2 Papers Concerning Automatic Domain Detection in Proteins

Protein Domain Example

The papers discussed today with take as input the 3 dimensional structures and output a proposed domain decomposition.

Motivation

1) Proteins have hierarchical organization. This may help us understand protein folding, evolution and function.

2) Efficiently maintain structural domain databases such as CATH.

An automatics method involving cluster analysis of secondary structures for the identification of domains in proteins

R. Sowdhamini and T. Blundell

Overview of Algorithm

1) Input: 3D structure of protein

2) Identification of secondary structures: alpha helices and beta sheets using the program SSTRUC

3) Calculate a distance measure (called the "proximity index") between every pair of secondary structures.

4) Cluster secondary structures based on proximity index and make a dendogram.

5) Choose where to cut the dendogram to find the domains.

Proximity Index

$$p_{i,j} = \frac{\sum_{k=1}^{k=n_i} \sum_{l=1}^{l=n_j} d_{k,l}}{n_i \times n_j}$$

$d_{k,l}$ = distance between residues k and l
ni and nj = the number of residues in secondary structure i and j

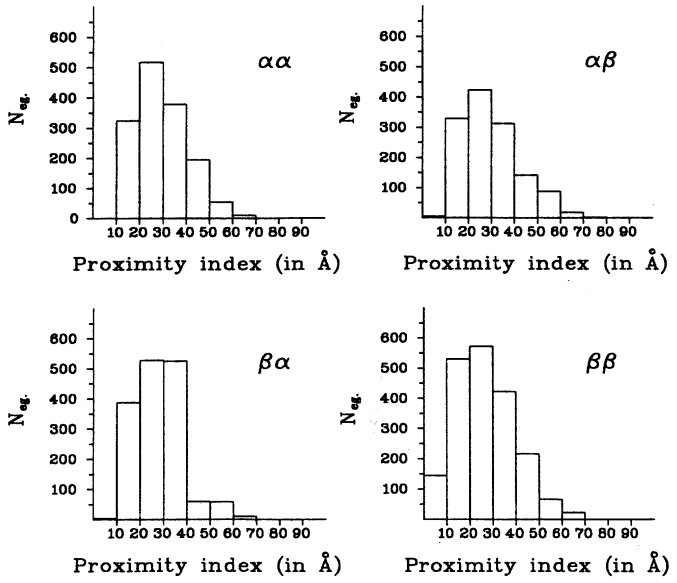
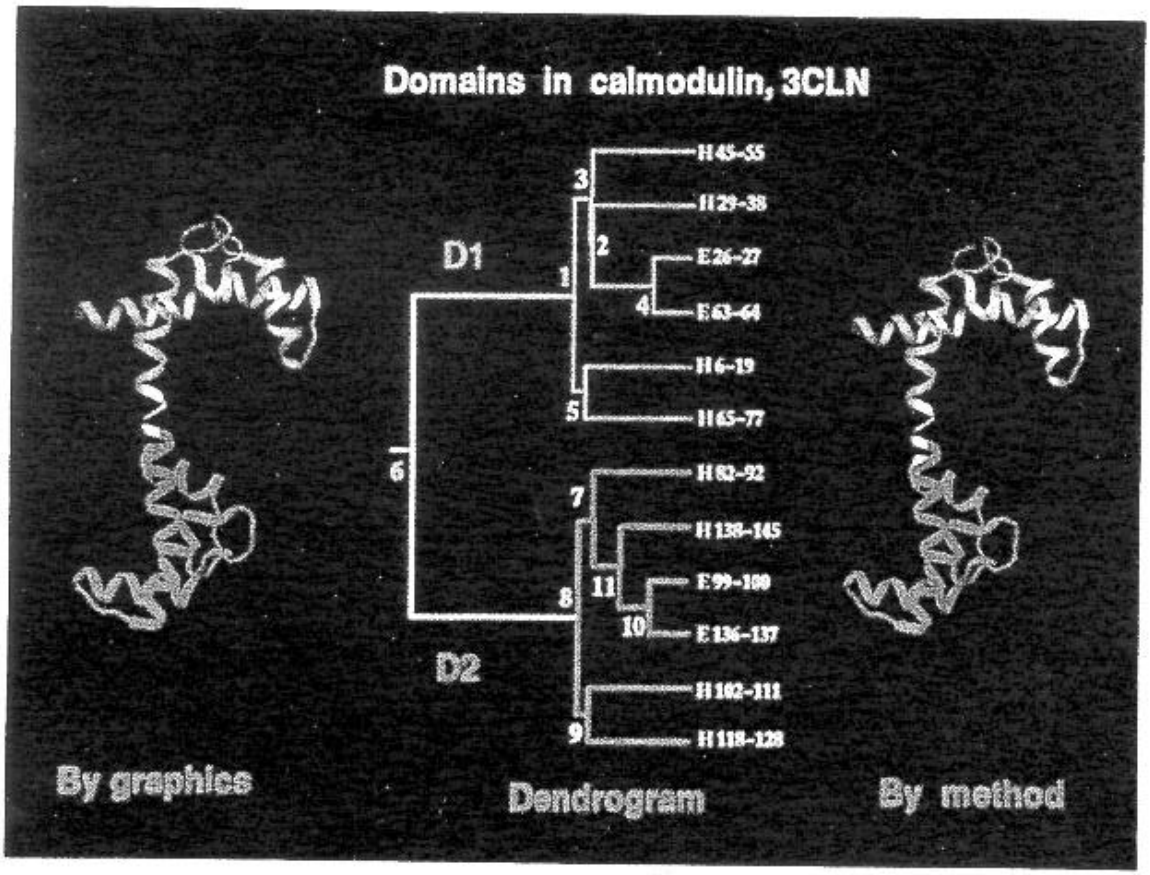


