Comparative / Homology Modelling

Andrew Torda, wintersemester 2008 / 2009, GST

Viciously abbreviated – one topic in depth

- rotamer optimisation
- remote sequence alignments
- loop prediction ---- my preference
- substitution matrices

My plan

- Quick overview of template selection
- loop prediction

- we keep track of topics for the exam

Who cares

Experimental structures are best, but

- not all proteins can be
  - expressed
  - crystallised
  - solubilised
  - labelled (for NMR)
  - assigned / phased …

Sometimes we know

- protein is vital to disease / function
  - from classical chemistry / biochemistry

Most basic rule

Mission

- make a model (guess for coordinates) from sequence information

Available information

- sequence always available
- possibly
  - some functional information
  - some chemistry

Guiding belief

- similar sequence gives similar structure
  - overall fold
  - local segments – think chemistry

Expectations of a model

Expectations

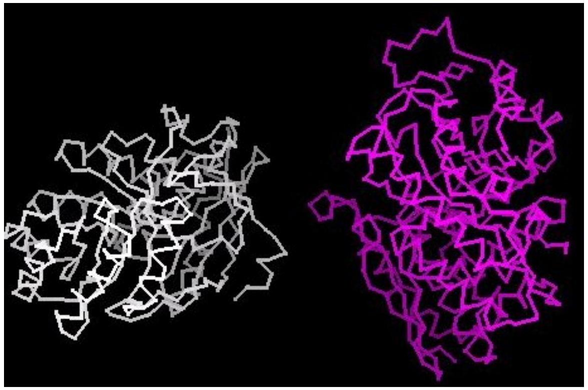
- is model enough ?
    - maybe for
        - designing a drug ? difficult
        - finding essential residues
        - locating differences compared to related structures

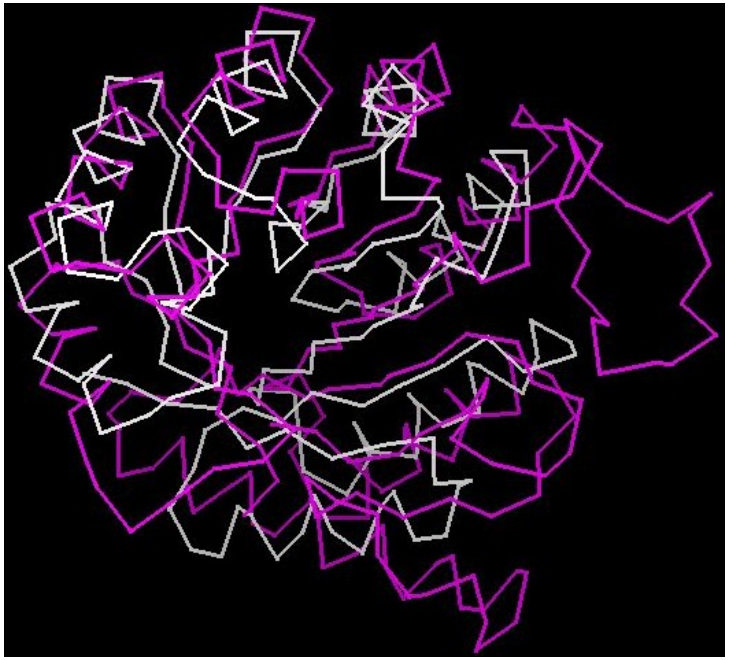
Fundamental hope

- If two proteins have a similar sequence, structures are similar
- we can build a good model for one protein using structure from a related one (of known structure)

Reasonable expectations

- two proteins, 2mnr, 4enl have easily detectable sequence homology
- could one have been modelled, knowing the other ?
- knowing the structures below, this is the limit of what could be done





Sequence and structure similarity

Two proteins with similar sequence

- how likely is similar structure ?
  - question of degree (how similar ?)

Reasons ?

- intuitive
- evolution
- physics (not today)

Intuitive

- people, pigs and horses have blood, breathe and need haemoglobin
- organisms are not identical, but similar
- there must be lots of haemoglobin like proteins

Evolutionary reasons

What does **NOT** happen

- living human, pig, e. coli
  - a single residue mutates
  - protein adopts a totally new structure
  - cannot carry out function
    - not a robust system

Consequence

- proteins must be able to tolerate mutations and keep working
- sequences must vary
  - structure and function do not change too much
- possible sequences are explored
  - continuously
  - randomly (almost)

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
6. verify

Finding a template / parent

How unique is my sequence ?

- given human haemoglobin, you would find horse, pig, and 100s of haemoglobins
- given a strange enzyme from an exotic virus, it may have no obvious homologues – it has evolved too much
- blast / psi-blast / fasta / hidden Markov models (Prof Kurtz lectures)

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

Why are these figures vague (≈ 15 to 25 %) ?

