Protein Domains

I619: Structural Bioinformatics

February 18, 2008

Last time: Structure Validation

- A structure can (and often does) have mistakes
- A poor structure will lead to poor models of mechanism or relationship
- Unusual parts of a structure may indicate something important (or an error)

Known bad structures

- Azobacter ferredoxin (wrong space group)
- Zn-metallothionein (mistraced chain)
- Alpha bungarotoxin (poor stereochemistry)
- Yeast enolase (mistraced chain)
- Ras P21 oncogene (mistraced chain)
- Gene V protein (poor stereochemistry)

Structure Validation

- Assess experimental fit
  - look at Resolution, R-Factor or RMSD (be careful with RMSD!)
- Assess correctness of overall fold
  - look at disposition of hydrophobic residues
- Assess structure quality
  - packing
  - stereochemistry
  - contacts...

A Good Protein Structure..

X-ray structure

- R = 0.59 random chain
- R = 0.45 initial structure
- R = 0.35 getting there
- R = 0.25 typical protein
- R = 0.15 best case
- R = 0.05 small molecule

NMR structure

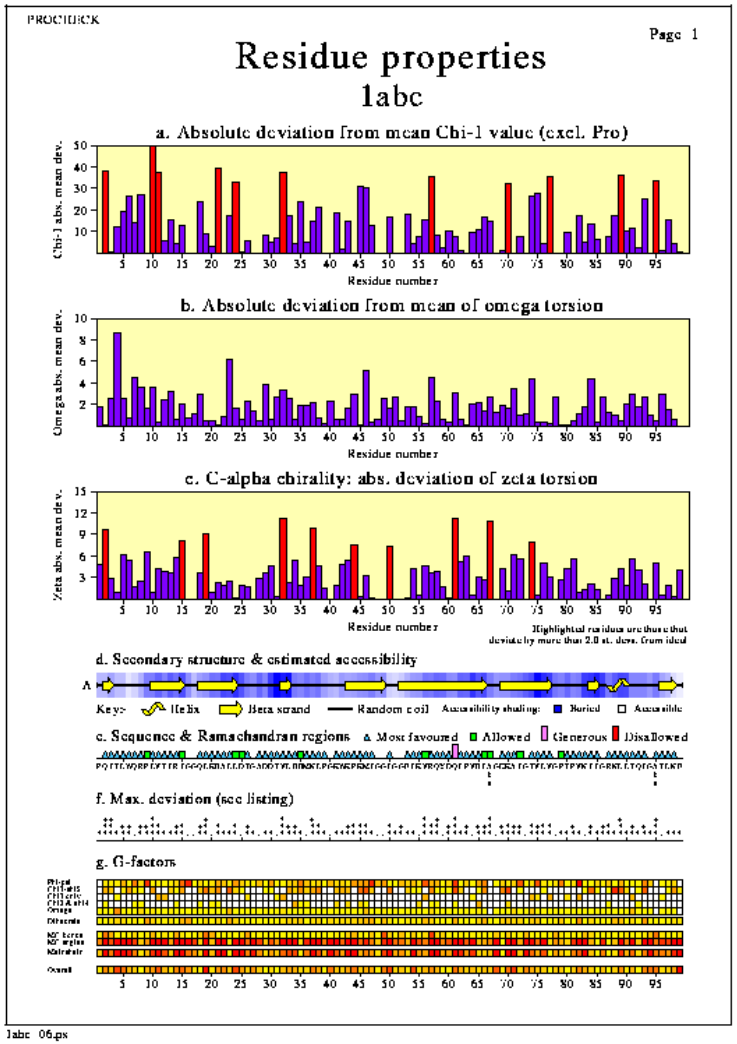
- RMSD = 4 Å random
- RMSD = 2 Å initial fit
- RMSD = 1.5 Å OK
- RMSD = 0.8 Å typical
- RMSD = 0.4 Å best case
- RMSD = 0.2 Å dream on

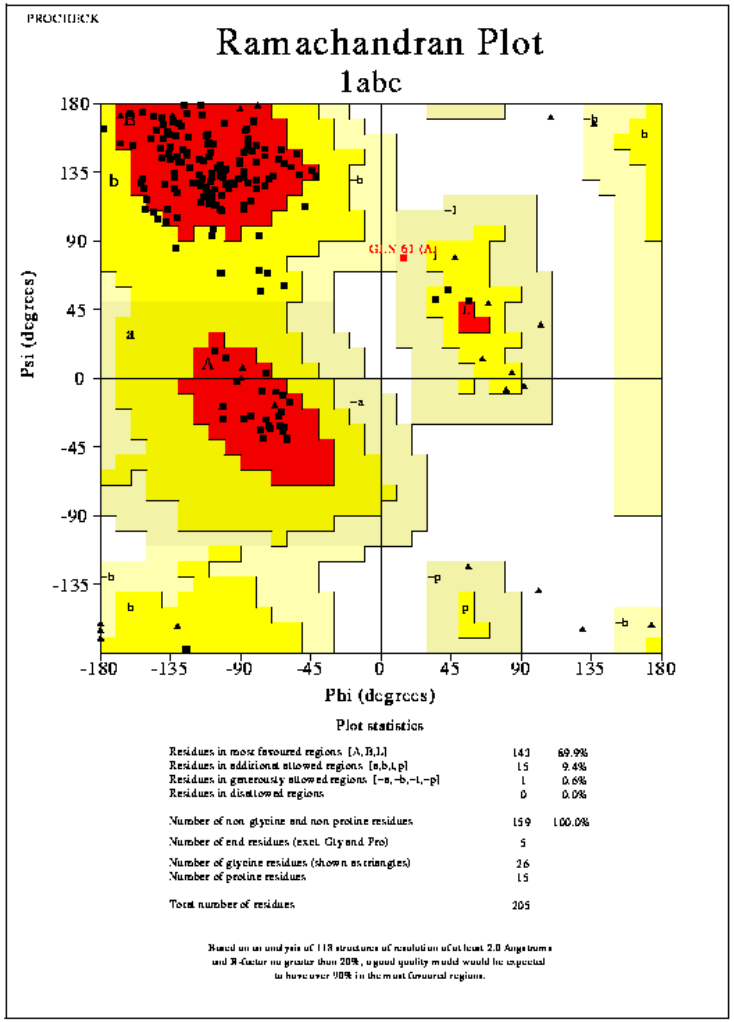
Structure Validation Servers/Programs

- WHAT IF
  - http://swift.cmbi.kun.nl/WIWWWI/
- Verify3D
  - http://www.doe-mbi.ucla.edu/Services/Verify_3D/
- VADAR
  - http://redpoll.pharmacy.ualberta.ca

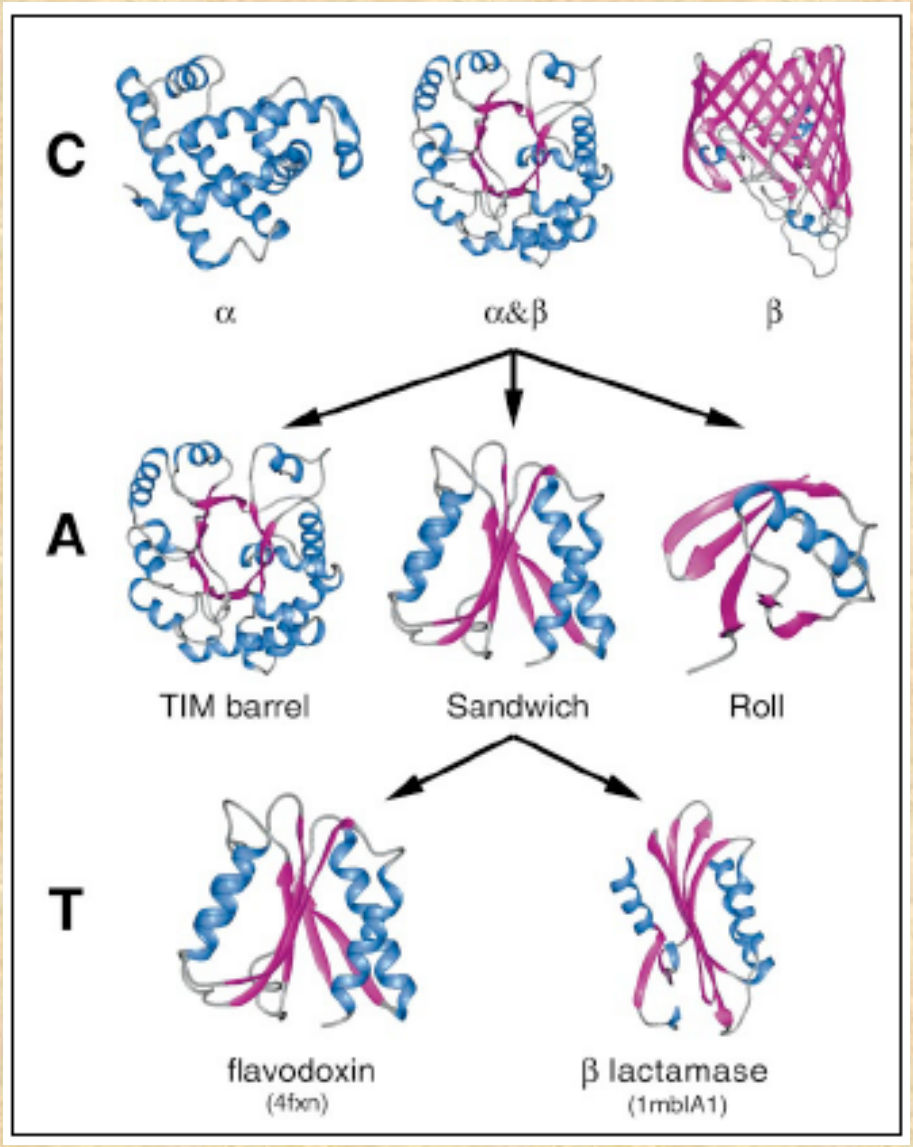
- PROCHECK
  - http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html
- VADAR
  - http://www.pence.ca/software/vadar/latest/vadar.html
- DSSP
  - http://www.cmbi.kun.nl/gv/dssp/

