Protein Domains

Two weeks for this topic

• many proteins have separate chains

enterotoxin (1lts) 7 chains

a haemoglobin (1h97)

• what about units within one chain?
4 domain protein

1cxl has 686 residues
Cleaves carbohydrate bond

- one solid lump but...
4 domain protein

1. α-amylase catalytic
2. α-amylase C-terminal
3. immunoglobulin like domain
4. starch binding

- even clearer example
3 domain protein

1qcf "src tyrosine kinase"

The domains really are common to other proteins

Number of domains is not absolutely defined
Plan

• chemistry, examples

• methods to automatically recognise domains (examples)

• chemistry – how common are domains of different sizes, types, ...
Earlier history

Term "domain" used before there were many structures

• Invented example: protein that
  • joins $\text{ADP} + P_i \rightarrow \text{ATP}$
  • performs some oxidation
  • responds to some regulator

• take protein + protease (splits protein in a few places)
• cleave / break protein - get a few pieces (2, 3, 4..)
• purify pieces

• pieces found that
  • can bind ADP/ATP
  • bind sugars, some regulators
Earlier history

Appeared that for some proteins
- different functions associated with different pieces
- refer to as "functional domains"

Belief / claim
- bigger proteins are made from units, combined over evolutionary time scales

• an example...
modular protein

Diptheria toxin (1ddt) middle of picture
- 3 domains
  - each similar to some different protein

• appears as if modules are mixed together
• should be visible at sequence level...

Sequence level domains

Align a group of sequences

a few hundred CNPases – red lines mark a domain

schematically..
Sequence level domains

Align a group of sequences

• appears to have 3 or 4 domains
• no reference to structures or function
### Domain definitions summary

<table>
<thead>
<tr>
<th></th>
<th>structure</th>
<th>sequence</th>
<th>biochemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td>functional</td>
<td>not necessary</td>
<td>not necessary</td>
<td>yes</td>
</tr>
<tr>
<td>sequence-based</td>
<td>not necessary</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>structure</td>
<td>yes</td>
<td>usually known</td>
<td>no</td>
</tr>
</tbody>
</table>

**How important?**

- > $\frac{2}{3}$ proteins have 2 or more domains
- part of definition
  - a piece of a protein which can fold and is stable

**Now**

- methods based on structure
Finding Domains

A definition leads to methods
• domain is a compact unit

Objective way to look for dense units?
• cluster analysis

Philosophy in cluster analysis
• look for dense groupings

Leads to dendrogram
Clustering

Approach

- need a (dis)similarity matrix between every object
- here: distance between C\(^\alpha\) atoms

\[
\begin{array}{c|cccc}
\text{residue} & 1 & 2 & \cdots & N \\
1 & 0 & \cdots & \cdots & \cdots \\
2 & 0 & \cdots & \ddots & \\
\vdots & \vdots & \ddots & \ddots & \\
N & & & & 0 \\
\end{array}
\]

- does this work?
Clustering

Clustering applied to concanavalin A

- bottom - small compact pieces
- higher – compact units
- looks like natural 3 domains

Number of domains is not absolutely determined

Very very very old method

Cuts / Surface area / volume

- Simple idea - cut chain in two pieces
- Density of part 1 / versus part 2
- Cut so as to maximise density

Problems - one cut is not enough

A method should be able to split with 1, 2, 3, ... cuts
- For 3 cuts with $N_{\text{res}}$ positions: $N_{\text{res}} \times N_{\text{res}} \times N_{\text{res}}$
  
  really $(N_{\text{res}})^{N_{\text{cut}}}$
Problems - density

I want to maximise density
• density of protein?
  • number of residues in a volume?
  • volume? not sphere

Contacts are easier than density
• within a domain there are many contacts
• between domains - few contacts
• an approximation
Counting contacts

Do I have many contacts compared to the number of atoms?

