

Identifying (Structural) Domains

Structural Bioinformatics (Part 4)

Presenter
Jose Lugo-Martinez
Lab Meeting

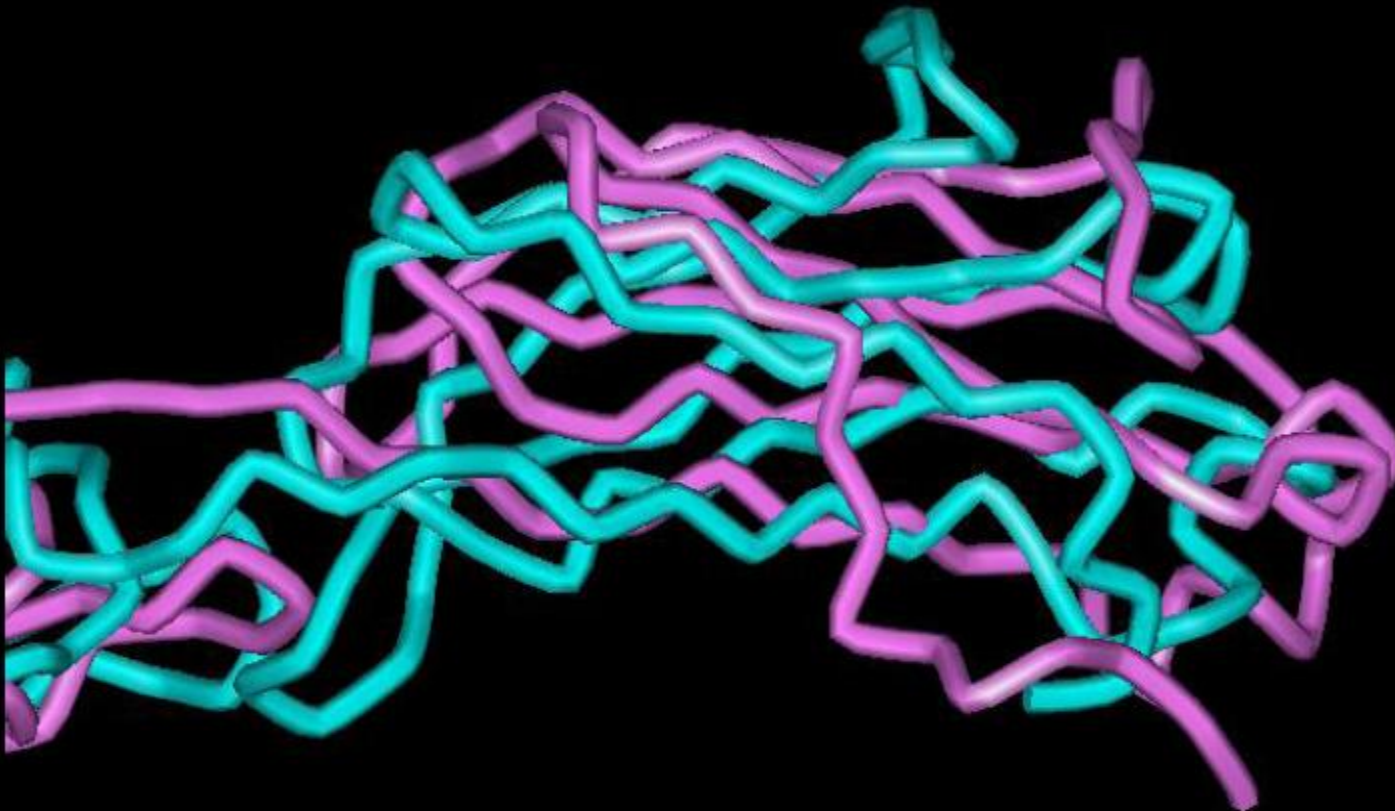
May 3, 2011





INDIANA UNIVERSITY

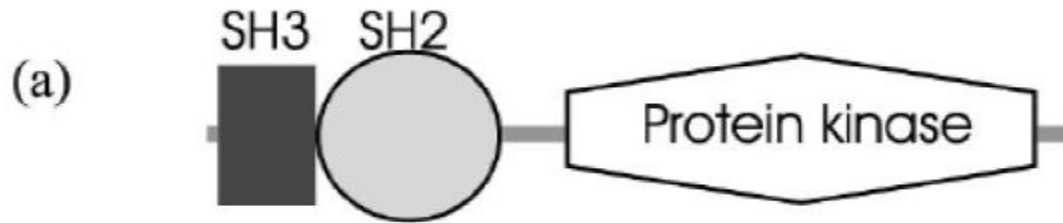
Immunoglobulins **1MCO:L** vs **3MCG:2** Seq. identity = 99%, RMSD = 5.5Å