Procheck


PROCHECK
Page 1
Residue properties
1abc
a. Absolute deviation from mean Chi-1 value (excl. Pro)
Chi-1 abs. mean dev.
Residue number
b. Absolute deviation from mean of omega torsion
Omega abs. mean dev.
Residue number
c. C-alpha chirality: abs. deviation of zeta torsion
Zeta abs. mean dev.
Residue number
Highlighted residues are those that deviate by more than 2.0 st. devs. from ideal
d. Secondary structure & estimated accessibility
Keys: Helix Beta strand Random coil Accessibility shading: Buried Accessible
e. Sequence & Ramachandran regions A Most favoured Allowed Generous Disallowed
f. Max. deviation (see listing)
g. G-factors
1abc 06.ps

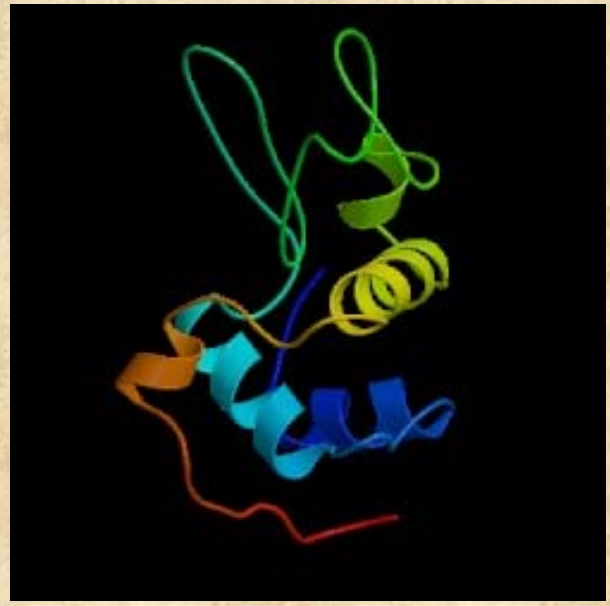

PROCHECK
Ramachandran Plot
1abc
Psi (degrees)
Phi (degrees)
GLN 61 (A)
Plot statistics
Residues in most favoured regions [A,B,L] 143 89.9%
Residues in additional allowed regions [a,b,l,p] 15 9.4%
Residues in generously allowed regions [~a,~b,~l,~p] 1 0.6%
Residues in disallowed regions 0 0.0%
Number of non-glycine and non-proline residues 159 100.0%
Number of end-residues (excl. Gly and Pro) 5
Number of glycine residues (shown as triangles) 26
Number of proline residues 15
Total number of residues 205
Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.
1abc 01.ps