- calculate distance between each $C_i^\alpha C_j^\alpha$ atoms = $d_{ij}$
- if $d_{ij} < 4$ Å, set $p_{ij} = 1$ else $p_{ij} = 0$
- for a given set of $N_{res}$ atoms (not whole protein)

$$
\frac{\sum_{i=1}^{N_{res}} \left( \sum_{j>i}^{N_{res}} p_{ij} \right)}{N_{res}}
$$

- not accurate, but easy to calculate
Cutting / contacts

A protein

Find close contacts

How can one find the best place(s) to cut?

Feed water into \( s \) (to \( t \))
- find the most blocked restrictive pipes
- not one, but all that are restrictive
Cutting / contacts

- Flow problem
- Many ways to cut the flow from $s$ to $t$

  - of all these "$st$ cuts" find the one with smallest capacity (flow)
  - more interesting - make the pipes different flow capacity
    - how are the residues really touching? $C^\alpha C^\alpha$ or $C^\alpha$ sidechain

Xu, Xu and Gabow, Bioinformatics, 16, 1091-1104, (2000), Protein decomposition using...
Cutting / contacts

Two steps

1. find the maximum flow from $s$ to $t$
2. cut $s$ from $t$ at the few most filled pipes
Maximum Flow

rule

- if every path from source to sink has one full edge
  - flow is maximum

keep trying every possible path, look to see if there is unused capacity
we can go backwards

---
Maximum Flow

- A B D F  
- note DF is not quite full
- add some A C E F  
- AC, EF are not full
- look at C, switch some capacity to CD (DF)

- with some numbers on the edges

Maximum Flow

Define an example system

- flow into A
- out of F

- capacities at each edge vary
Maximum Flow

Find a new path (possibly with backwards flow)
- what is the smallest unused capacity on the path? $\Delta f$
  - $> 0$? send flow $\Delta f$ in this path
Alternative

- also ends with flow of 12
- look why this this is definitely time to stop ...
Alternative

• are there any paths with unfilled pipes?
• start at A
  • to the left is filled
  • try A → C
    • both routes out of C are filled

• more solutions?
• definitely different ways to find solutions
  • different order of visiting paths
Maximum Flow

- path = any route from A to F
- is there any path where all edges have extra capacity?
  - finished - flow is maximum
- algorithm (not optimal)
  
  while (flow not maximum / path found)
  
  add flow to path

Our definition - finished when

- every path from source to sink has at least one edge (pipe) which is full

Is this efficient?
Efficiency

Worst possible selection of path order would require 2000 iterations

First part of procedure finished
- flow is maximum
- next
  - where to cut graph
Find ways to cut network, max flow = 12

- AB, AC capacity = 14
- BD, BE, AC capacity = 17
  - both bigger than flow (12)

Better
- for each path
  - find first full pipe - cut
  - AB, CD, CE capacity = 12
  - = max flow
  - best cut
Cutting graph

If the capacity across our set of cuts = maximum flow
- it is a "minimum cut"

- smallest connection between two parts of graph
- graph / network / protein is broken into two parts / domains

Useful yet?
- no mention of finding source $s$ and sink $t$
- details - efficiency not mentioned
Network flow and proteins

Source
• find a surface residue
• connect an s
• connect to nearby surface residues

Sink
• find a surface residue far away, connect to $t$

Ad hoc? arbitrary? optimal?
• maybe not critical

Multiple domains?
  while (domains not too small)
    keep trying to split

Xu, Xu and Gabow, Bioinformatics, 16, 1091-1104, (2000), Protein decomposition using...
Splitting - near neighbours / Ising spins

Background story - Ising spin model
• energy of spin \( i \) depends on \( i, i + 1 \)
• energy can be good or bad

For lots of spins
• islands of same spin
• can be generalised to 2D, 3D

Finding low energies? Simplest method
• try to flip a random spin
  • accept flip if energy improves
  • sometimes accept if energy goes up (probabilistic)
Splitting - near neighbours / Ising spins