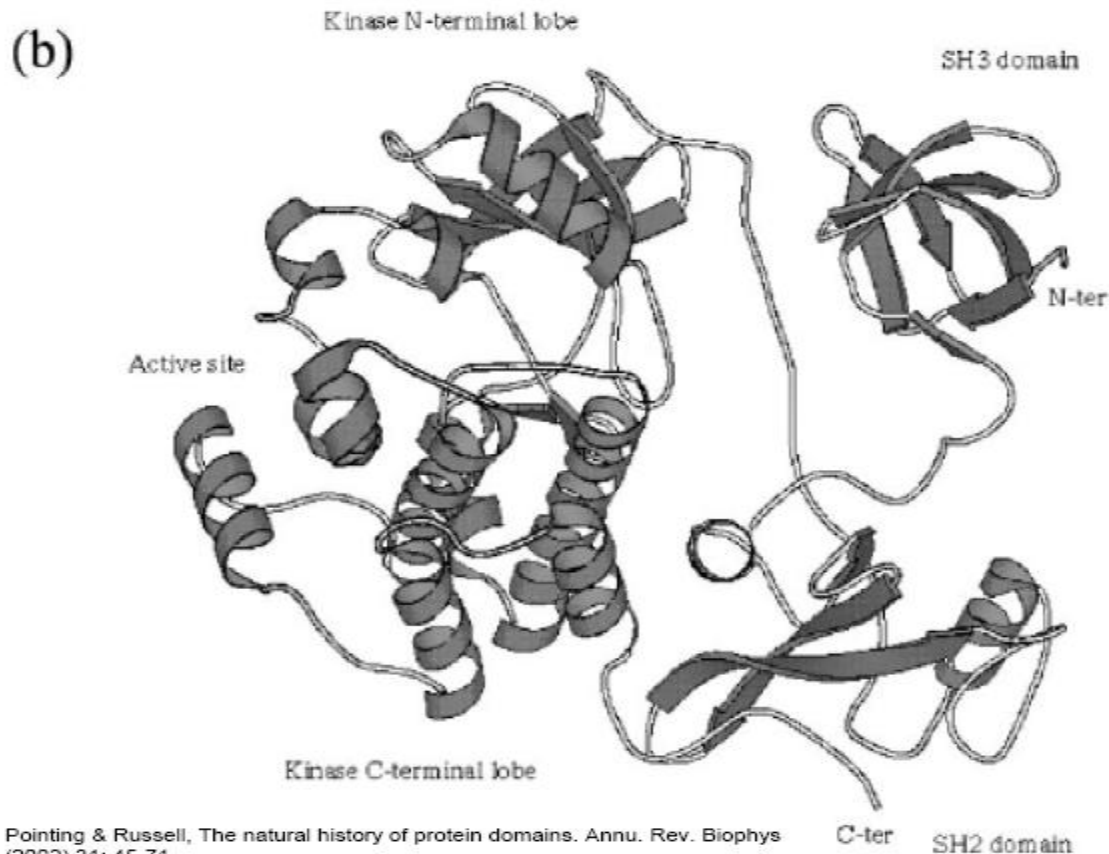
Structural Domains

- At first, domain concept used to describe distinct regions of 3D structures
- Nowadays, usually defined as: *spatially distinct structures with a well-defined hydrophobic core that could fold and function in isolation*
 - stability
 - compactness
 - presence of hydrophobic core
 - ability to fold independently



An Example

Part of the structure of hematopoietic cell kinase (1qcf)





Why Domains?

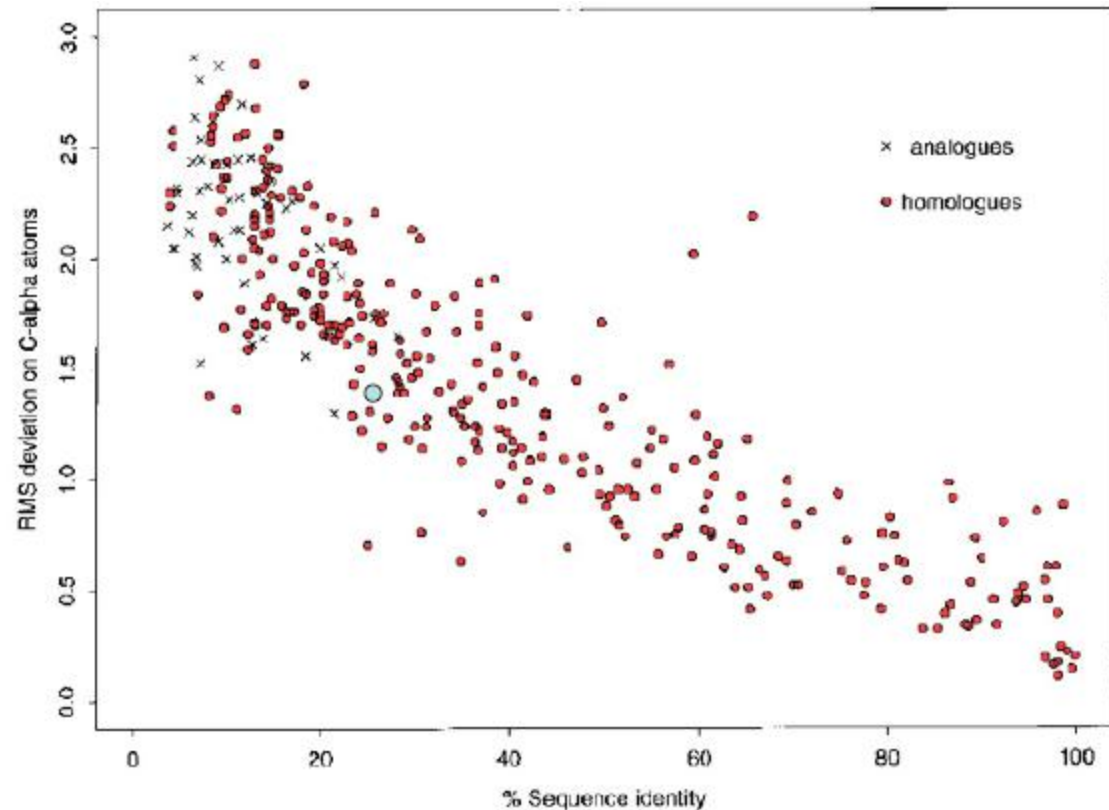
- Protein structure can elucidate its function and evolutionary history
 - domains are correlated with protein function
 - particular arrangements can be used for functional variability
 - domains provide insights into evolution of organisms
 - e.g. existence of STY kinases lead to extinction of H kinases in eukaryotes for intracellular signaling
- Databases that map protein structure space at domain level
 - SCOP, CATH



Homology and Structural Similarity

(Russell et. al., 1997)

Proteins that diverge in evolution maintain their global fold!



homology - occurrence of identical or similar features as a result of common ancestry

analogy - apparently similar yet phylogenetically independent (evolved separately)



Question: How to Identify Domains?