CATH Classification



Domains in 3-D Structures

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

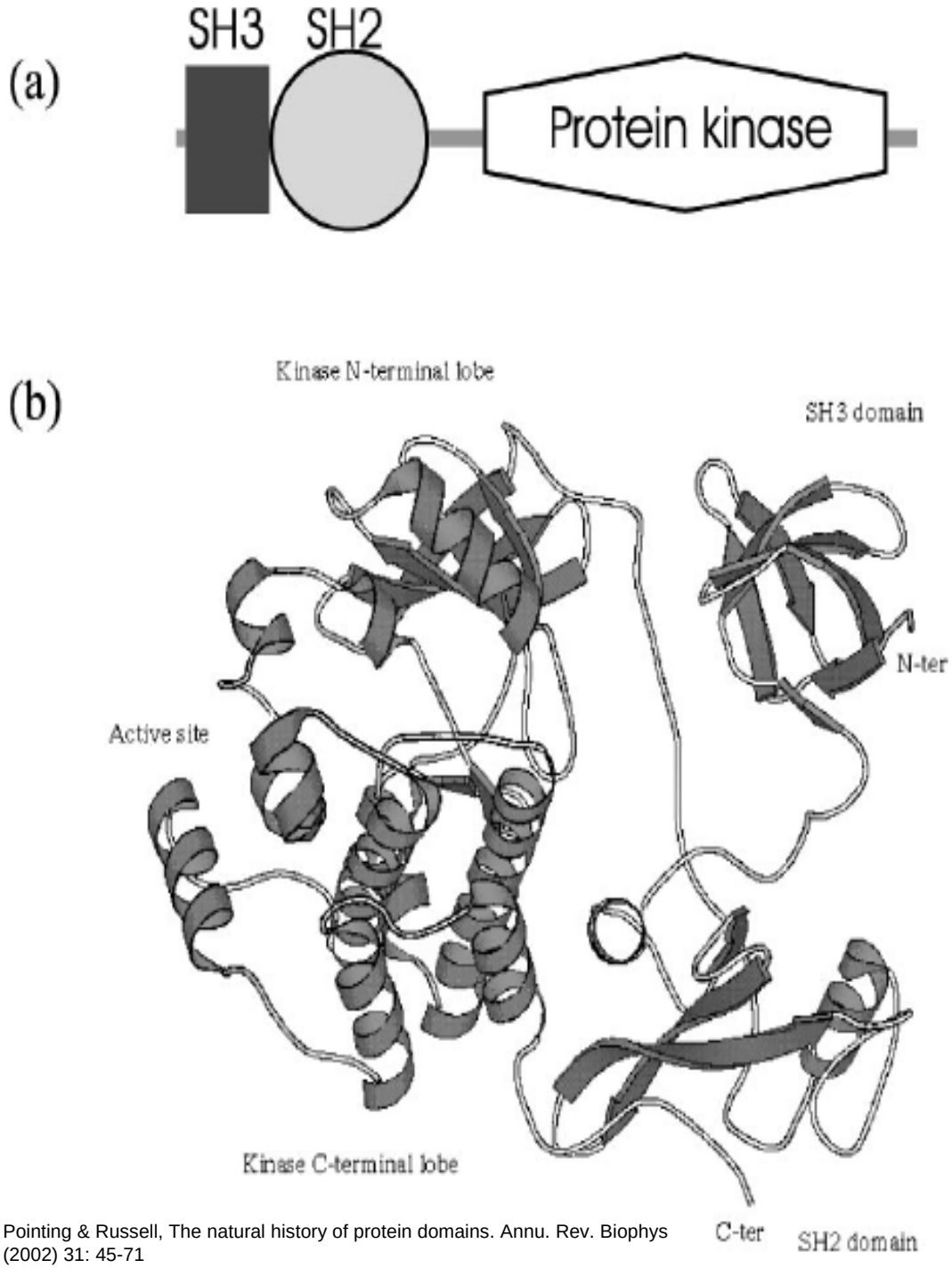


PDB ID 1IY3

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



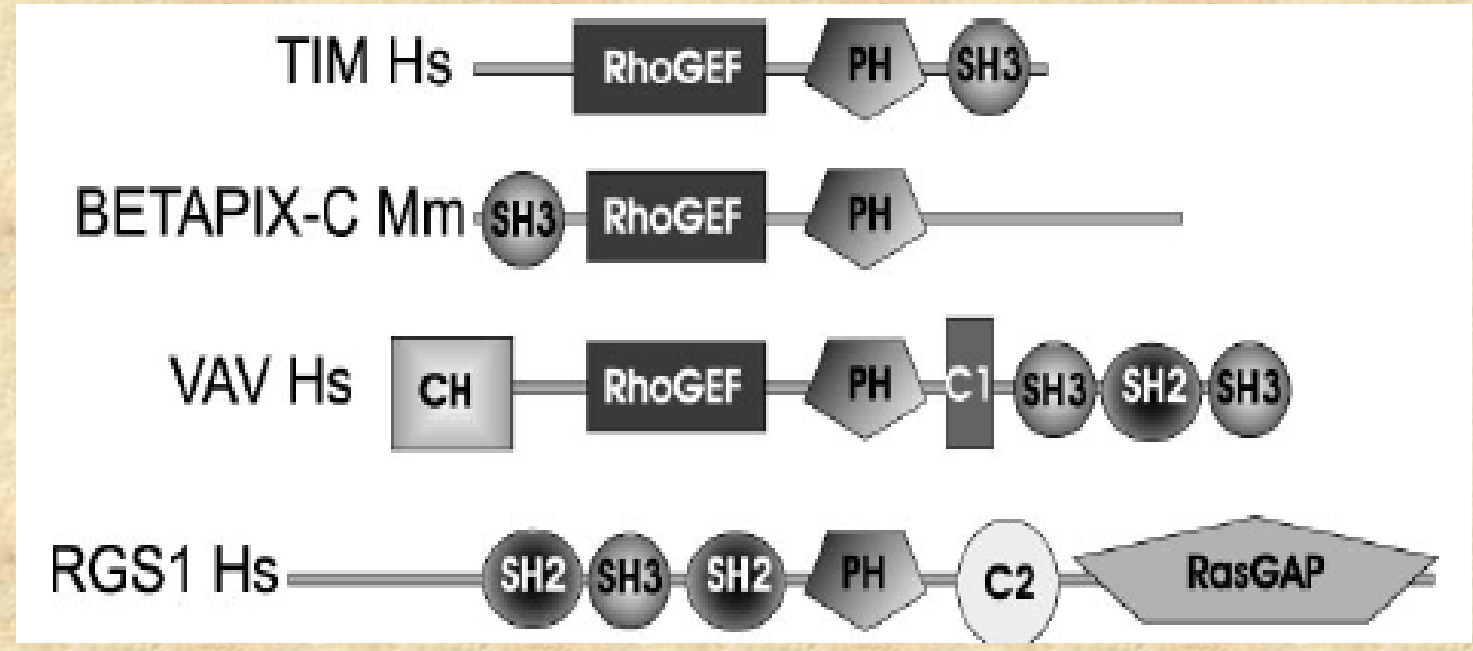
Pointing & Russell, The natural history of protein domains. Annu. Rev. Biophys (2002) 31: 45-71

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

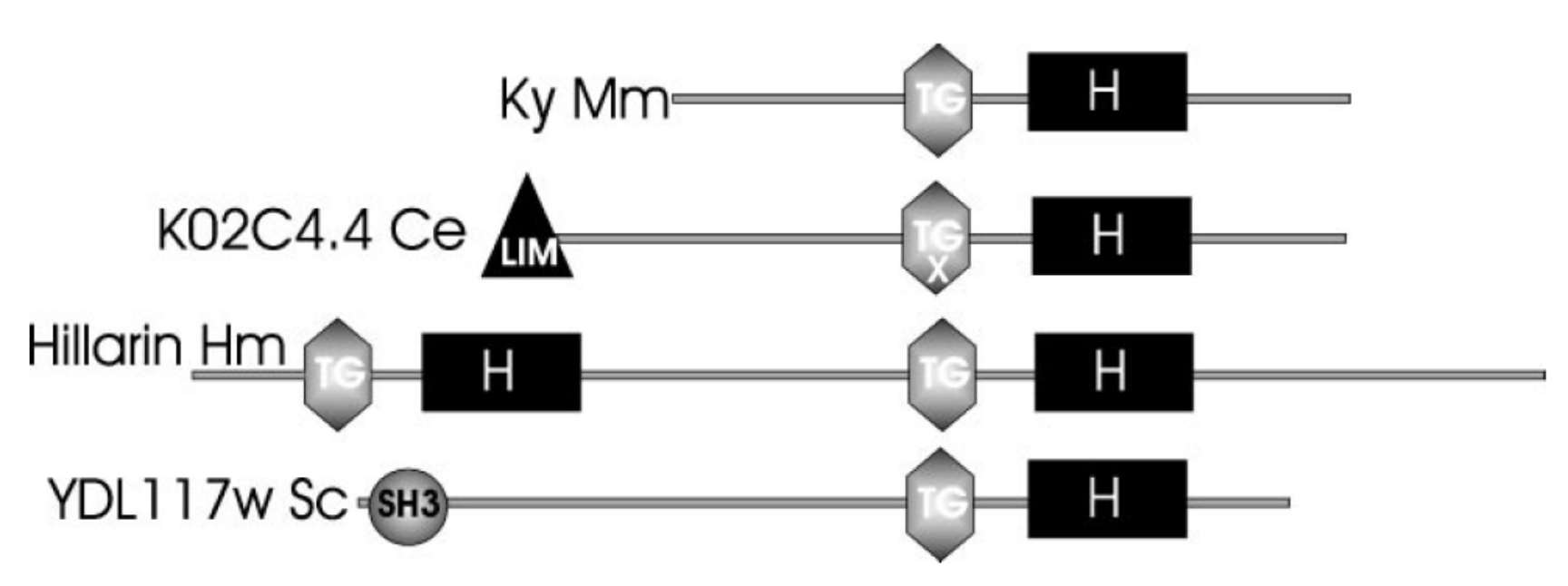
- the arrangement of different domain types in proteins may cause problems in sequence analysis



Pointing & Russell, The natural history of protein domains. Annu. Rev. Biophys (2002) 31: 45-71

Domains and Homology

- gene duplication has been a major evolutionary process in the acquisition of novel function
- domains are orthologous, or proteins?



Pointing & Russell, The natural history of protein domains. Annu. Rev. Biophys (2002) 31: 45-71

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

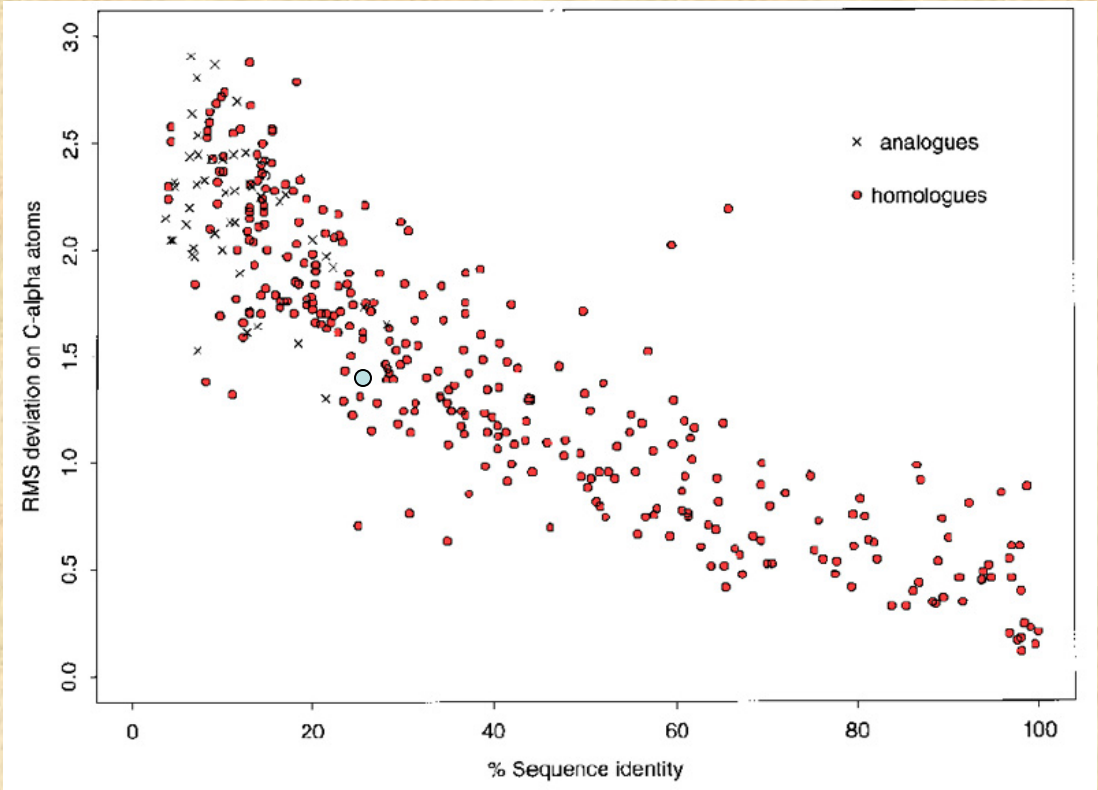
- Libraries
  - Pfam (HMMs)
  - PROSITE (patterns and profiles)
  - SMART (HMMs, eukaryotes)
  - TIGRFAM (HMMs, prokaryotes)
  - InterPro (integrates other databases)
  - CDD (profiles of conserved domain cores)

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 (convergent folds)

Homology and Structural Similarity

Proteins that diverge in evolution maintain their global fold!