Fig. 1. Distribution of proximity indices between pairs of secondary structures (α: helix; β: extended strand) of 20 different proteins of varying sizes and folds. These 20 proteins form a subset of the 101 proteins used for analysis. $N_{eg.}$ is the number of examples.

Cluster and Dendogram



Automate Dendogram Cutting

Try all combinations of clusters and compute disjoint factor.
Choose combination withthe highest disjoint factor.

$$D_f = \alpha * W_{1,2} * W_{1,3} * ... * W_{n-1,n}$$

Where α is a ration between the mean proximity indices of all secondary structures to the mean proximity indices of within clusters and Wi,j are weighting factors to make sure clusters i and j aren't too close.

Empirically,
$D_f > 1.5$ implies the domains are disjoing.
$1.25 \leq Df \leq 1.5$ implies the domains interact.
$1.0 \leq Df < 1.24$ implies the domains are conjoint.

Formulas for α and $W_{i,j}$

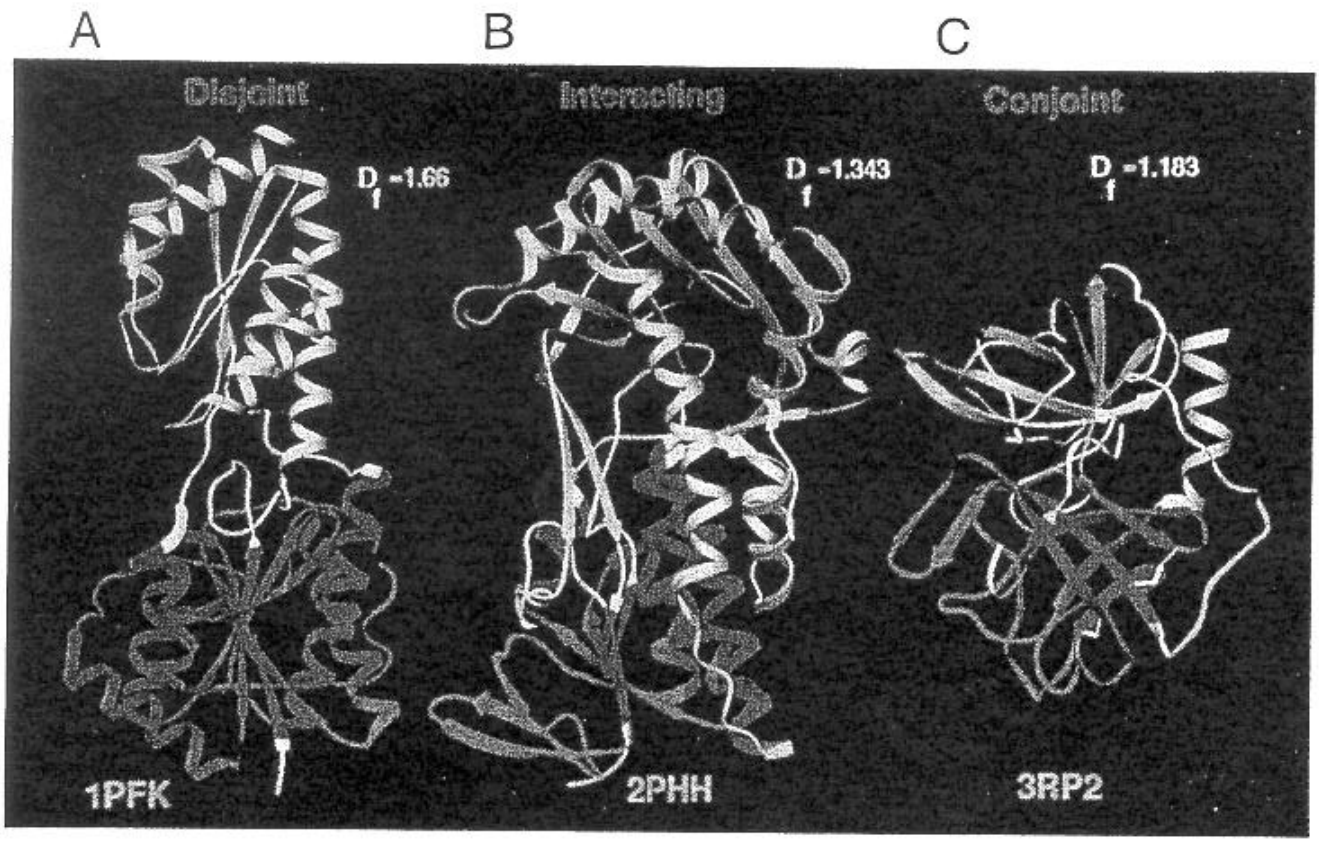
$$\alpha = \frac{\dfrac{\sum_{i=1}^{i=nt-1} \sum_{j=i+1}^{j=nt} p_{i,j}}{\dfrac{nt(nt-1)}{2}}}{\dfrac{\sum_{k=1}^{k=n_s} \sum_{ii=1}^{ii=ist(k)-1} \sum_{jj=ii+1}^{jj=ist(k)} p_{ii;k,jj;k}}{\dfrac{ist(k)(ist(k)-1)}{2}}}$$

n_s = number of clusters
nt = number of secondary structures
ist(k) = number of secondary structures in cluster k

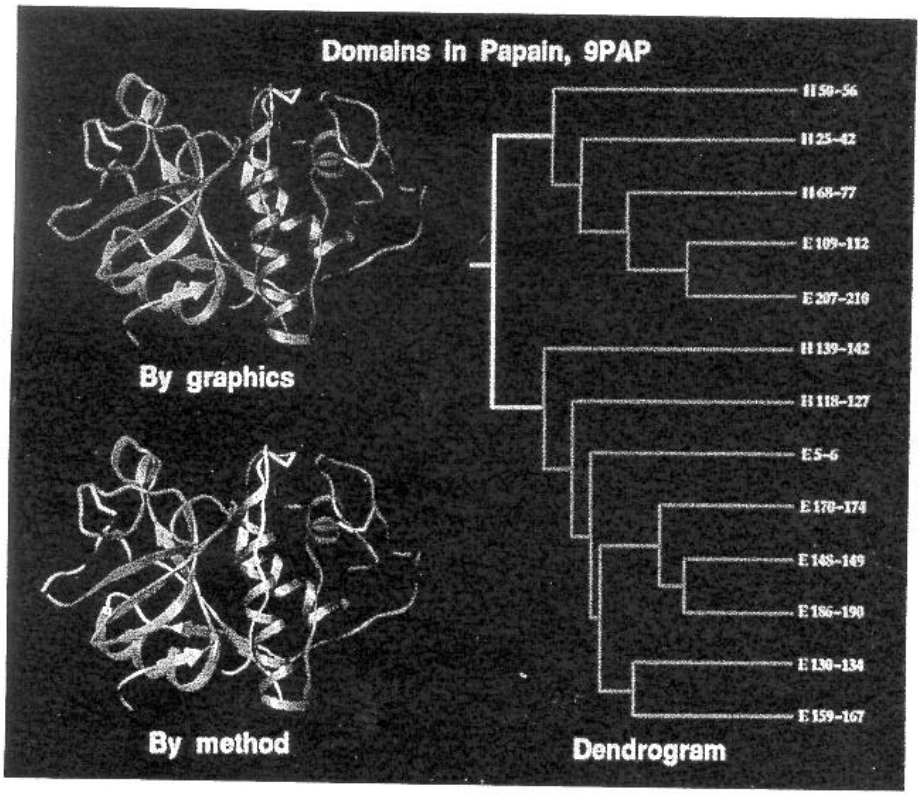
$$W_{1,2} = \frac{\sum_{i=1;1}^{i=1;l} \sum_{j=2;1}^{j=2;m} n(i) \times n(j) - \sum_{i=1;1}^{i=1;l} \sum_{j=2;1}^{j=2;m} \sum_{ii=1}^{ii=n(i)} \sum_{jj=1}^{jj=n(j)} d^2_{1;i;ii,2;j;jj}}{\sum_{i=1;1}^{i=1;l} \sum_{j=2,1}^{j=2,m} n(i) \times n(j)}$$