- Important factors
  - length and degree of similarity
  - number of similar sequences

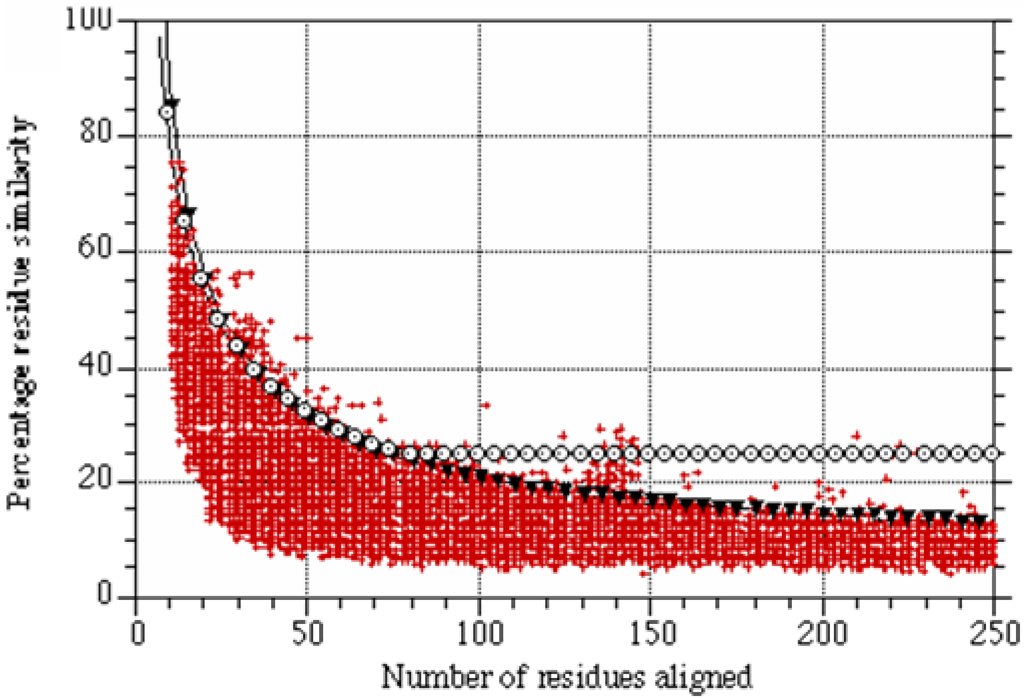
Template reliability

Length and degree of similarity

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

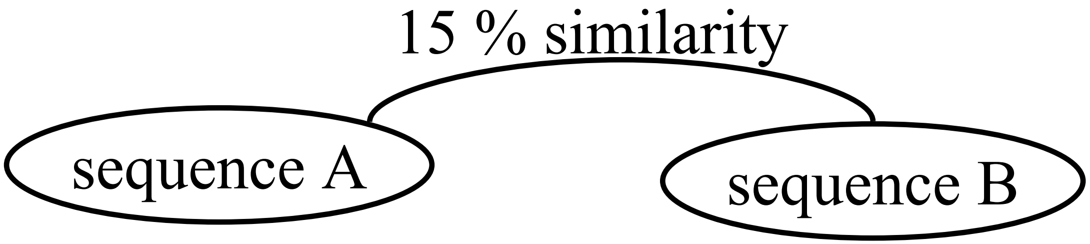
- Figure from last semester (purely empirical)
  - we know the size of an alignment, how often are the two proteins not (structurally related)