- Given a protein sequence and/or 3D structure
 - How many domains are there?
 - Where are the domain boundaries?
- Still not yet completely resolved
 - e.g. ~80% agreement between methods
- Need both sequence- and structured-based approaches
 - From Sequence
 - sequence data >>> structure data
 - From Structure
 - protein structure more conserve than sequence
 - studies developed to detect homology for sequence-dissimilar proteins that share the same fold (e.g. comparison of only structurally equivalent positions) reveal that >12% sequence identity typically indicates homology



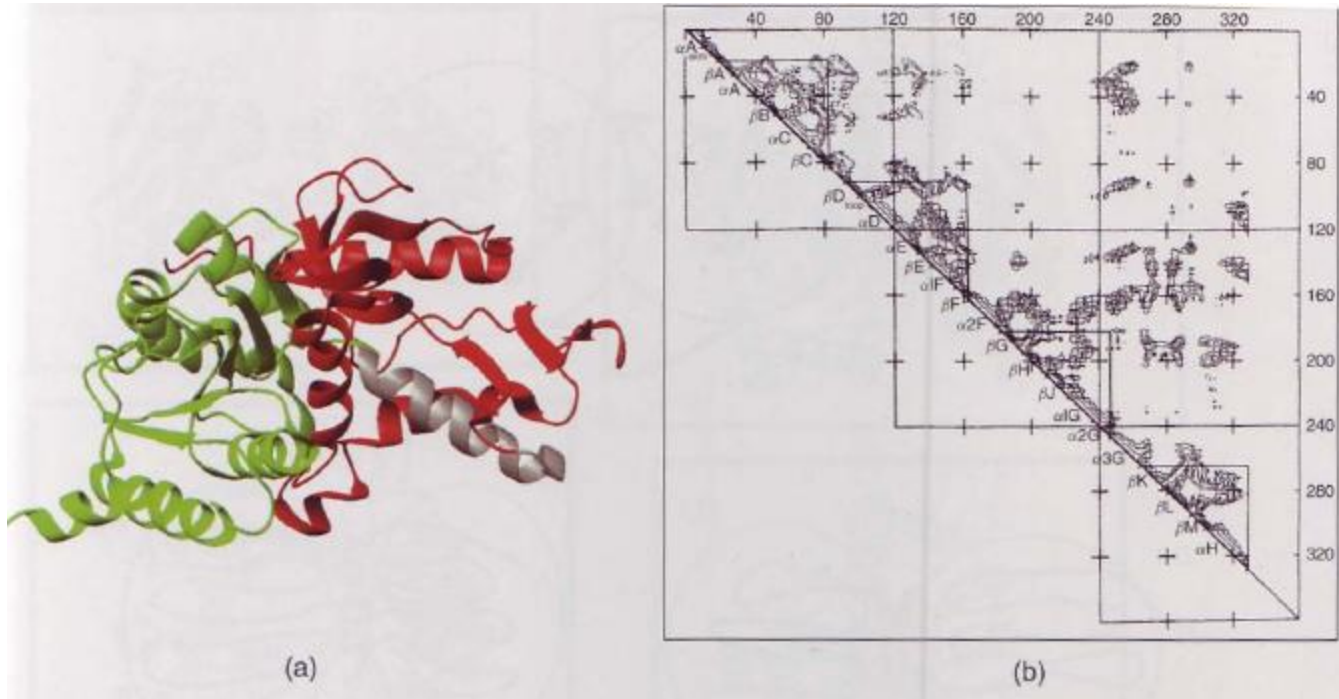
Identifying Domains from Structure

- Started during 1970s
- Two approaches:
 - Manual inspection
 - relies on human experts either exclusively or partially (SCOP)
 - nearly always superior to its counterpart
 - structure data throughput >>> human experts
 - Computational methods (top-down and/or bottom-up)
 - try to identify domains in an automated way by encoding basic rules and assumptions
 - fast and consistent
 - cannot reconcile contradictions within structural data due to biological complexity (mainly at domain boundaries)



INDIANA UNIVERSITY

First Manual Survey of Structural Domains



domain structure of dogfish lactate dehydrogenase

Ooi and Nishikawa; Phillips suggested that C α -C α distance plots were useful for protein domain identification