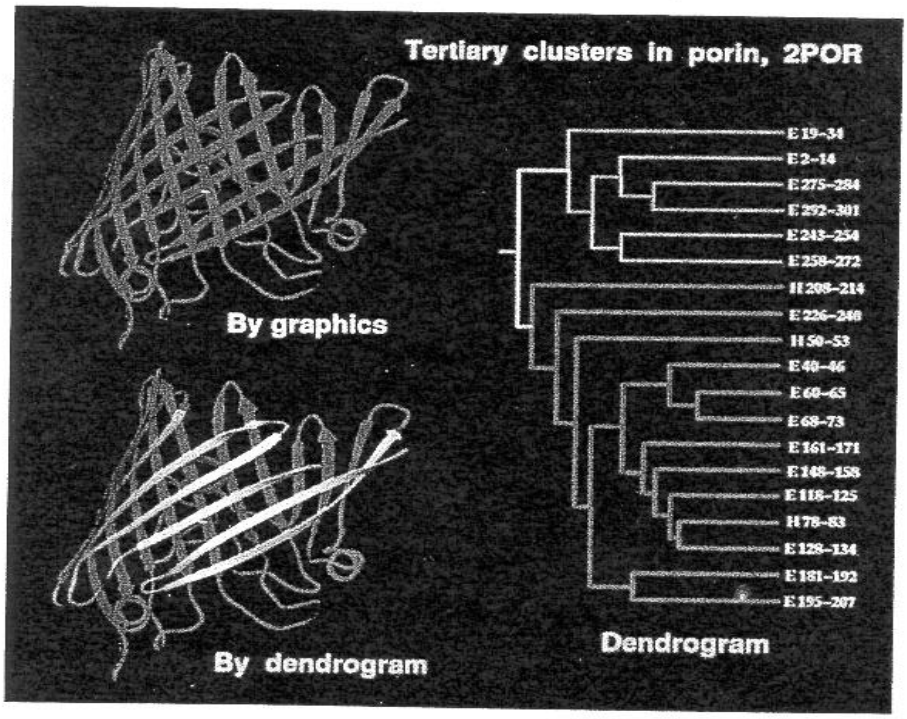
di,j = number of residues within 7 A between secondary structure i and j
n(i) = number of residues in secondary structure i

Disjoint, Interacting, Conjoint



Results


Domains in Papain, 9PAP
By graphics
By method
H 50–56
H 25–42
H 68–77
E 109–112
E 207–210
H 139–142
H 118–127
E 5–6
E 170–174
E 148–149
E 186–190
E 130–134
E 159–167
Dendrogram


Tertiary clusters in porin, 2POR
By graphics
By dendrogram
E 19–34
E 2–14
E 275–284
E 292–301
E 243–254
E 258–272
H 208–214
E 226–240
H 50–53
E 40–46
E 60–65
E 68–73
E 161–171
E 148–158
E 118–125
H 78–83
E 128–134
E 181–192
E 195–207
Dendrogram

Results

1) Visually method looks reasonable.

2) Can find domains that are not a continous sequence.

3) Often gets number of domains correct though it sometimes overestimates.

4) Boundary borders are not tight.