Russell *et al.* (1997) *J Mol Biol* **269**: 423-439

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; still convergent evolution appears to be true

Why Domains are Interesting

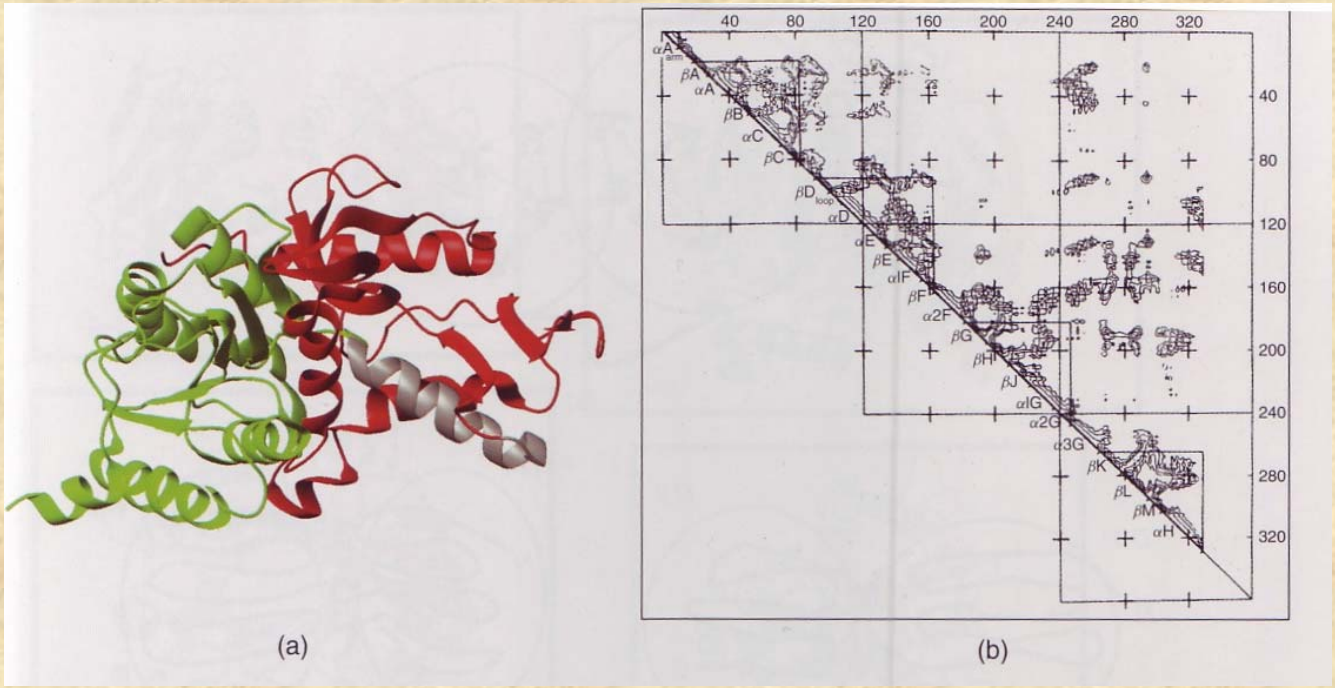
- domains are correlated with protein function
  - particular arrangements of domains can be used for functional variability
- domains provide insights into evolution of organisms
  - e.g. existence of STY kinases lead to extinction of H kinases in eukaryotes for intracellular signaling
- identification 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

- Cα-Cα distance plots were suggested to be useful for protein domain identification (Ooi and Nishikawa; Phillips)

Identifying Domains in Proteins

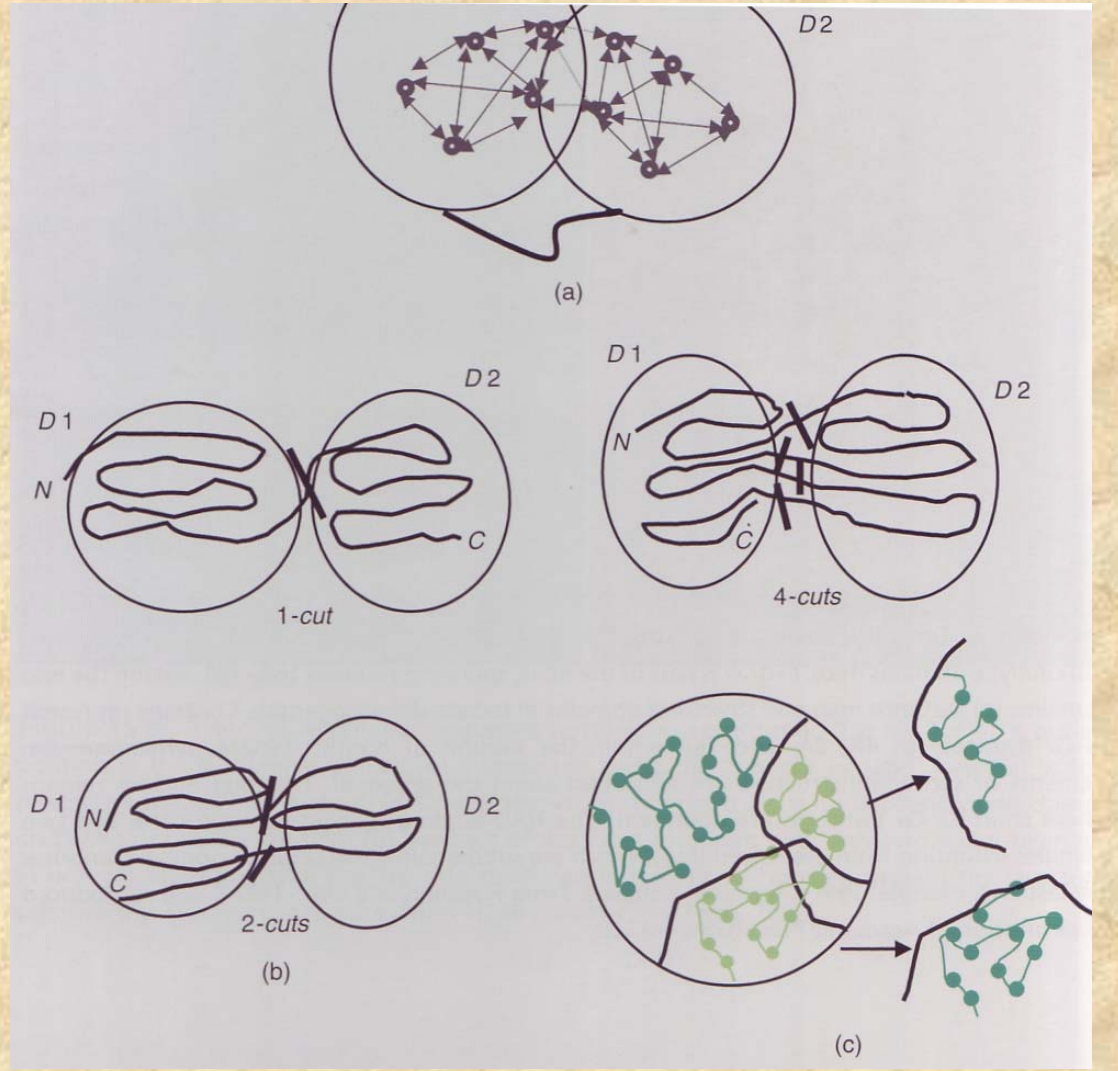


domain structure of dogfish lactate dehydrogenase determined using the distance map

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?