- but there is more to deciding whether or not similarity is significant

http://cubic.bioc.columbia.edu/papers/1999_twilight/paper.html

More to reliability

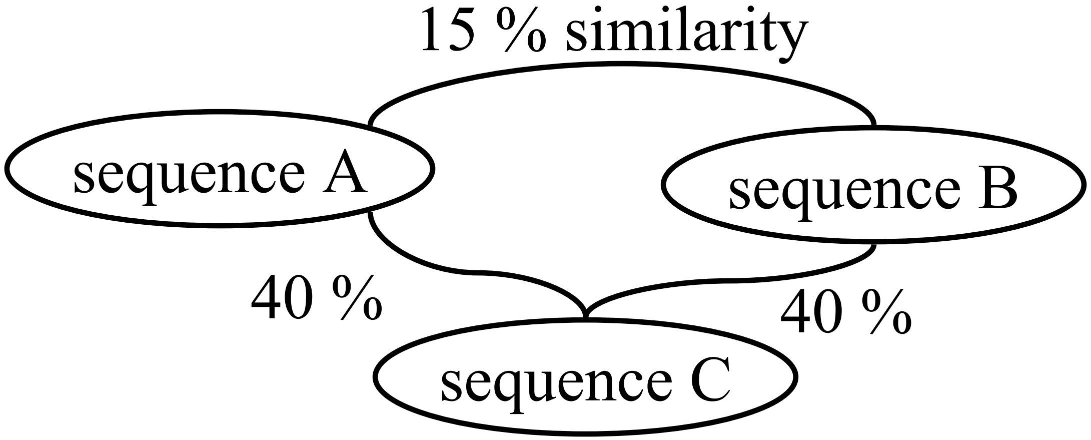


15 % similarity

sequence A

sequence B

not so reliable



15 % similarity

sequence A

sequence B

40 %

40 %

sequence C

much more reliable

- how significant is the similarity between two proteins ?
  - does not only depend on the two proteins
- reminder of psi-blast method ...

Blast, Fasta, Psi-blast reminder

We have a database of all protein sequences
+ a list of all structures

- search database of structures to find closest known structure
  - scan every sequence using fast method (blast, fasta)
  - do not do full optimal alignment
- psi-blast decoration (important / effective)

  ```
  while (not converged)
      scan database of all sequences (not just structures)
      collect close homologues
      build profile / modify score matrix
  ```
- maybe database includes structure files or homologue sequence information can be used on structures

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
6. verify

- we need an alignment
- how does this differ from the style described in other lectures ?
    - not scanning a database (10^6 sequences)
    - one or few alignments
        - we can do best possible alignment

Careful alignments

- Database scanning uses approximations
- Now, computer time not a problem
- Use
  - most expensive alignment algorithm, could be one of
    - Needleman and Wunsch
    - Gotoh
    - Smith and Waterman
  - careful selection of substitution matrix
  - careful selection of gap penalties
- example..

Difficult alignment example

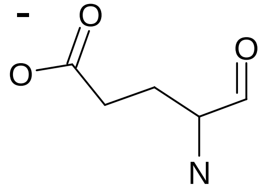
- unknown sequence ANDREW
- sequence of structure ANDRWQANDRKWSANDRWWC
- reasonable alignments

```
ANDR-WQANDRKWSANDRWWC
ANDREW---------------  guess 1 [ includes gap
-------ANDREW-------C  guess 2
--------------ANDREW-  guess 3
```

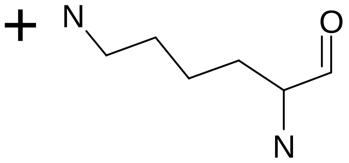
- How do they differ ? Is one correct ? More likely to be correct ?
- guess 1 means that a residue has disappeared (difficult to model)
- guess 2 involves K->E, guess 3 W-> E
- Intuitively ?
- Quantitatively ? substitution matrices ...

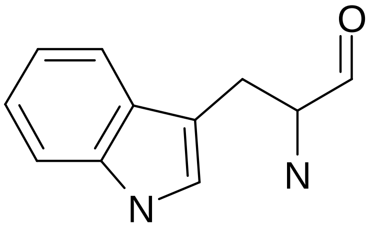
Amino acid substitution matrices

- Intuitively
  - measure amino acid similarity
  - as a chemist ask is this glu like another charged sidechain ?



or a huge hydrophobic sidechain ?





Think evolution

- if a "-" residue mutates to a "+", will it kill the organism ?
  - maybe
- if it mutates to a large, greasy, insoluble residue will it kill you ?
  - more often

Substitution matrix - what should it say

- a boring matrix (like DNA)

	A	C	G	T
A	1	0	0	0
C		1	0	0
G			1	0
T				1

- a more interesting matrix
- it tells us that
  - cys (C) is special (does not want to mutate to anything)
  - glu and asp are similar
  - phe and tyr are similar
  - real matrices
    - 20 x 20 (at least)
- Where do real matrices come from ?
  - chemistry ? No
  - evolution ? yes

	A	C	D	E	F	..	Y
A	5	0	1	1	1	..	1
C		10	0	1	1	..	1
D			6	3	0	..	0
E				6	0	..	0
F					10	..	8
...						..	..
Y							10

Building substitution matrix (collect data)

- Similar sequences are easy to align (by hand ?)
- count how often a residue changes to each other typ

ANDRWSANDRK and **WPANLHREWAN**

ANERWSANDRK and **WPLNLHREHAN**

- there is no question about alignment (obvious)
- immediately collect rate of change data
- some residues almost never change to anything
- some pairs change often
- turn into similarity matrix ?
    - take log(M_{ij})