Problem: Identifying Structural Domains

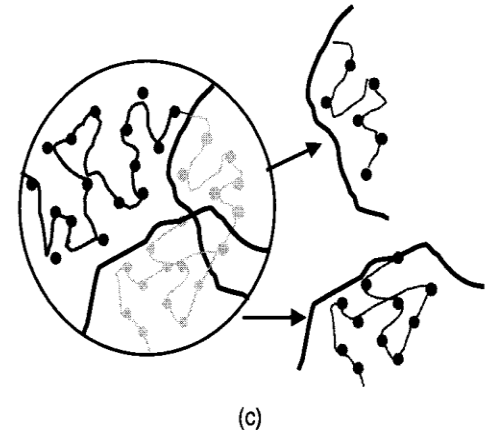
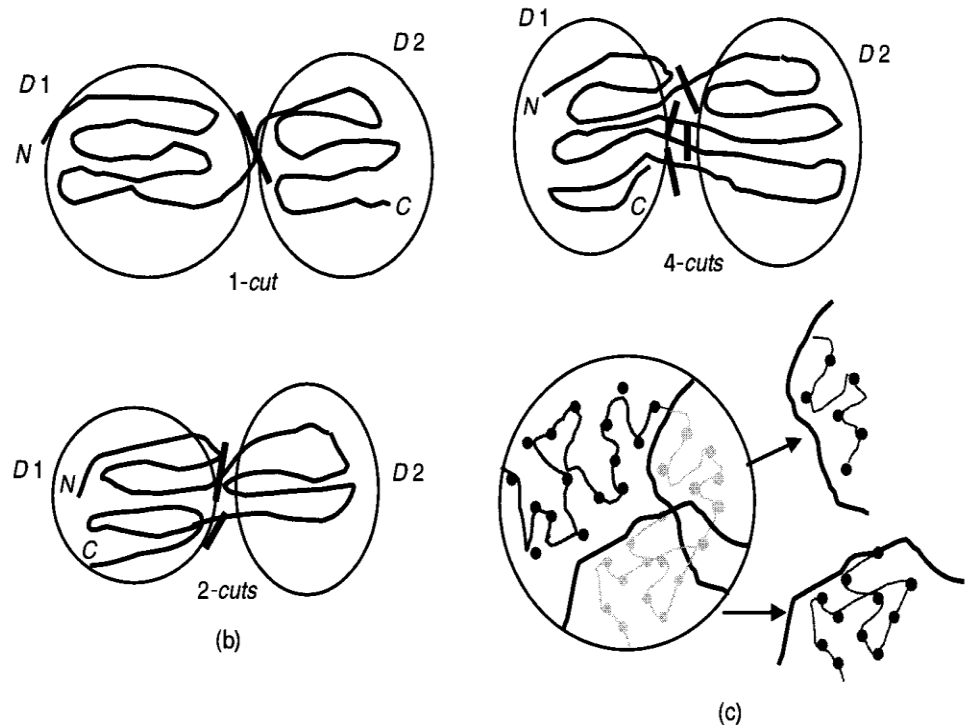
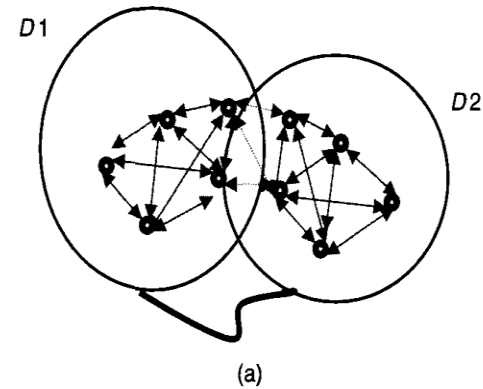
- **Domains “Key Property”:** *atomic interactions within domains are more extensive than between domains*
- Domain identification becomes an optimization problem!
 - Given 3D protein structure
 - **Idea:** look for groups of residues with a maximum number of atomic contacts within a domain, but a minimum number of contacts between domains
- How to approach non-contiguous domain problem?
 - stretches of residues that are distant on the polypeptide chain but found in close proximity in the folded structure
 - may arise as a result of gene insertion events or domain swapping
 - alternative problem: how many cuts? (Next slide)



How many cuts?

Domains composed of one or more chain segments

Any domain assignment method must be able to cut chain as many times as necessary



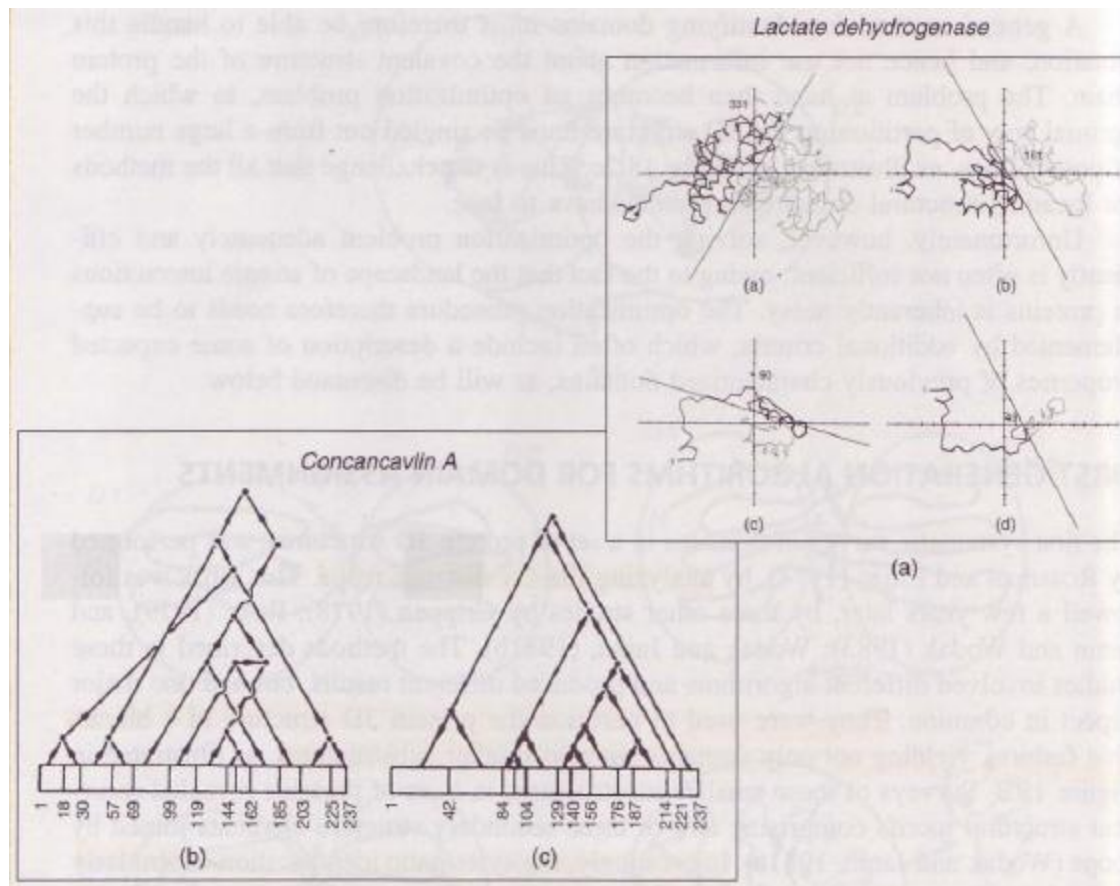


Structural Domains: Automatic Analysis

- First generation methods (1974-1995)
 - Rossman and Liljas (1974): 1st algorithmic survey of domains
 - based on C α -C α distance maps
 - Crippen (1978); Rose (1979); Janin and Wodak (1980s)
 - all studies share one aspect - partitioning of 3D structure into a hierarchical fashion, yielding domains and smaller substructures which revealed recurrent structural motifs comprising 2 or 3 SSE joined by loops
 - Lack of separation between training and testing data