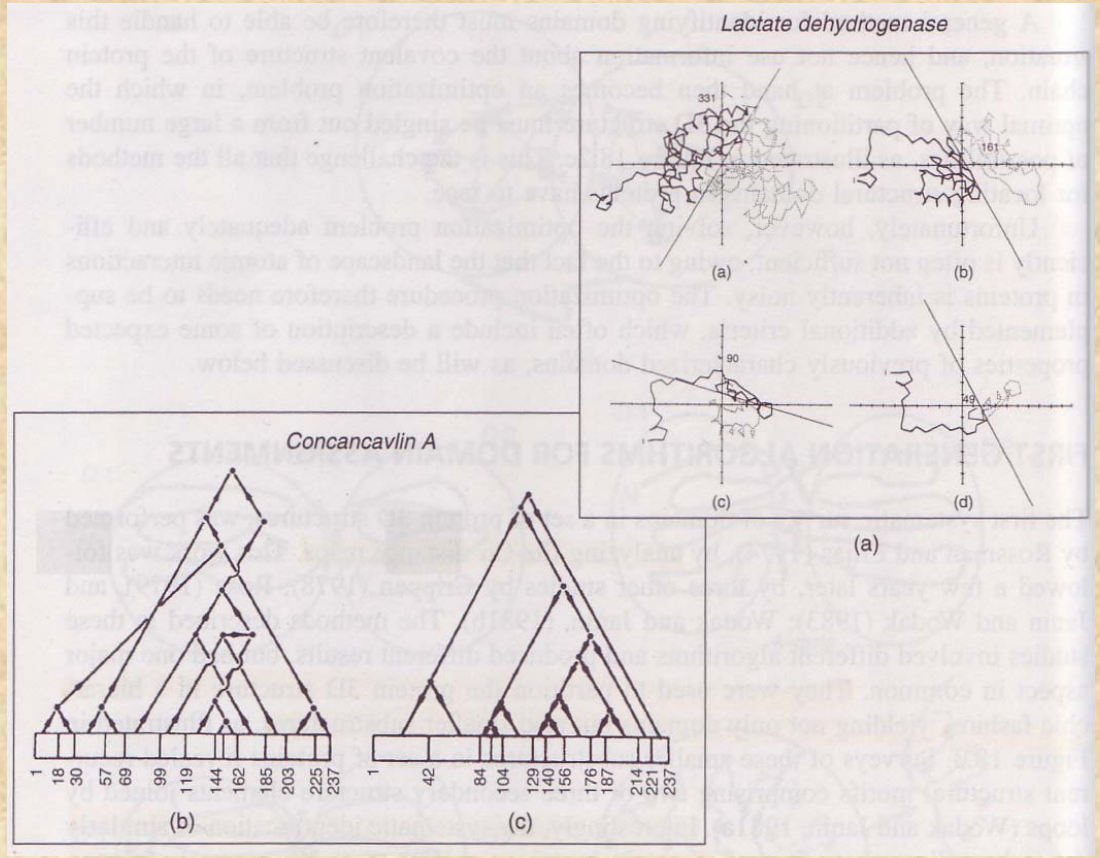
Identifying Number of Cuts



First Generation Methods - Partitioning

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

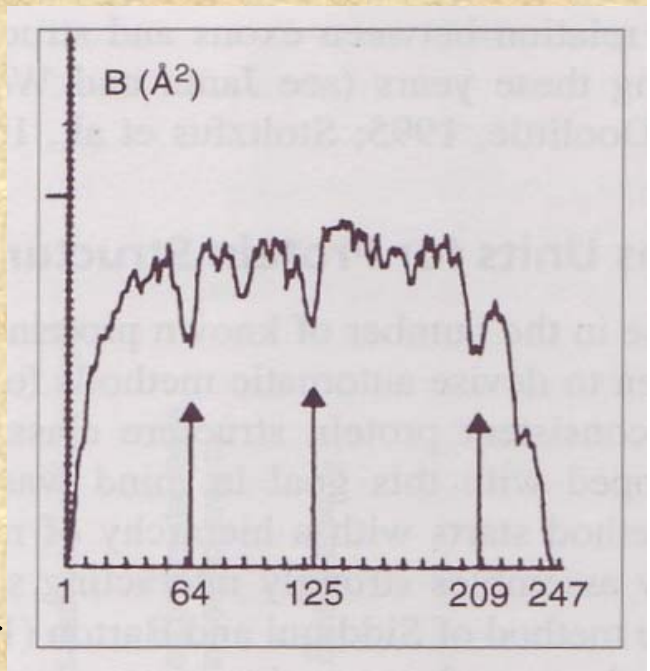
First Generation Methods



← Rose, 1979

↑
Crippen, 1978

↑
Wodak and Janin, 1981 →



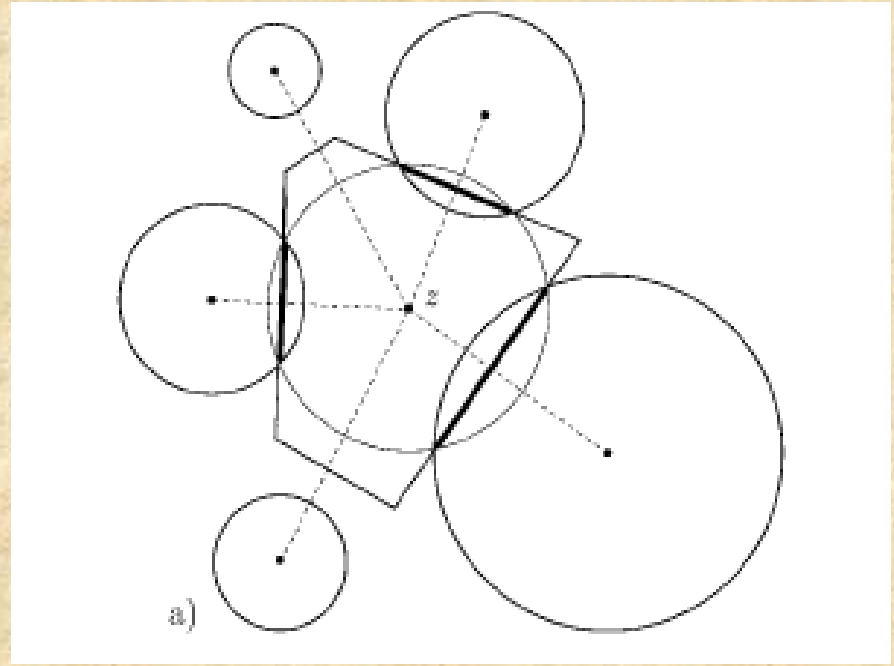
Second Generation Methods

- methods from other disciplines are adopted
  - graph theory
  - physics
  - statistics
- more general
- computationally more efficient

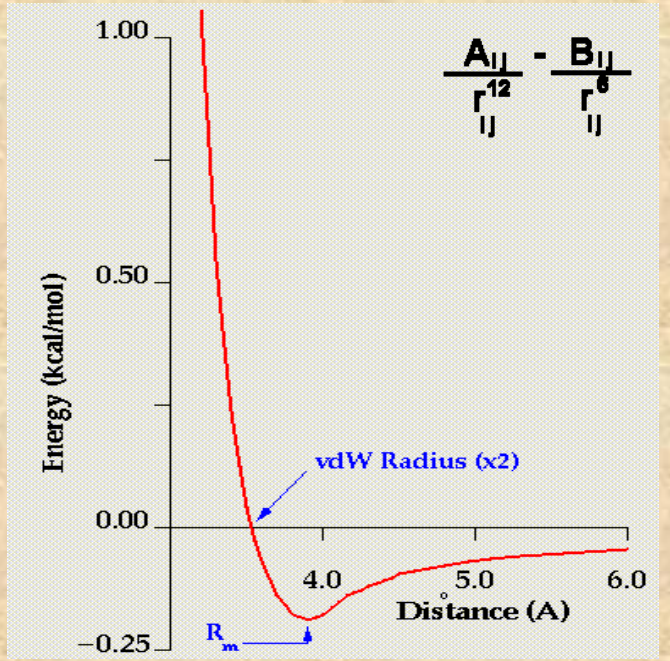
STRUDL

- 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
  - graph 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