	A	L	W	H
A	3	1	0	0
L		1	0	0
W			2	1
H				1

log-odds scores

- will re-appear in a chemical context later
- look at mutations and see **X**-> **A**
  - is this interesting ?
    - how common is **A** ?
- general logs-odds probability $$score = \log\left(\frac{N_{obs}^{AB}}{N_{exp}^{AB}}\right)$$
  - must define N_{exp}
  - for substitution frequency f_{AB}
  - $$N_{exp}^{AB} = \frac{N_A}{N}\frac{N_B}{N}N$$
  - log vs $\log_2$ vs ln
    - not important

Substitution matrix – remote homologues

Recipe above

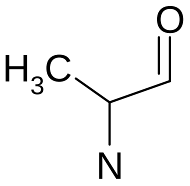
- based on reliable data
- good for similar sequences – maybe not remote homologues

Remote homologues

- A<->B easily
- B<->C easily
- A<->C less frequent
- so between close neighbours
  - AC change much less than AB or BC
- remote homologues, more evolution, more A<->B<->C
  - over long time, A<->C will seem more frequent
- use different matrices – depending on how remote homologues

Sidechain replacement

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
6. verify



How reliable are any sidechains ?

- depends on
  - size
  - interactions
  - temperature
  - location (buried, accessible)

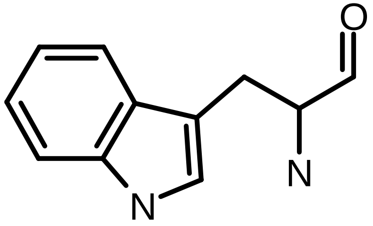
Sidechains – should we worry

When do we not care ?

- for some residues, not meaningful (ala example)
- 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 ?

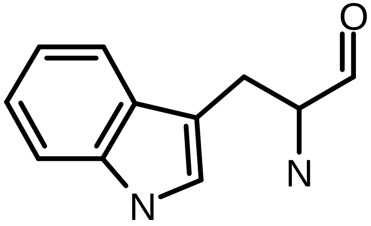
- sidechain is big and buried
- sidechain is charged and buried (salt bridge ?)
- example – trp usually
  - big
  - buried
  - hydrophobic
  - not very mobile

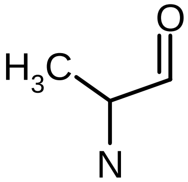


Sidechain placement

How to place sidechains

- if identical to parent
  - re-use parent coordinates
- in all cases C^{β} is known from backbone
- 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 (more later)

Can we solve this like a conventional formula ?

- no...

Fix structures from a formula ?

- You are asked to minimise $y = (x - 5)^2$
- easy
- Our function
  - variables are hundreds of (*x*,*y*,*z*) coordinates
  - many almost similar answers
  - no analytic solution
- Energy functions in detail soon

What can one do ?

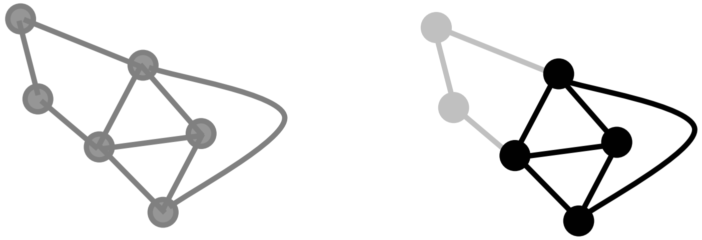
- there are ways to reduce energy of a structure..

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

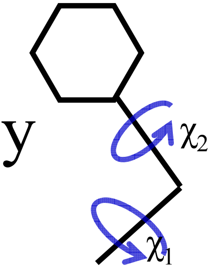
Rotamers and cliques

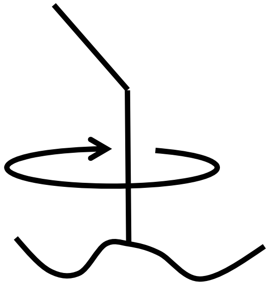
- Many ways to optimise side chains
  - annealing, simulations, self-consistent mean field optimization
- Clique detection
  - just one example (not best, fastest, …)
- Ingredients
  - side-chain rotamers (discretisation)
  - score for energies / clashes
- definition
  - clique – subgraph where each point is connected to all others