First Generation Methods



→ Rose, 1979

Crippen, 1978

Wodak and Janin, 1981



Structural Domains: Automatic Analysis (2)

- Second generation methods (1994-2007*)
 - increase availability of structure data enabled performance comparisons against a representative dataset
 - use of techniques from other disciplines
 - Graph theory (Kernighan-Lin graph heuristics), physics (rigid body oscillation), statistics (Ising model)
 - more general (able to deal with non-contiguous domains)
 - computationally more efficient
- Second generation+ methods
 - if incorporated cross-validation in testing



Identifying Structural Domains: STRUDL

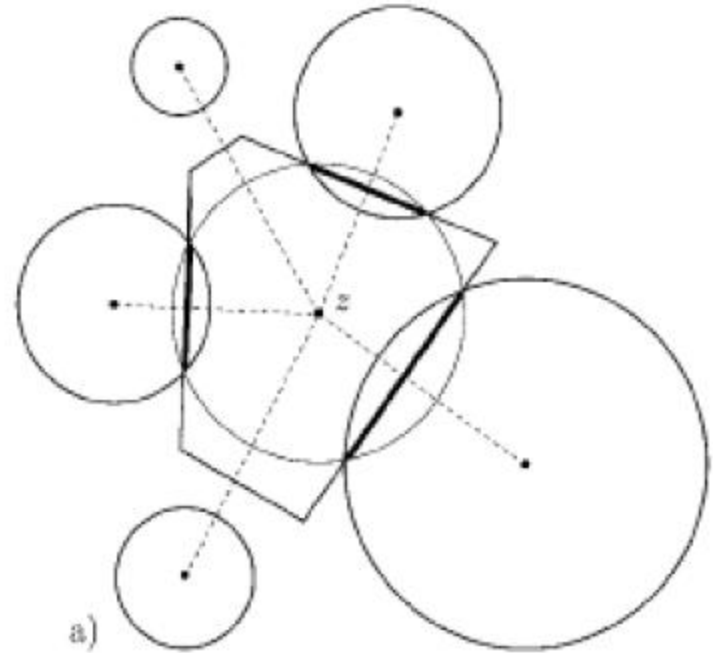
- STRUDL (STRuctural Domain Limits)
- Proposed by Wernisch, Hunting and Wodak (1999)
 - protein structure as 3D graph of interacting residues
 - partitioning graph into sets of residues such that interactions between sets in minimum (i.e. contact area between domains should be minimized)
 - NP-hard
- Novel aspects:
 - modified Kernighan-Lin heuristic for graph partitioning
 - atom contact area is used to measure domain interactions
 - authors claimed it is more robust than counting atomic contacts due to lower sensitivity to distance thresholds

Contact Area

Area of intersection of the van der Waals sphere around each atom and the faces of its weighted Voronoi polyhedron

Contact area with the neighboring atom: area of polygonal faces obtained as intersections of Voronoi planes and spheres of radii R

- symmetric measure



$$R = R_{\text{vdW}} + 1.4\text{\AA}$$

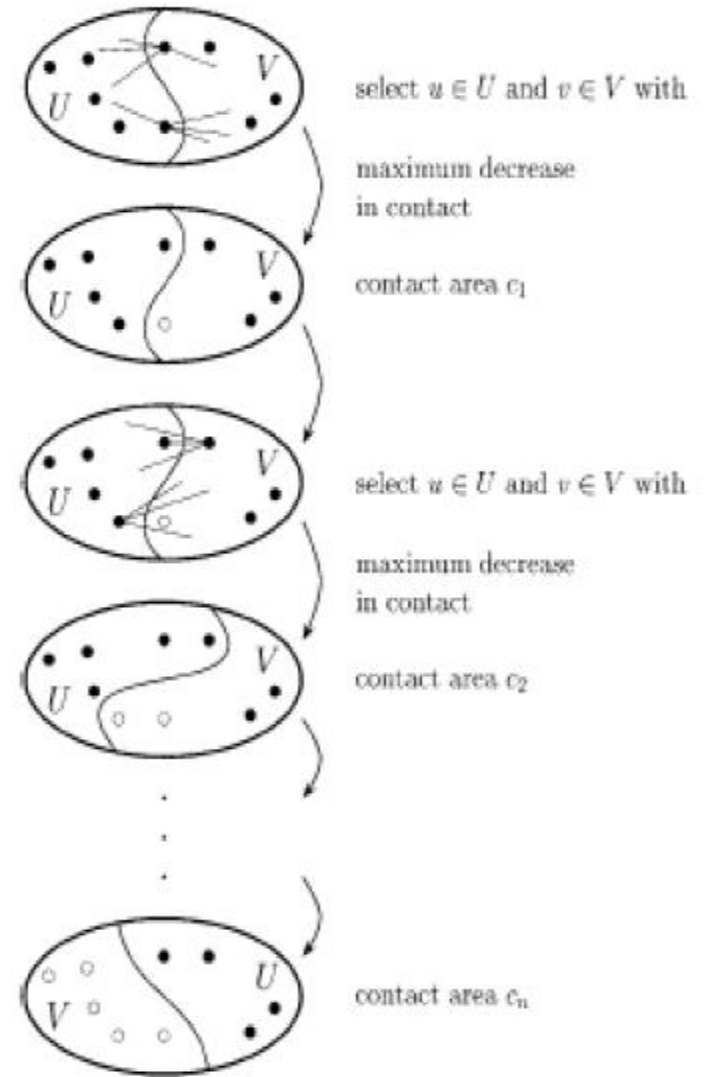
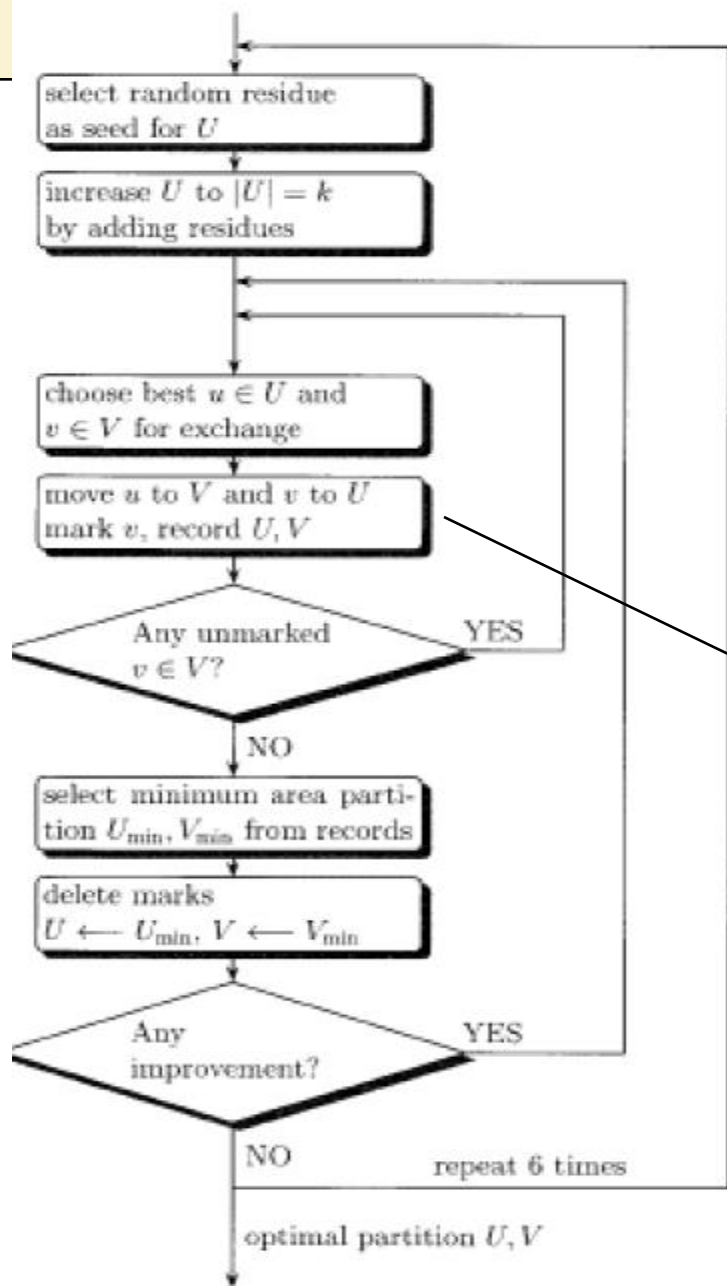
Voronoi cell - the smallest polyhedron created by the set of planes perpendicular to the lines connecting atoms center to those of its neighbors, and positioned at the intersection of the accessible sphere of the atom and those of its neighbors