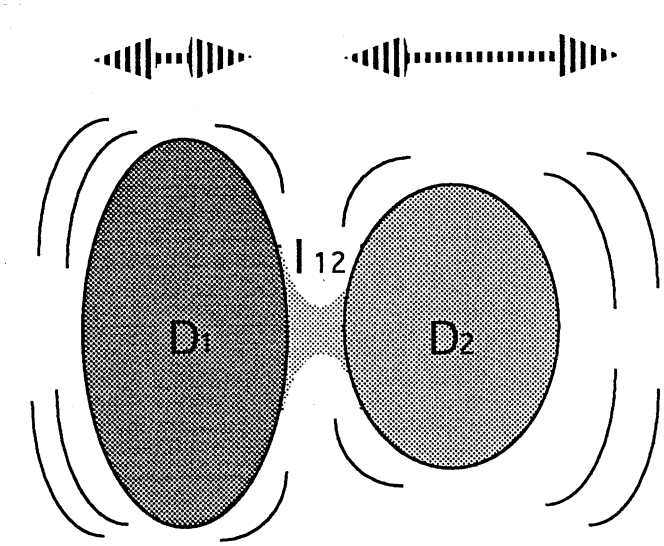
Parser for Protein Folding Units

Liisa Holm and Chris Sander

Perspective

This paper considers domains as independent folding units. Trys to answer question: If a protein was slowly unfolding what parts of the protein would seperate from each other first. This could give insight into autonomous folding units.

Physical Model



$$\tau 2 \sim \frac{\mu}{I_{12}}$$

1) Assume we propose 2 domains.
2) We model there movements as a harmonic oscillator. The potiential energy is then:

$$V(x) = .5 * V_0 * x^2$$

and the square of the oscillation time (period) is:

$$\tau^2 = (2\pi)^2 \mu / V_0$$

where V0 is the contact potiential and μ is the reduced mass.
3) If the oscillations are slow enough the the proposed domains are reasonable.

Approximation for V_0

V_0 is the force constant of the interface.

1) Each contact pair ($\leq$ 4.0 Angstroms) contributes
   1.0 Kcal/mol/(Angstrom)^2

2) Each Hydrogren Bond contributes 15.0 Kcal/mol/(Angstrom)^2

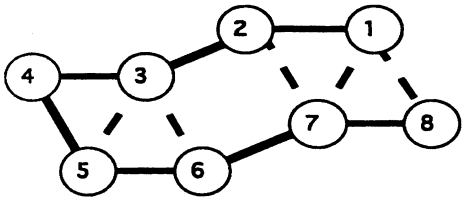
3) V_0 for the interface is the sum of all the contributions.

Algorithm

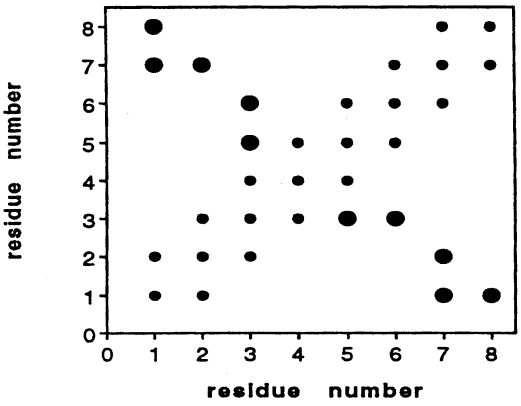
1) Input: 3D structure of protein
2) Make Contact Matrix
3) Find ordering of amino acids which make proposing domains easier.
4) Based on ordering of amino acids find the best way to choose 2 domains.
5) Apply steps 1 - 4 on the subdomains found in step 4.
6) Terminate when subdividing no longer is reasonable.

Contact Matrix

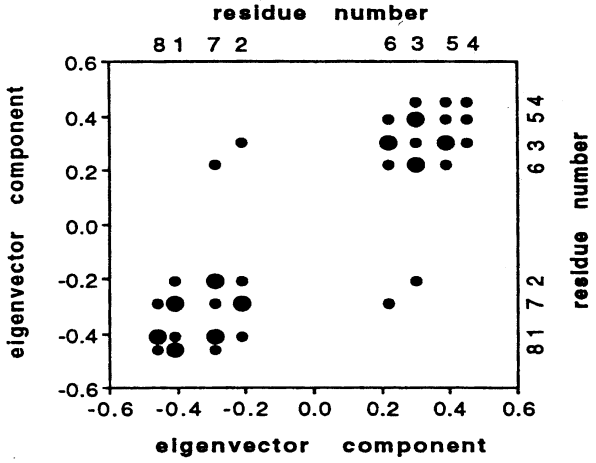
A.



B.



C.



Finding Domains from Reordering

1) Attempt to cut protein into domains after every amino acid in reordered sequence.

