Protein Domains

I690/B680: Structural Bioinformatics

February 12, 2006
Domains in 3-D Structures

• Initially
  – the concept of a domain was used to describe distinct regions of 3D structures
  – 1960s
    • lysozyme
    • ribonuclease
  – both lysozyme and ribonuclease contained spatially distinct structural units (termed domains)

PDB ID 1IY3
Solution Structure of the Human lysozyme at 4°C
Domains in 3-D Structure

• Over time it emerged that such domains could occur
  – in a variety of structural contexts
  – in multiple copies in the same polypeptide chain

• Today
  – domains are usually defined as spatially distinct structures that could fold and function in isolation
An Example

- part of the structure of hematopoetic cell kinase (1qcf) containing protein kinase and src homology 2 and 3 (SH2 and SH3) domains

Domains in Protein Sequences

- the arrangement of different domain types in proteins may cause problems in sequence analysis
Libraries of Domain Sequences

• various libraries of domain or motif alignments (and profiles associated with them) are available
• it is easy to automatically assign a domain in protein sequences based on these libraries

• Libraries
  – Pfam (HMMs)
  – PROSITE (patterns and profiles)
  – SMART (HMMs, eukaryotes)
  – TIGRFAM (HMMs, prokaryotes)
  – InterPro (integrates other databases)
  – CDD (profiles of conserved domain cores)
Structure and Sequence Conservation

• domain recurrences among 3-D structures reveal that protein structure is more conserved than sequence

• proteins sharing a common fold and have indications of the same function lie in the same superfamily
  – conservation of catalytic or binding sites indicates functional conservation

• various methods were developed to detect homology for sequence-dissimilar proteins that share the same fold
  – e.g. comparison of only structurally equivalent positions reveals that >12% sequence identity typically indicates homology

• homology vs. analogy
Evolution of Domains

- some domains occur in all three kingdoms (archaea, bacteria, eukarya)
- such domains may be essential for cellular processes or just very adaptable to variety of functions
- many domains are eukaryotic only
  - cell-cell communication roles, various types of regulation
  - plants and animals have many independent extracellular domains
  - immune response
- all in all, most modern folds may have arisen from only a few ancestors
  - e.g. domains may be descendants of short polypeptide segments that together were capable of folding and conveying a beneficial function
Exons and 3-D Structure

- once exons were identified, it has been hypothesized that they correspond to the protein structural/functional units

- **Introns-early** hypothesis
  - introns were present in the progenitor of all living organisms and were subsequently lost in bacteria and archaea

- **Introns-late** hypothesis
  - introns are eukaryotic inventions

- conservation of introns typically implies the same functional class
Fold Changes During Domain Evolution

- fact: protein adopting similar folds differ from each other outside of the conserved core

- circular permutation: presumably occurs during gene duplication, fusion, and partial deletion and can make changes to the topology of a protein
  - fusion of the N and C termini and a cleavage at a different location

- protein regions can significantly change in structure with conventional mutation and insertion events
Convergent Evolution

• Example
  – Ser/His/Asp catalytic triad is found in at least 5 different folds, which cannot be considered homologous

• Example
  – thermosylin and mitochondrial processing peptidase: share high similarity in active sites and 3-D structures
  – arrangement and packing of the core secondary structure elements is completely different

• There is growing evidence that a number of apparently different structures may share a common ancestor, convergent evolution appears to be true
Why Domains are Interesting

• domains are correlated with protein function
  – particular arrangements of domains can be used for functional variability

• domains provide insights into evolution of organisms
  – e.g. existence of STY kinases lead to extinction of H kinases in eukaryotes for intracellular signaling

• identification of orthologs and paralogs can have implications in understanding molecular basis of disease
  – model organisms can play a key role in understanding the effects on phenotypes

• practical interests for structural genomics projects
Identifying Domains in Proteins

- the work on structural domains started as early as it could (1970s)
- visual inspection of X-ray structures
- Wertlaufer: domains are regions of the polypeptide chain that form compact globular units, sometimes loosely connected to one another
- $\text{C}_\alpha$-$\text{C}_\alpha$ distance plots were suggested to be useful for protein domain identification (Ooi and Nishikawa; Phillips)
Identifying Domains in Proteins

domain structure of dogfish lactate dehydrogenase
determined using the distance map
Formulating the Problem

• domain: atomic interactions within domains are more extensive than between domains

• data: 3-D protein structure
• domain identification: optimization problem
  – maximize atomic contacts within a domain, but minimize outside of a domain

• a good approach to identifying stable structural units that would self-fold

• how to approach non-contiguous domain problem?
  – alternative problem: how many cuts?
Identifying Number of Cuts

(a)

1-cut

2-cuts

4-cuts

(b)