$R = R_{vdW} + 1.4Å$

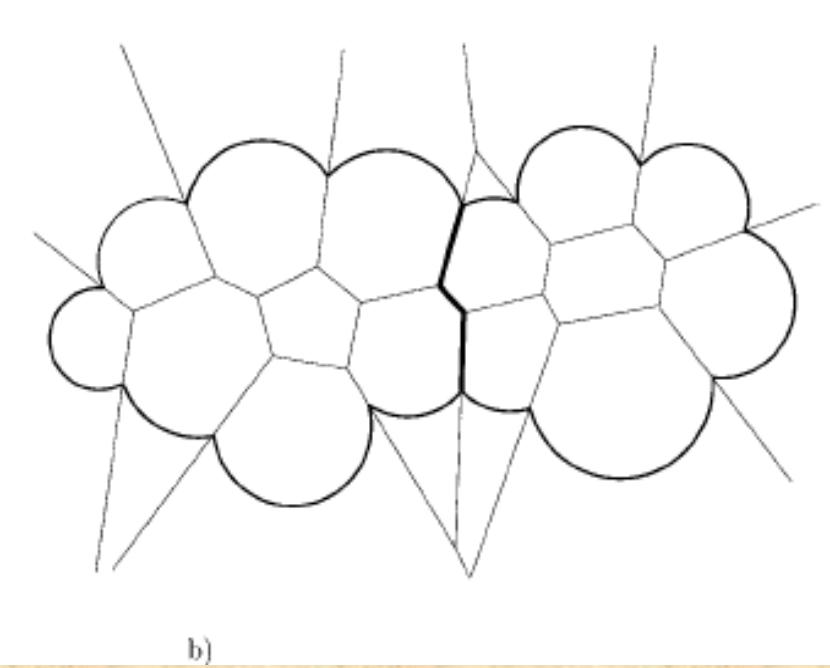


www.cgl.ucsf.edu/.../ tutorial/App/ljplot.g

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



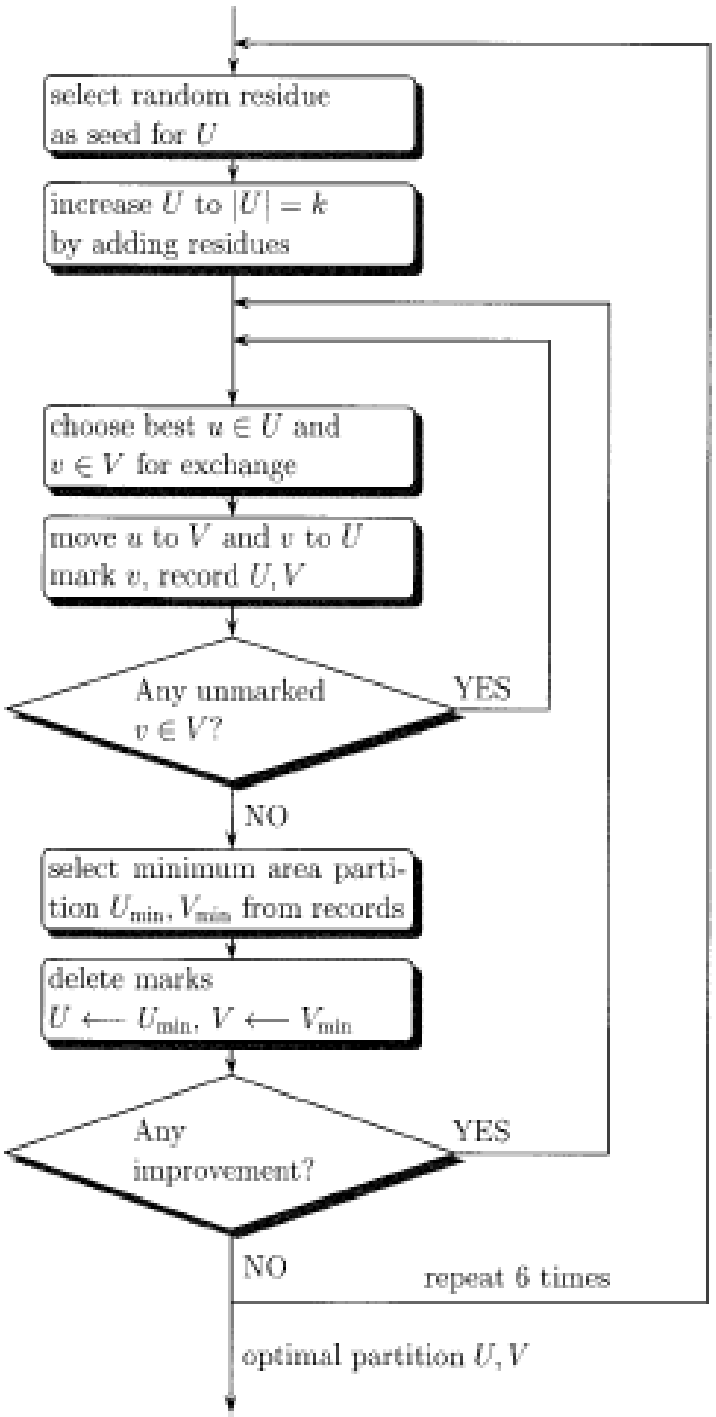
Contact area between two groups of atoms, all individual contact areas are simply added.

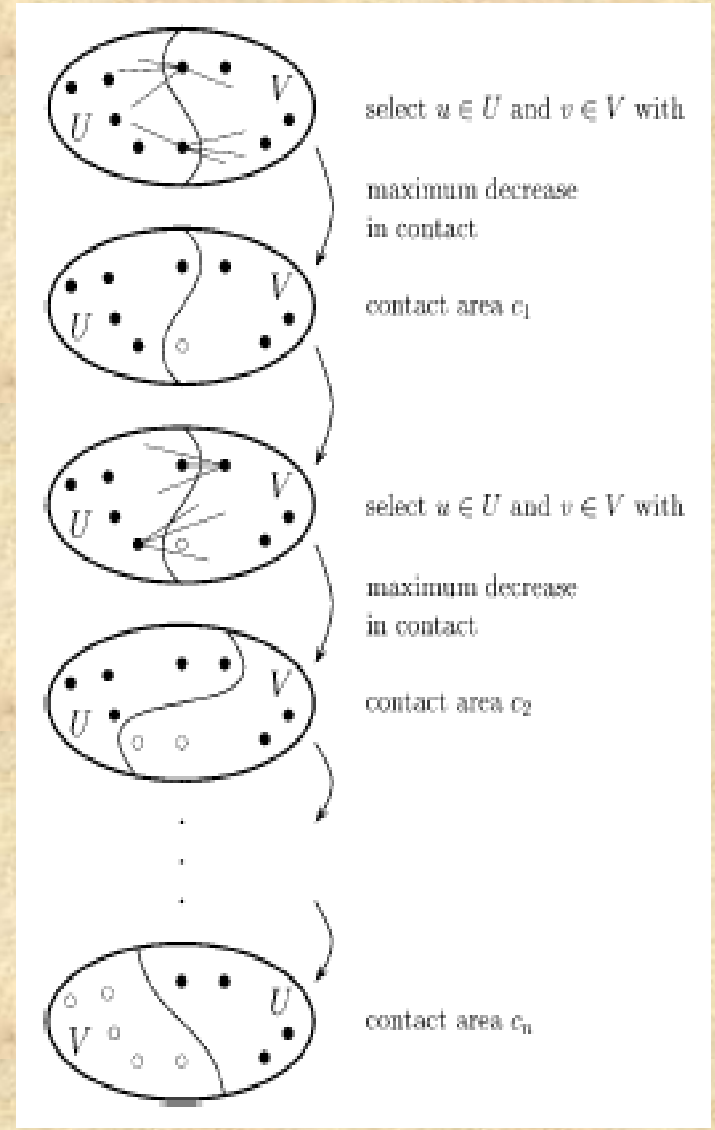
Contact map can be created based on pairwise contact areas!

STRUDL Algorithm

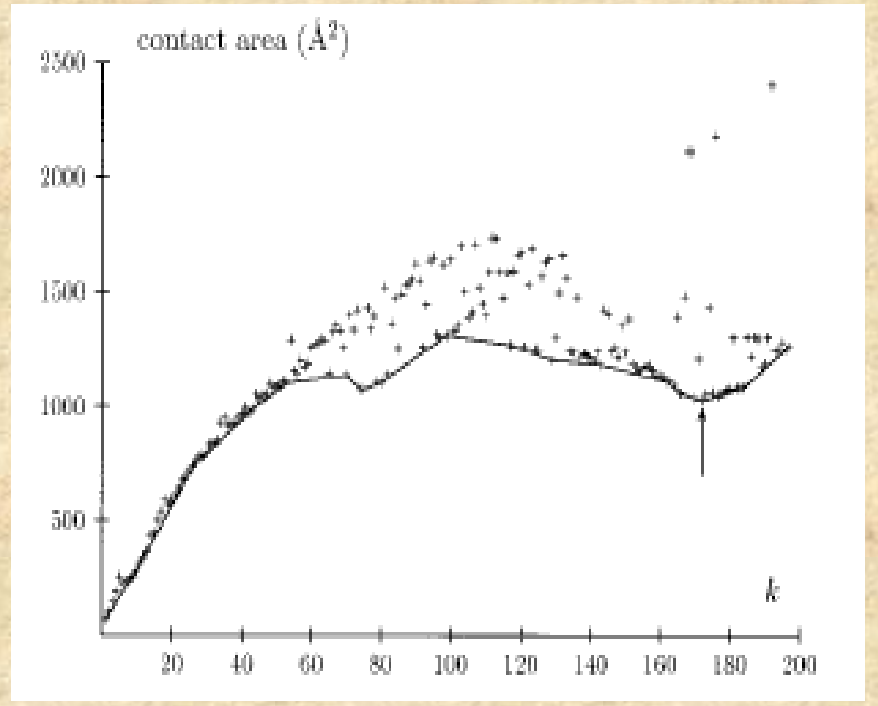
1. Start with a subset *U* consisting of 1 randomly chose residue (complement of *U* is *V*)
   1. *U* has size of *k* (predetermined);
   2. $1 \leq k \leq N/2$
2. Perform “exchanges”, moving 1 residue from *V* to *U* and one from *U* to *V* so as to maximally decrease contact between two sets
   1. 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