STRUDL Algorithm

- Start with a subset U consisting of 1 randomly chose residue (complement of U is V)
 - U has size of k (predetermined);
 - $1 \leq k \leq N/2$
- Perform “exchanges”, moving 1 residue from V to U and one from U to V so as to maximally decrease contact between two sets (if not possible, minimal increase CA between U and V)
 - flag moved residues
 - save contacts for each step and corresponding partition
- Find best of all recorded partitions (“optimum”)
- Repeat 2 and 3 until no further reduction in contact map is observed

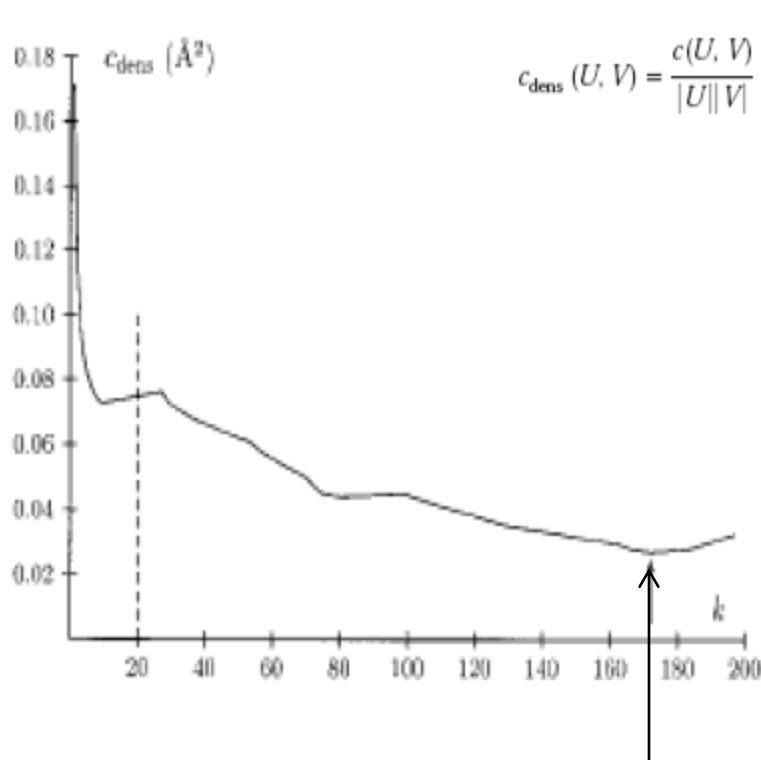
Partitioning Method



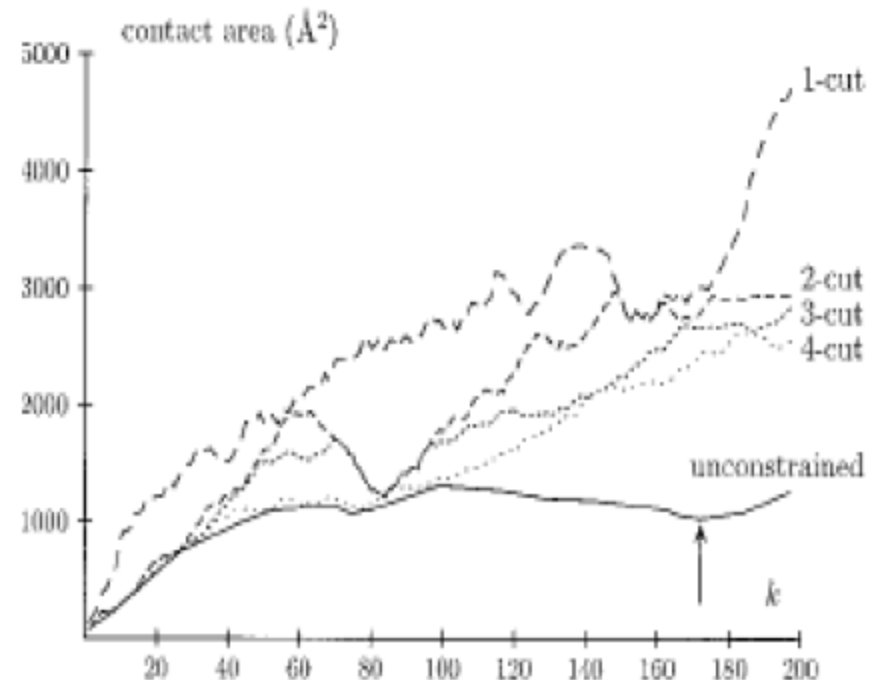


Minimum Contact Profile

P-hydroxy-benzoate hydroxylase, 394-residue protein with 2 discontinuous domains



global minimum
at $k = 172$





Observed Limitations

- Performance comparisons of the algorithms
 - 70-80% of available structures can be solved correctly
 - remaining ones show complex, multi-domain structures
- Typical errors
 - # of assigned domains is over-predicted (overcut)
 - # of assigned domains is under-predicted (undercut)
