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
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
• ...

Andrew Torda
07.12.2015 [2]
Modelling

...ADEFINEH I KH-GED...

- do a blast search ... find
...AQDEF-HIKKGFED

replace original
..AQDEF_H..  

with your sequence
..AADEFGH..
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

Andrew Torda
07.12.2015
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?
  • 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

<table>
<thead>
<tr>
<th>high sequence identity</th>
<th>low sequence identity</th>
<th>very low</th>
</tr>
</thead>
<tbody>
<tr>
<td>(&gt; ~20-25 %)</td>
<td>(&lt; ~20-25 %)</td>
<td></td>
</tr>
<tr>
<td>blast, fasta, anything</td>
<td>psi-blast, HMMs</td>
<td>psi-blast, optimism</td>
</tr>
</tbody>
</table>

Why so vague?
Template reliability

Length and degree of similarity

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

• not very good
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 there is more to deciding whether or not similarity is significant

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

15 % similarity

sequence A  sequence B  not so reliable

40 %  40 %

sequence A  sequence B  much more reliable

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

- we need an alignment
- difference compared to database searches?
  (different to Georgio & Prof Kurtz)
  - 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
- careful selection of substitution matrix
- careful selection of gap penalties

How important?
Alignment errors

ANDREW
ANQEW

two reasonable alignments

ANDREW or ANDREW
ANQ-EW or AN-QEW

difference?
• from $C_i^\alpha$ to $C_{i+1}^\alpha$ almost 4 Å
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

- Is one correct? More likely to be correct?
- guess 1 - a residue has disappeared (difficult to model)
- guess 2 - K→E; guess 3 W→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
Sidechain placement

How to place sidechains

• if identical to parent
  • re-use parent coordinates
• in all cases $C^\beta$ 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 (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

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...
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

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
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 $\omega$
  • $\phi, \psi$  $12 \times 12 = 144$
• 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 / ...

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
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?
• interaction with experiment (predictions tested)

• 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

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

None of the vague statements
- quality