Slightly better method

while (energy still high)
    for each spin
        change to be same as average of $n$ neighbours

Protein version

• for any known structure
  • easy to make list of neighbours of each residue

• residues close in space should be in similar domains
Splitting - near neighbours / Ising spins

protein

make a list of neighbours for each residue

label all points with a number
Splitting - near neighbours / Ising spins

label of a residue is $m_i$
while (labels changing)
for each residue $j$

$$m_{av} = \frac{\sum_{i \in \text{neighbours}} m_i}{n_{\text{neighbour}}}$$

if ($m_{av} > m_j$)
  $m_j \ (\text{new}) = m_j \ (\text{old}) + 1$
else if ($m_{av} < m_j$)
  $m_j \ (\text{new}) = m_j \ (\text{old}) - 1$

<table>
<thead>
<tr>
<th>step</th>
<th>residue number</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1  2  3  4  5  6  7  8  ...</td>
</tr>
<tr>
<td>1</td>
<td>2  3  3  4  6  7  8  8  ...</td>
</tr>
<tr>
<td>2</td>
<td>2  2  4  3  5  8  9  9  ...</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Splitting - near neighbours / Ising spins

Properties of Taylor / Ising spin-inspired method

Optimism

- will converge and become stable

Requires threshold - what is a neighbour

- can use sophisticated averaging - distance dependent
- may converge to 2, 3, ... domains
Methods so far

1. simple - look for single cut points and maximise density

2. Crippen / hierarchical clustering

3. Network flow

4. Ising spin / Taylor

• All methods have arbitrary numbers
Why are methods so complicated?

If we cut protein chain once

- methods are easy - use density criterion

Cut protein twice? more? remember $(N_{res})^{N_{cut}}$

How many domains?

- Crippen / clustering method - whatever you want
- Network flow - repeatedly split and eventually stop
- Taylor / Ising - may converge to > 2 domains
Crippen / hierarchical clustering

At what level of hierarchy do I cut tree?

Network flow

- what constitutes a contact? (any 2 atom < 4 Å?)
- give pipes (edges) more weight for different kinds of contacts
- are solutions unique?
  - probably in practice
- when do we stop splitting domains?
Taylor / Ising spin method

- what constitutes a contact? how many Å?
- type of averaging to get $m_{av}$?
- when does one converge?

Elegance

Do methods work as described? not really

All authors report problems - example
- Taylor finds different results for α-helical and β-sheet regions
- simple explanation? distances within / between secondary structure are very different
Do methods work?

With many fixes and tuning - yes
- distance criteria, thresholds

Do methods agree?

Only ask question if you agree to think in terms of structure
- answer will be different in terms of evolution or sequences

Criteria
- how many domains inside a protein?
- where are the domain borders?
Number of domains

- test set of few hundred proteins
- compare against author's estimate
- 80-90% agreement

How many domains per protein?

Same set of 467 proteins

- authors split into several domains
- "SCOP" prefers smaller number of domains

Agreement?

Lots of room for differences

Some statistics

How big is a protein domain?

Peaks near 60 and 130 residues

How complicated are domains?

Justification for complicated domain recognition

- single cuts, double cuts in chains are not enough

What percentage of domains are built from
- 1 chain?
- 2 chains? ...

In "DALI", 23 % of domains are not continuous (multiple crosses of chain)

Evolutionary picture

Original claim
• domains are units that move as a module in evolution

If we see multiple cuts 10-20% of time
• picture is much less clear
Summary

Domain definitions
- functional, structural, sequence based

Finding domains
- relies on contacts, density
- method must be able to handle multiple crossings of chain

We considered
- clustering / hierarchical
- network flow
- Taylor / Ising spin-inspired

- Methods do not agree with each other
- Some trends in size and number of domains
- Real proteins are not as simple as evolutionary picture