 - sometimes debatable assignments



Identifying Domains from Sequence

- Given a new sequence:
 - “Easy”
 - if homology can be identified by sequence alignment (to domain databases/profiles, e.g. Pfam)
 - “Not-so-easy”
 - if homology cannot be reliably identified
- Another challenge: *non-contiguous domains!!!*



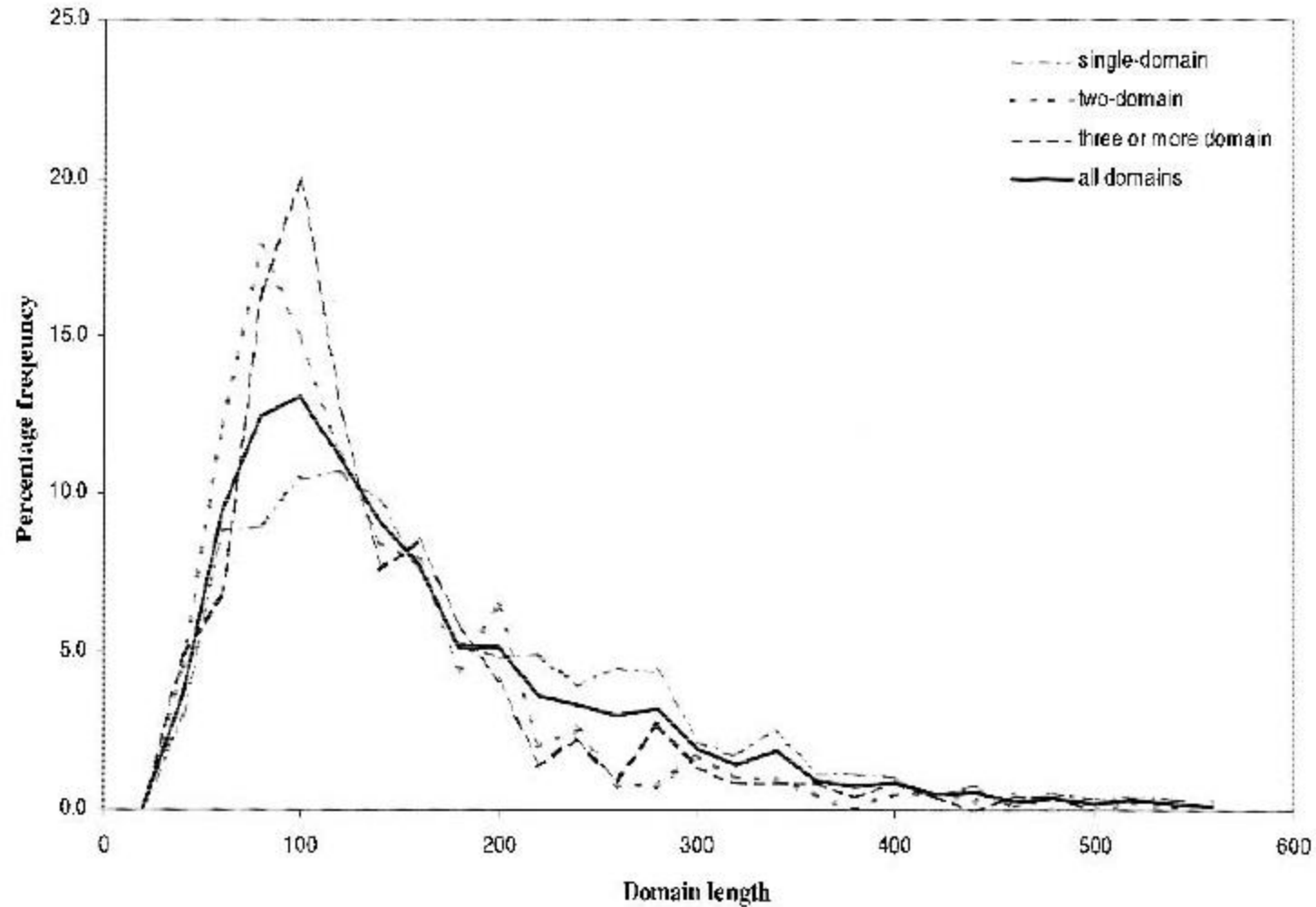
Domain Prediction from Sequence Info

- Based purely on domain and protein sizes
 - no sequence or structure information used
 - predictions based exclusively on size of protein
 - e.g. Domain Guess by Size (Wheelan, Marchler-Bauer, 2000)
- Based on single sequence information
 - use of aa propensities for linker regions associated with domain boundaries (i.e. linker regions distinct to rest of protein)
 - e.g. CHOPnet (Liu and Rost, 2004)
- Incorporate structure prediction from sequence
 - add information content inherent in structural data predictions (either secondary or tertiary structure)
 - DomSSEA (Marsden, McGuffin, and Jones, 2002), SnapDRAGON (George and Heringa, 2002)