2) Calculate τ for every cut and choose the cut which maximizes τ.

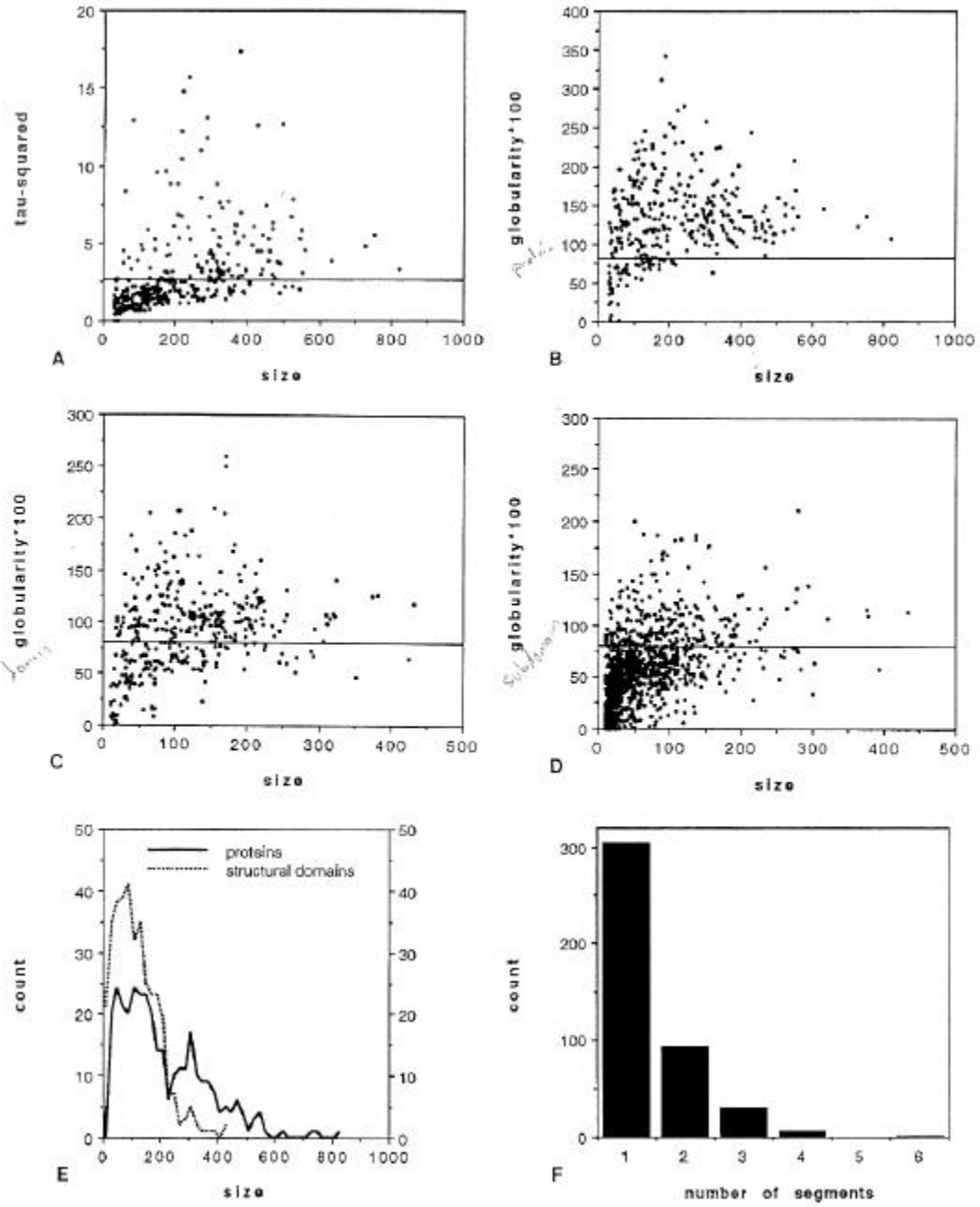
3) Some rules for splitting a protein into subdomains:

   a) Lower limit of domain size is 40 residues
   b) Highly flexible units ($\tau > 2.6$) are always cut
   c) Highly cooperative β-sheet networks are never cut
   d) A cut is accepted if both subdomains are compact. That is $\gamma > 0.80$ where

