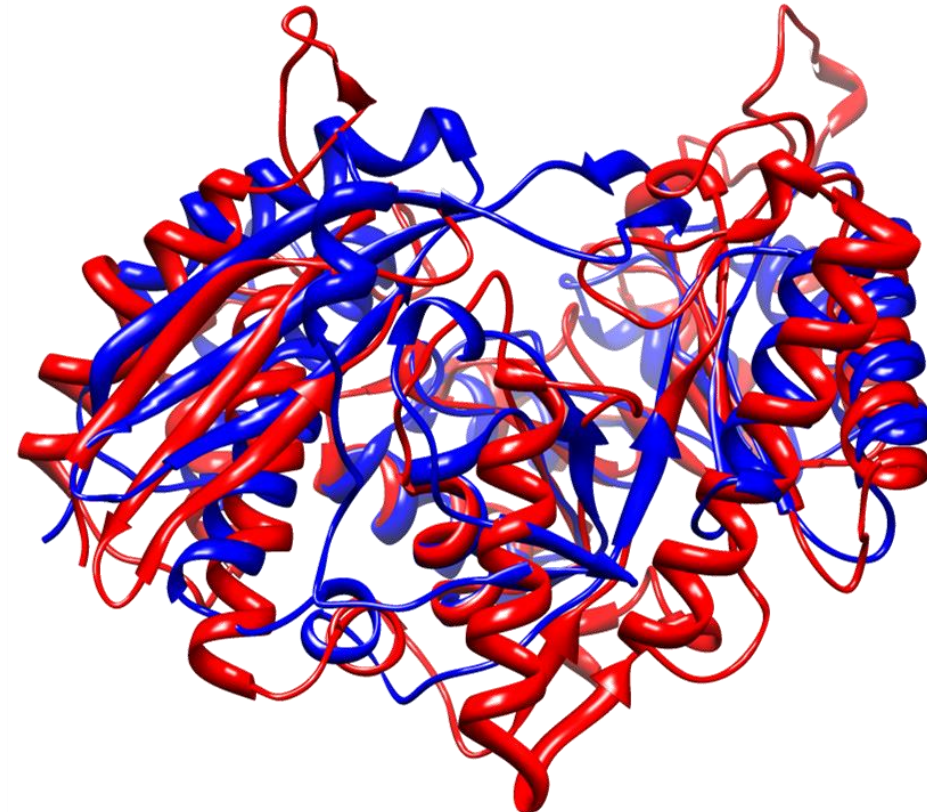
Why I do not like modelling

You build a model – decide that the grey residues are important

- Has the model helped you ?

`aacsdecgh...` known structure

`aacfdefgh...` your sequence

but we could look in original structure and mark all the residues near active site

`aacsdecgh...` known structure

and look for corresponding residues in your sequence

`aacfdefgh...` your sequence

Summarise für Klausur

Ideas of sequence similarity

Technical issues
- loops
- sidechain placement

None of the vague statements
- quality