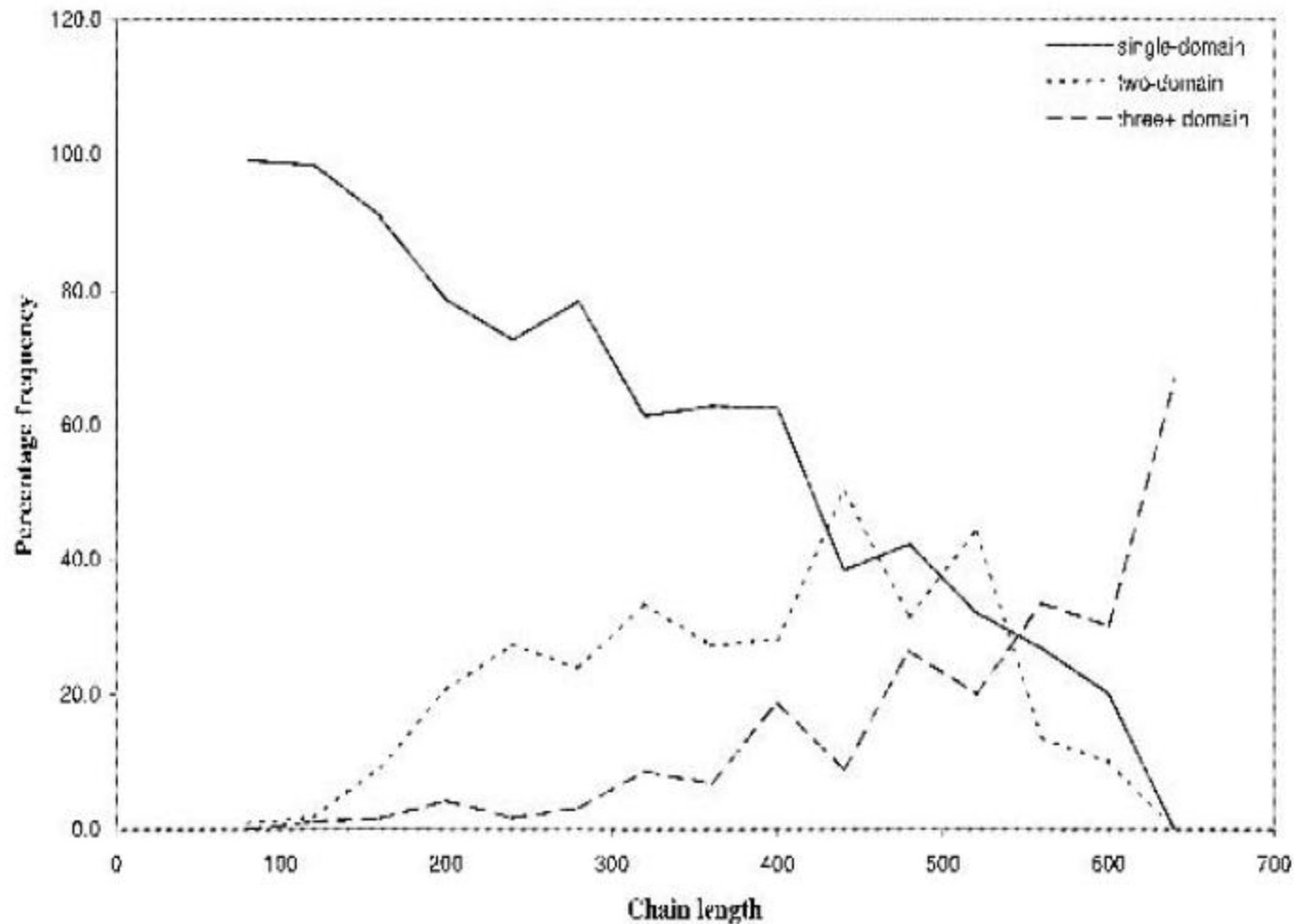
INDIANA UNIVERSITY

Domain Guess by Size (1)





Domain Guess by Size (2)



57% accuracy for identifying domain boundaries



Single Sequence Info-Based Methods

- **CHOPnet** (Liu and Rost, 2004)
 - uses aa composition, many predicted features, sequence alignments and combines them using neural networks
 - 69% accuracy for # of domains
 - 50% cut prediction (for 2 domains)



Adding Predicted Structure Methods (1)

- **DomSSEA** (Secondary Structure Element Alignment)
 - predict secondary structure on a new (query) protein
 - align query sequence of predicted secondary structures to all sequences in the database (of domains)
 - collect top hit from the pair with the highest score
 - make a cut within a coil region
 - 73.3% acc. for # of domains
 - 50% overall cut prediction (domain boundaries) for 2 domains and 24% for multi-domain proteins



Adding Predicted Structure Methods (2)

- **SnapDRAGON** (George and Heringa, 2002)
 - *ab initio* protein folding method: DRAGON
 - predicts secondary structure, makes sequence alignment, then folds each chain based on conserved hydrophobicity
 - each folded chain is evaluated for domain boundaries
 - large number of structures are evaluated for domain consistency
 - 72% acc. for # of domains
 - 63.9% for domain boundary predictions in multidomain proteins



INDIANA UNIVERSITY

Thank You!

QUESTIONS?