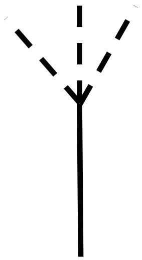


Rotamers

- 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

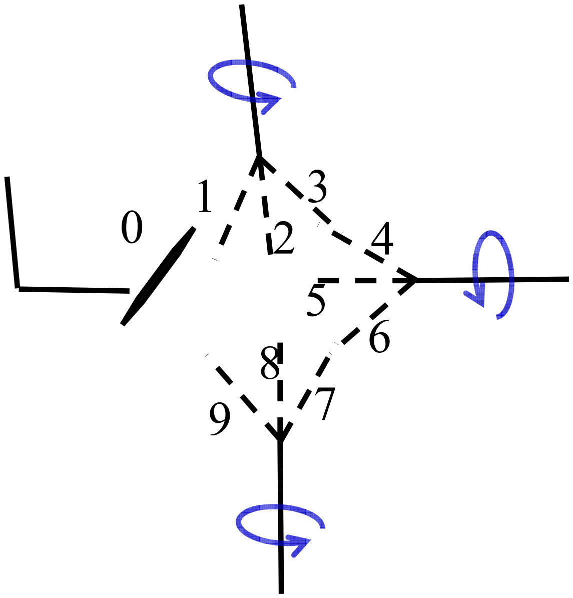






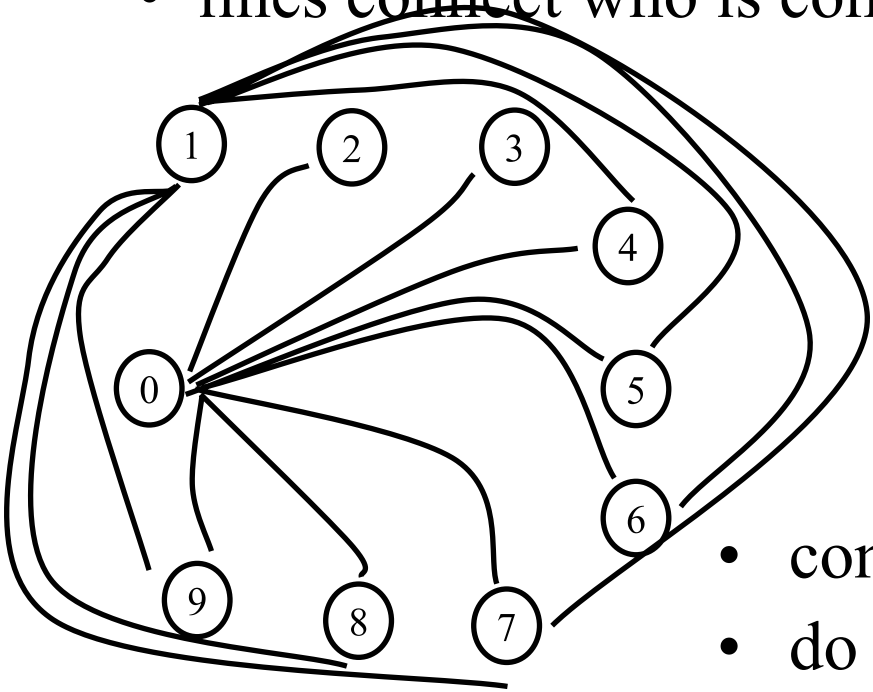
Rotamers

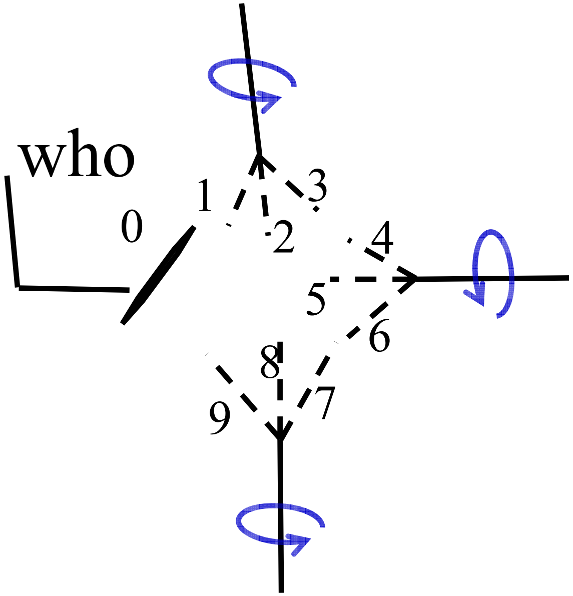
- Fitting rotamers in a protein
- simple quasi-energy function
    - atoms may not clash
    - imagine 0 is fixed
    - 0 does not fit with 1
        - OK with 2 or 3
    - 1 is not OK with 0, 2, 3
        - OK with 4, 5, …9
- what we want – lists of who is compatible with who