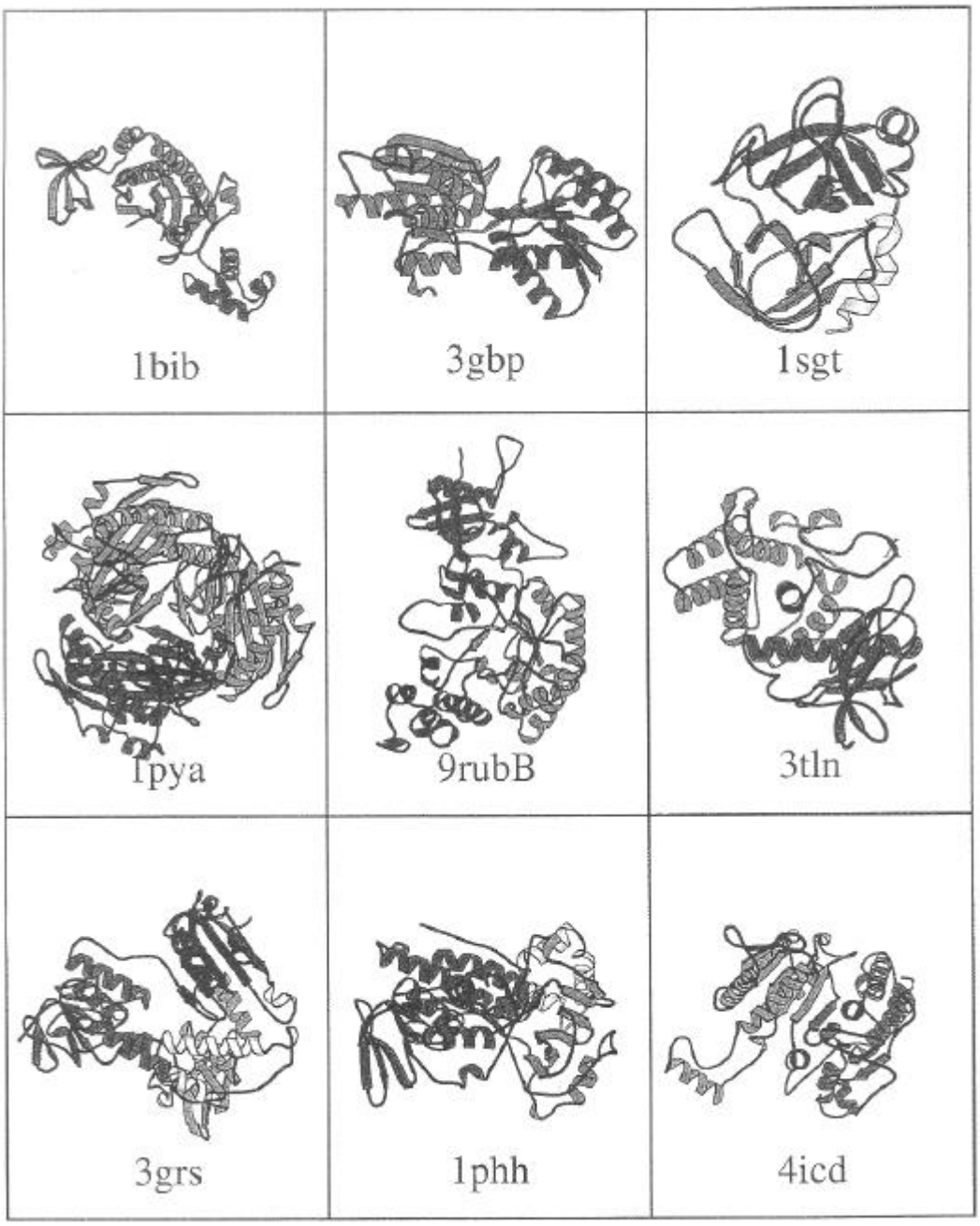
$$g = \frac{1}{N} \sum_i \sum_{j<i-3} a_{ij}$$

   where a_{ij} is the contact strength between residues i and j.
   e) A cut which results in a small nonglobular unit is accepted if the larger domain is then cut when algorithm is applied recursively on it.

Tables


tau-squared
size
A
globularity*100
size
B
globularity*100
size
C
globularity*100
size
D
count
protsins
structural domains
size
E
count
number of segments
F

Results



Results

1) Can find domains composed of a noncontinuous chain. Though 75% of the domains it finds are continuous chains.

2) There is some experimental evidence that domains can fold independently.

3) Method has problems with ambigous structures such as TIM barrels.