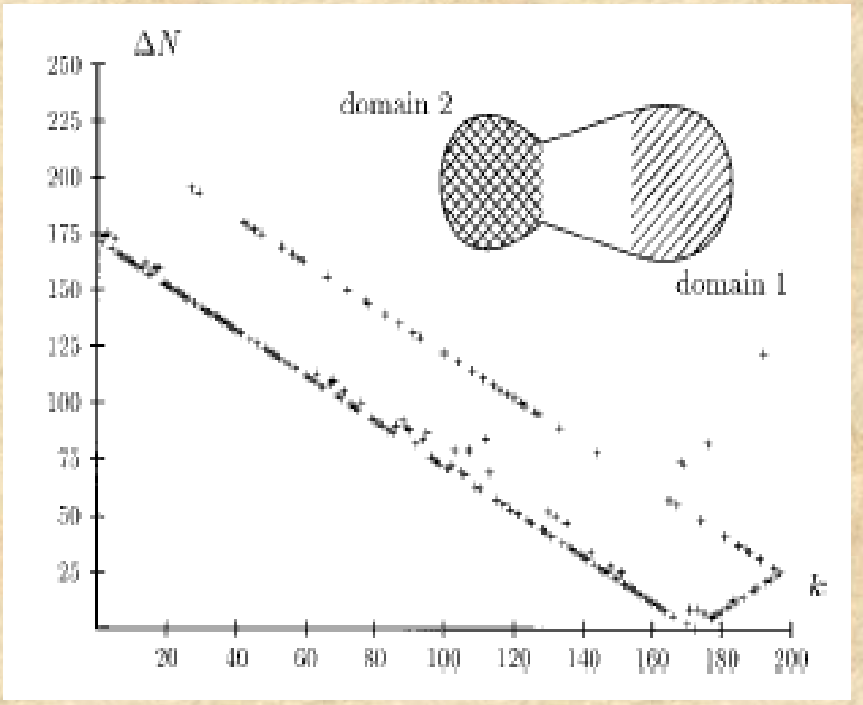
Partition Method



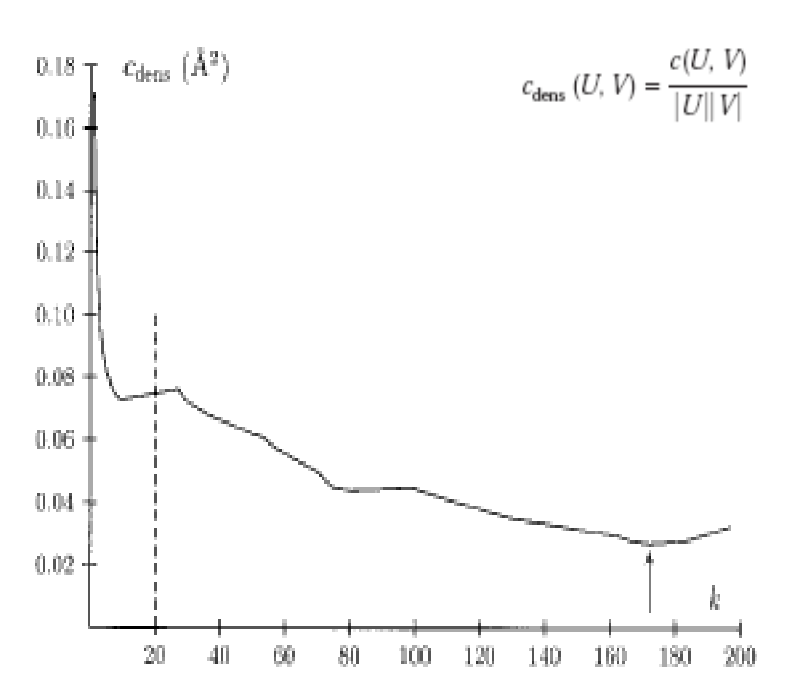


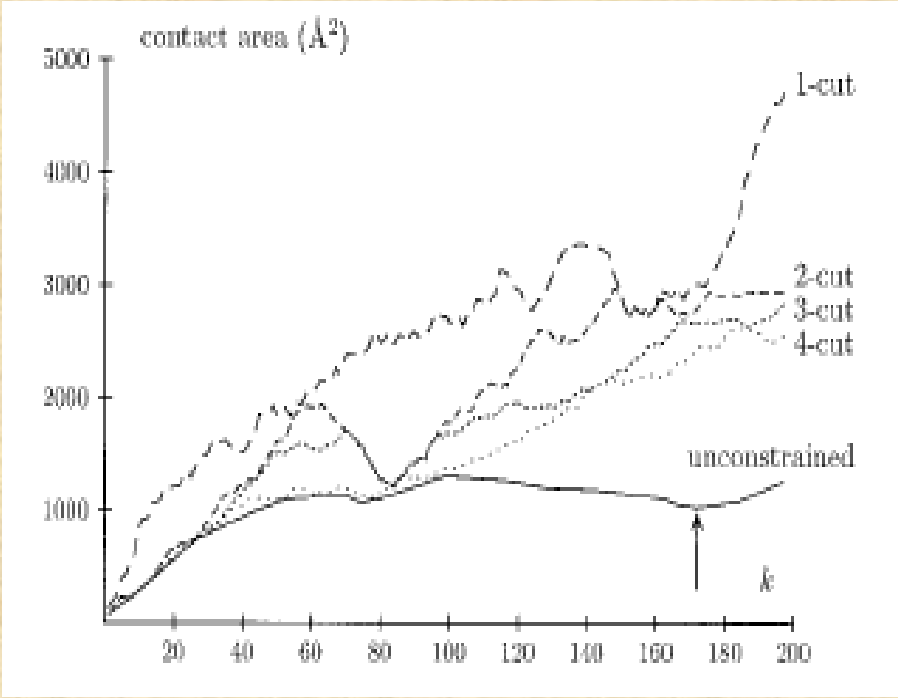
Behavior of the Algorithm





Minimum Contact Profile





contact area density

Note: short domains, where $k < 0.05N$, not allowed

Accepting/Rejecting Domains

- highly heuristic
- based on 9 different parameters
  - contact area density

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

  - normalized contact area

$$c_{\min}(U, V) = \frac{c(U, V)}{\min(|U|, |V|)}.$$

  - sum of interresidue contacts

$$c_{\text{prop}}(U, V) = \frac{c(U, V)}{\sum_{\text{all } i,j} c_{ij}}$$

  - compactness

$$b_{\text{tot}}(U) = \frac{1}{|U|} \sum_{i,j \in U} c_{ij}$$

Accepting/Rejecting Domains

- based on 9 different parameters
  - average interresidue contact

$$b_{\text{mean}}(U, V) = \tfrac{1}{2}\,(b_{\text{tot}}(U) + b_{\text{tot}}(V)).$$

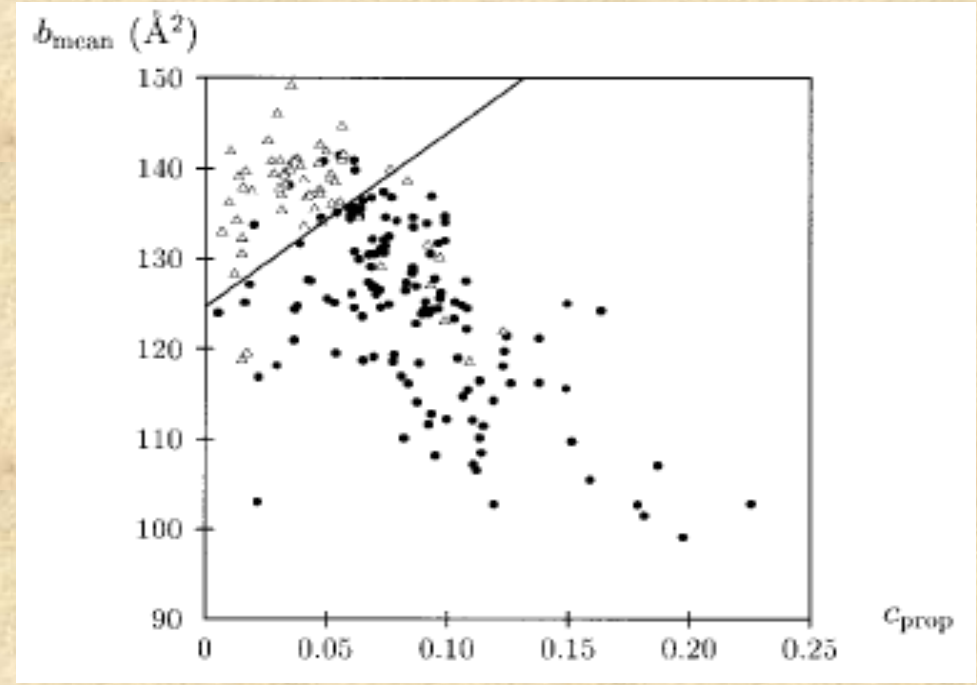
  - average slope of the profile
  - proportion of residues with positive slope
  - depth of the contact density profile at the minimum
  - total number of residues in the protein and the number of residues to the left of the minimum in the contact density partition

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)

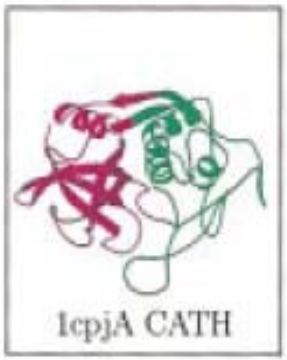
Tests

par.	c_{dens}	c_{min}	c_{prop}	b_{tot}	b_{mean}	ρ_{avg}	ρ_{prop}	d	n_{opt}
c_{dens}	1.00								
c_{min}	0.77	1.00							
c_{prop}	0.83	0.83	1.00						
b_{tot}	−0.32	−0.10	−0.08	1.00					
b_{mean}	−0.69	−0.59	−0.53	0.84	1.00				
ρ_{avg}	0.67	0.77	0.78	0.10	−0.36	1.00			
ρ_{prop}	0.21	0.46	0.44	0.12	−0.16	0.63	1.00		
d	−0.37	−0.55	−0.50	0.08	0.37	−0.62	−0.69	1.00	
n_{opt}	−0.59	−0.58	−0.43	0.47	0.69	−0.51	−0.15	0.43	1.00
n_{total}	−0.62	−0.53	−0.52	0.42	0.62	−0.49	−0.03	0.30	0.92