Rotamers

- draw as a graph
  - lines connect who is compatible with who

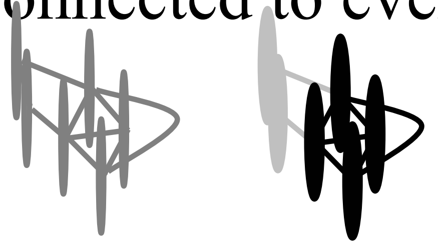


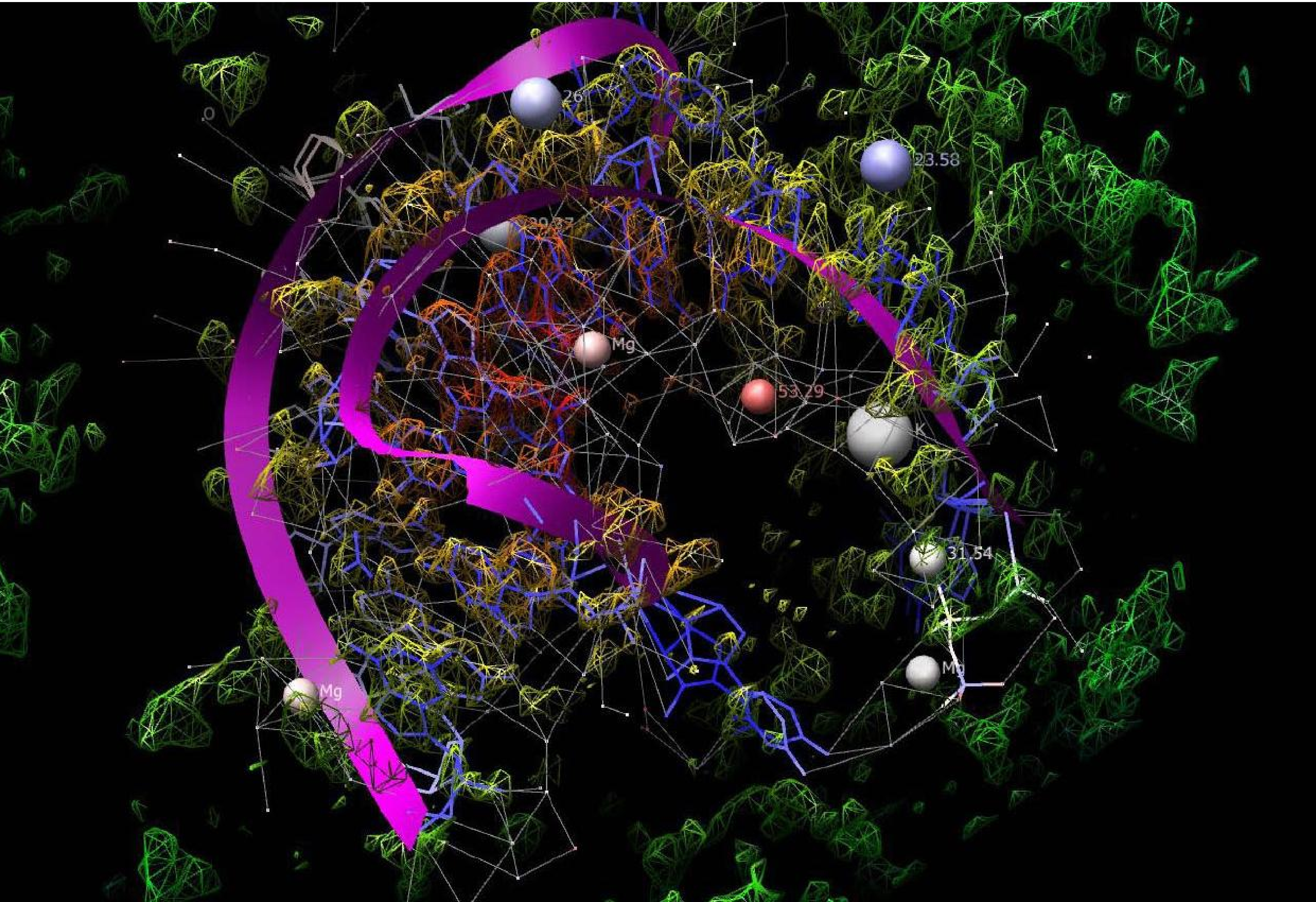


- connections for 0 and 1 drawn
- do for all other nodes (rotamers)
- no edges between nodes for 1 residue

Rotamers

- imagine there is only one possible set of rotamers
  - every node (rotamer) will be connected to every other
    - = 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…




0
26
23.58
Mg
53.29
K
31.54
Mg
Mg

Rotamers – problems with cliques

- Killer problem
  - finding maximal cliques is very very difficult
- Rotamer concept
  - side chains do not exist at 0, 120, 240°
- Better energy functions are more complicated
  - not compatible/incompatible
  - requires thresholds