(c)
First Generation Methods

- based on contact maps (Rossman and Liljas, 1974)

- partitioning methods (Crippen, 1978; Rose, 1979, Janin and Wodak, 1980s)
  - hierarchical
  - structure partitioning

- based on surface area scan (Janin and Wodak, 1981)
  - surface area scan was done between N-terminal part and C-terminal part, where position $i$ divides N and C terminus
  - the interface area plot can help identify domains
First Generation Methods

← Rose, 1979

↑ Crippen, 1978

↑ Wodak and Janin, 1981 →
Second Generation Methods

• methods from other disciplines are adopted
  – graph theory
  – physics
  – statistics

• more general
• computationally more efficient
### STRUDL

- **STRUctural Domain Limits**
- proposed by Wernisch, Hunting and Wodak; Proteins 1999
- starts from the premise that partitioning of the residues is an optimization problem in which contact area between domains should be minimized
  - recursive partitioning
  - highly heuristic, uses Kernighan-Lin algorithm for partitioning while accepting/rejecting partitions is based on a series of (fitted) parameters
- **Novel aspects**
  - heuristic graph partitioning
  - atom contact area is used to measure domain interactions
Contact Area

\[ R = R_{vdW} + 1.4 \text{Å} \]
Definition of Contact Area

- each atom represented by its accessible sphere of radius R (see previous slide); 1.4Å is added to the van der Waals radius to accommodate 1 water molecule
- Voronoi cell
  - the smallest polyhedron created by the set of planes perpendicular to the lines connecting atoms center to those of its neighbors, and positioned at the intersection of the accessible sphere of the atom and those of its neighbors

- **Contact area** with the neighboring atom is defined as the area of polygonal faces obtained as intersections of Voronoi planes and spheres of radii R
  - symmetric measure
Contact Ares for Partitions

Contact area between two groups of atoms, all individual contact areas are simply added.

Contact map can be created based on pairwise contact areas!
STRUDL Algorithm

1. Start with a subset $U$ consisting of 1 randomly chose residue (complement of $U$ is $V$)
   1. $U$ has size of $k$ (predetermined);
   2. $1 \leq k \leq N/2$

2. Perform “exchanges”, moving 1 residue from $V$ to $U$ and one from $U$ to $V$ so as to maximally decrease contact between two sets
   1. flag moved residues
   2. save contacts for each step and corresponding partition

3. Find best of all recorded partitions (“optimum”)

4. Repeat 2 and 3 until no further reduction in contact map is observed
Partition Method

1. select random residue as seed for $U$
2. increase $U$ to $|U| = k$ by adding residues
3. choose best $u \in U$ and $v \in V$ for exchange
4. move $u$ to $V$ and $v$ to $U$
5. mark $v$, record $U, V$
6. Any unmarked $v \in V$?
   - YES
8. select minimum area partition $U_{\text{min}}, V_{\text{min}}$ from records
9. delete marks
10. $U \leftarrow U_{\text{min}}, V \leftarrow V_{\text{min}}$
11. Any improvement?
   - YES
12. repeat 6 times
13. optimal partition $U, V$

Examples:
- select $u \in U$ and $v \in V$ with maximum decrease in contact
- contact area $c_1$
- select $u \in U$ and $v \in V$ with maximum decrease in contact
- contact area $c_2$
- ...
- contact area $c_n$
Behavior of the Algorithm
Minimum Contact Profile

$c_{\text{dens}}(U, V) = \frac{c(U, V)}{|U||V|}$

Note: short domains, where $k < 0.05N$, not allowed
Accepting/Rejecting Domains

- highly heuristic
- based on 9 different parameters
  - normalized contact area
  - sum of interresidue contacts
  - compactness
  - average interresidue contact
  - average slope of the profile
  - proportion of residues with positive slope
  - depth of the contact density profile at the minimum
  - total number of residues in the protein and the number of residues to the left of the minimum in the contact density partition
Data Sets

- 787 proteins from PDB as the representatives of the fold families in the CATH
  - 524 single domains
  - 263 multi-domain proteins

- 192 structures to fit additional parameters
  - 132 single domain structures
  - 60 multi-domain structures

- initial parameters adjusted on the training set using cross-validation
Tests

- each parameter was individually assessed
- best performing parameters were assessed in pairs etc.
- only two parameters were finally used, based on classification error
  - single vs. multi

Threshold optimization for a pair of parameters (mean burial vs. the contact area ratio)
Results

- Voronoi contact area worked well compared to the inter-atomic contacts (the number of atom pairs within 8Å)
- Correct assignments
  - same number of domains as in CATH with >85% residues in correct domains
- 635 chains (80.7%) are classified correctly
- types of errors
  - undercut
  - overcut
  - debatable assignments
Errors?
Analysis