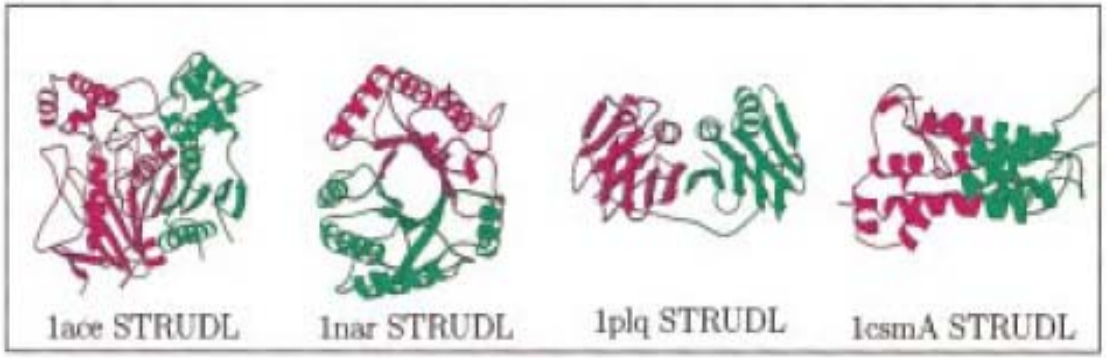
Results

- 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

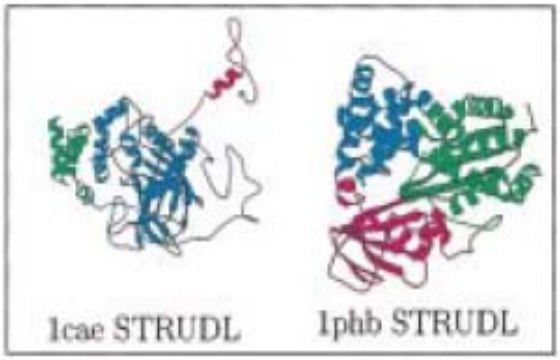
Errors?



a) Undercut

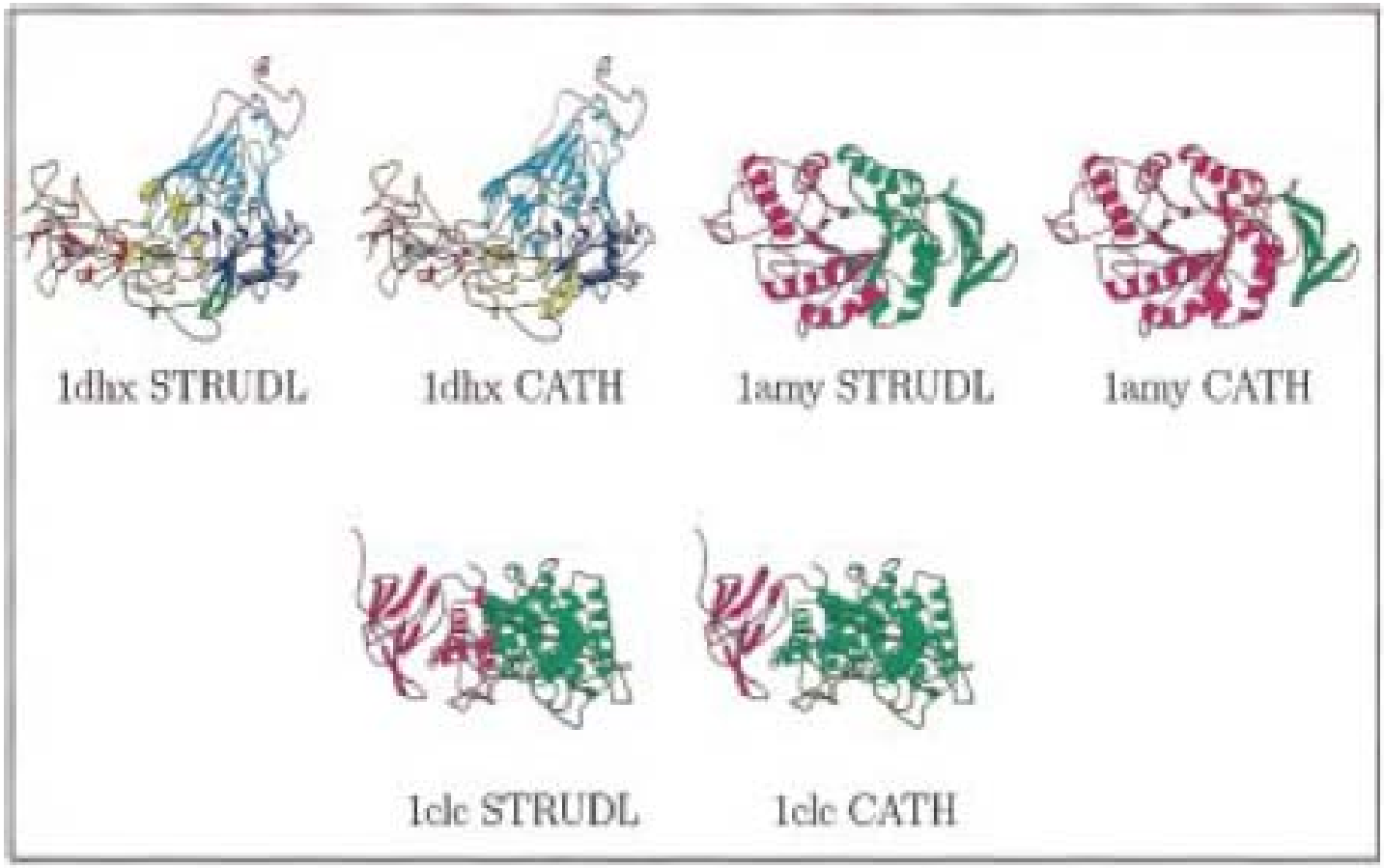


b) Overcut with single architecture



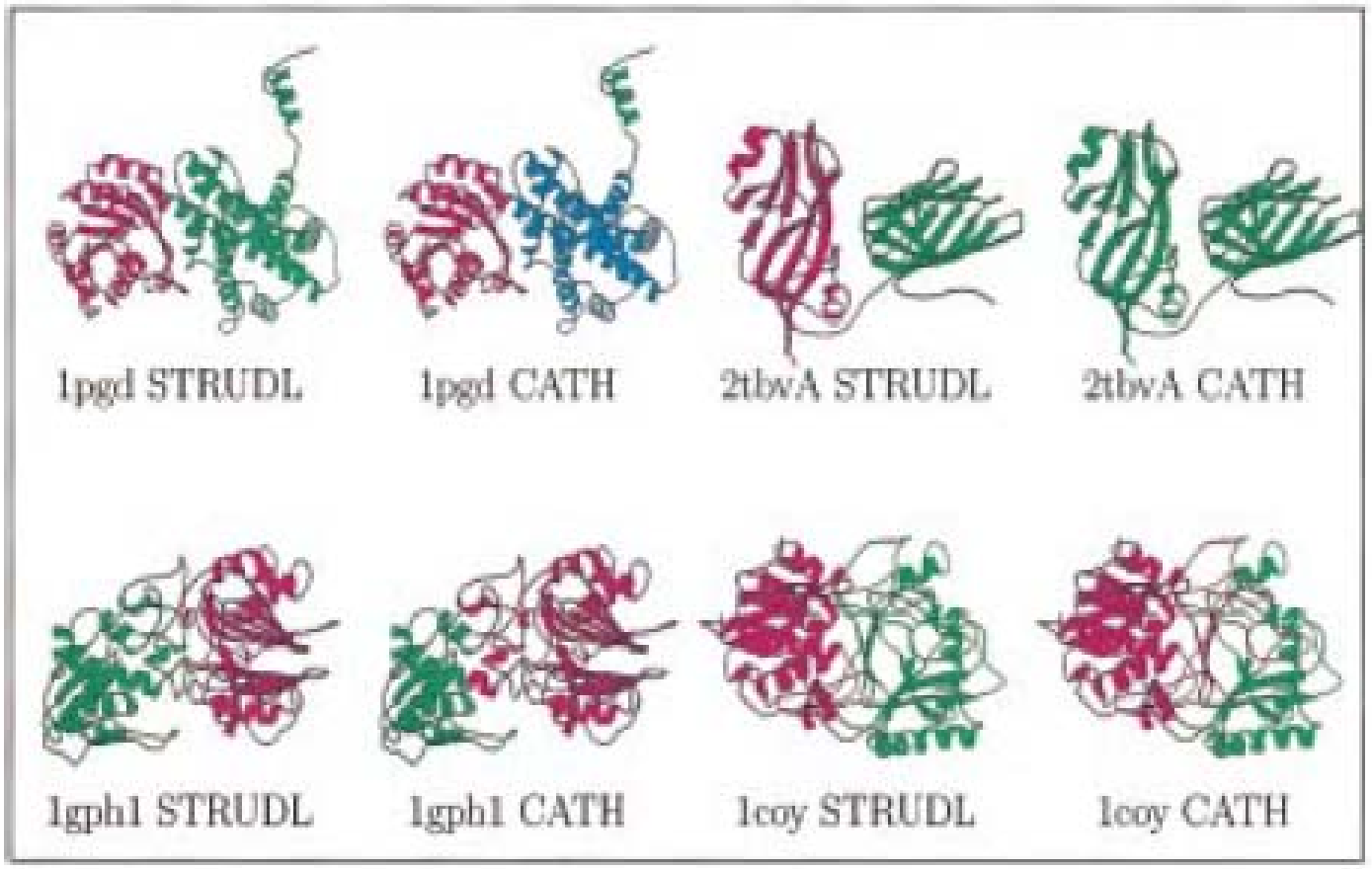
c) Overcut with appendices and multiple architecture

Analysis



d) Complex

Analysis



e) Debatable