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
6. verify

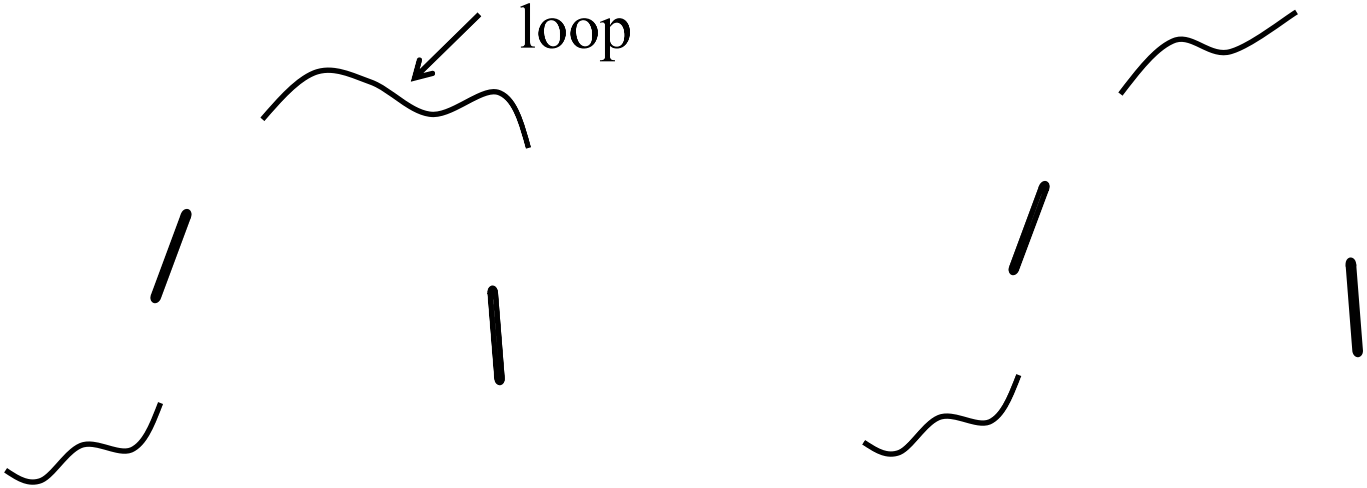
Broken main chain

- Typical situation

`ANDR-WQANDRKWSANDRWWC` parent

`ANDREW---DRKWS--DRWWC` model

our model…

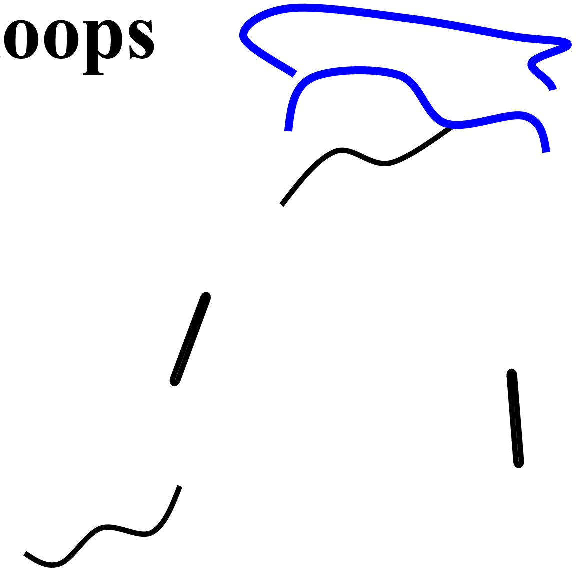


- basic problem…
  - pieces of unknown structure
  - endpoints relatively fixed
  - should be joined

Loop modelling

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

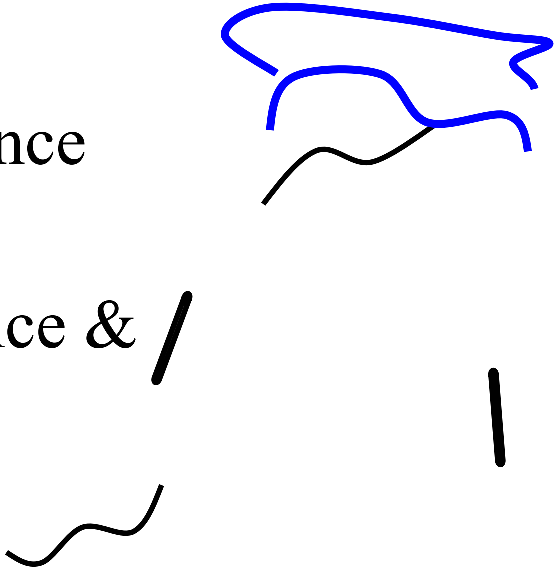
Results ?

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

Loops Database searching

Database searching

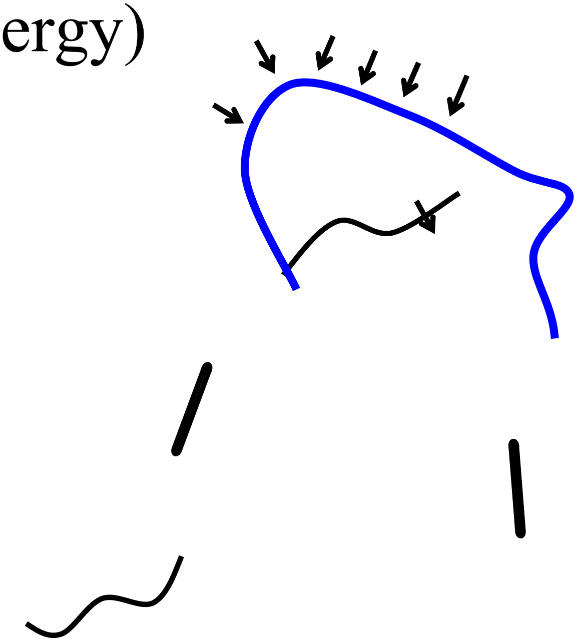
- imagine we have a 9 residue loop
- take protein data bank
- collect coordinates of all 9-residue loops
- insert those with correct end to end distance
- refinement…
  - insert those with almost correct distance &
  - similar sequence to loop residues



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

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**
6. verify

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 / …

Verification

General vs specific

- all proteins have some characteristics
- your protein may have some specific properties

General properties (from previous slide) – easy to check ?

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

 +

- backbone angles / ramachandran plot

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, ..)
- weakness ?
  - may be poor at overall structure

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

Specific protein properties

Collect known properties

- mutation data
  - are any residues vital ? does the model disagree
  - does it disagree with known facts ?
    - a set of residues are known to be vital in every related protein
    - are they disturbed in model ?
- sequence motifs ?

Chemical predictions (examples)

- only interesting if you can predict
  - something new / testable
    - predict a charged residue is buried (asp, glu)
      - must have a changed pK_a
  - active site is changed
  - changed susceptibility to
    - reduction / oxidation…

Real world exercises

Recipe on these slides ?

- too simple
  - steps combined / repeated
  - usually many models generated and checked
  - interaction with experiment (predictions tested)

Expectations

- Easy cases – near Å accuracy
  - your sequence is 90 % to something of known structure
  - part of a large family of proteins
- Hard
  - less than 25 % homology + few homologues
  - consequence – alignment will not be perfect
    - some predictions will be wrong
- Worse
  - membrane bound / interacting

What does one achieve ?

Very easy cases ?

- not much change from parent – could work there

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

Back to first example

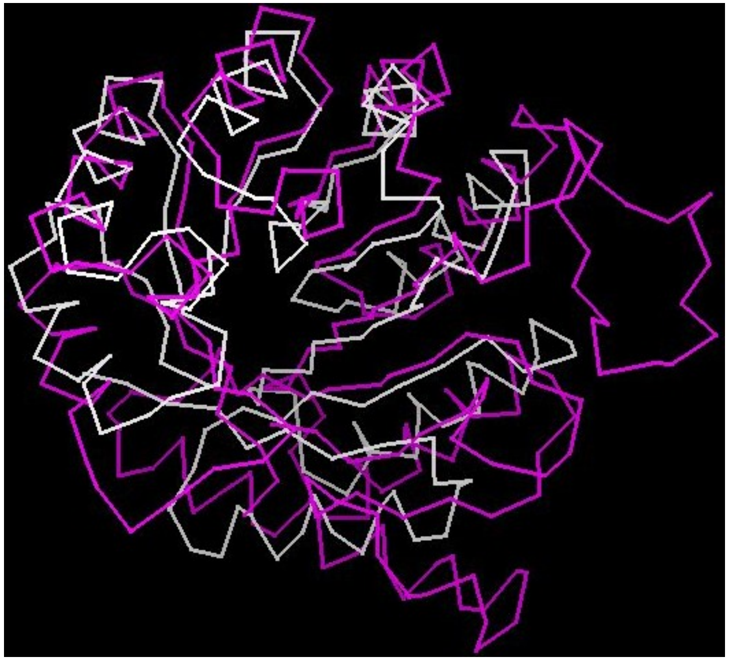
- 2mnr and 4enl
- would be a typical modelling target
- in real world
  - alignment would not be perfect
  - loops may be quite wrong

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--lawdidagr------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

- this does not give best structures
- this alignment does not correspond to the nice picture



- next semester… energy functions