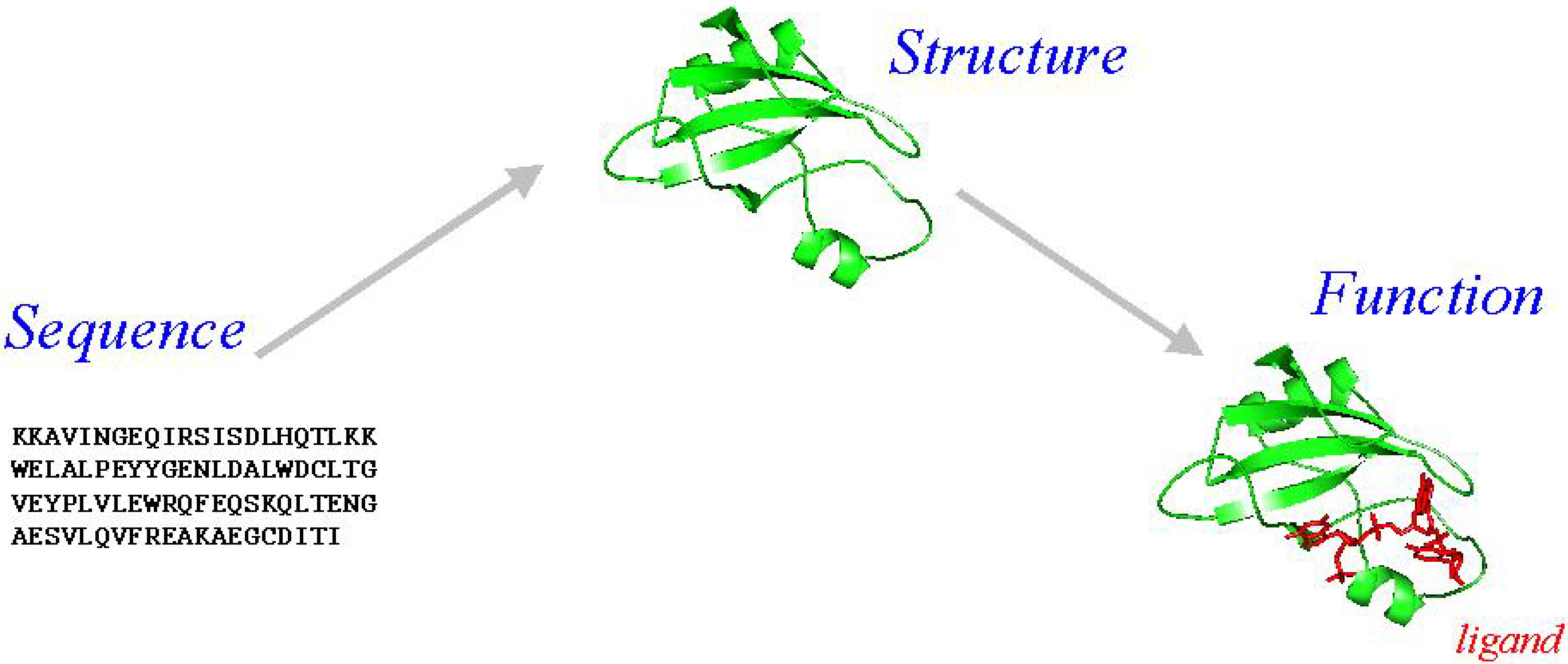
Protein Shape Descriptors

Patrice Koehl

Stanford University

http://csb.stanford.edu/koehl/

From Sequence to Function and Back…



Sequence

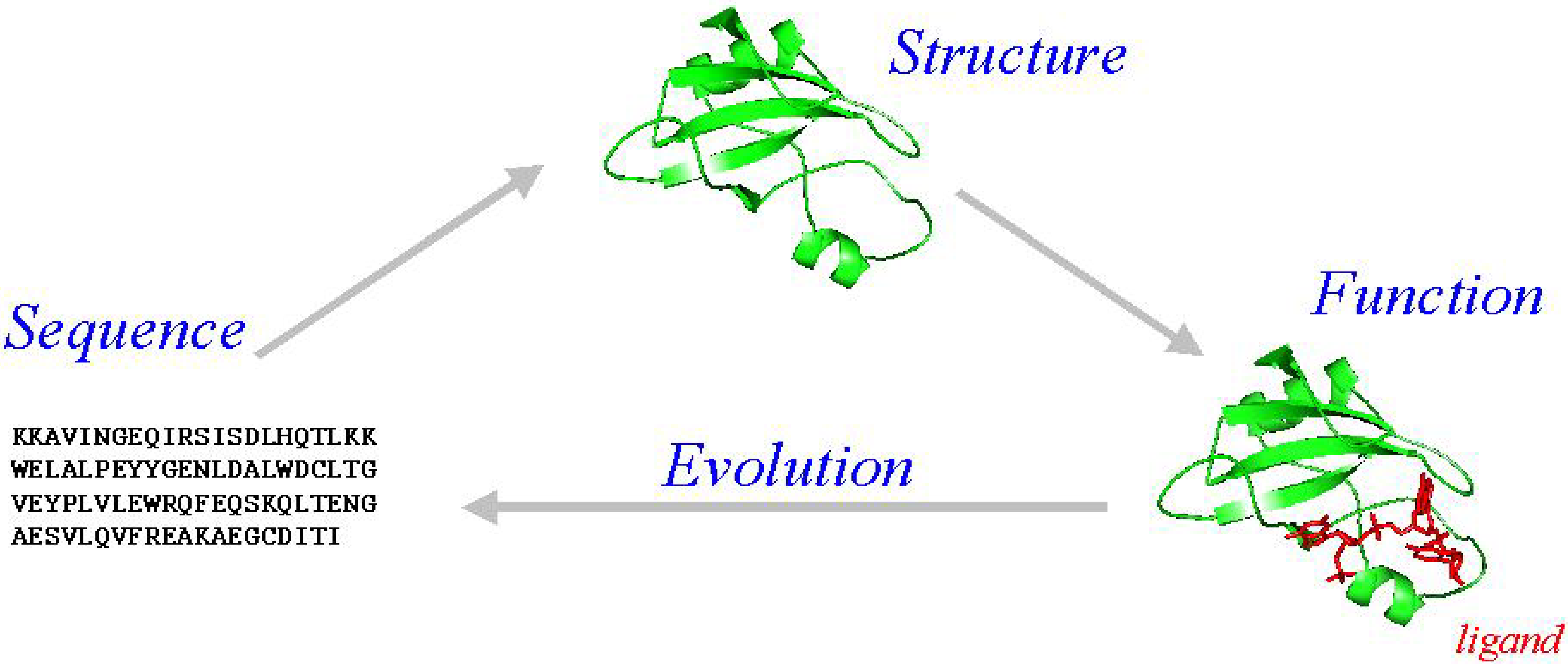
KKAVINGEQIRSISDLHQTLKK
WELALPEYYGENLDALWDCLTG
VEYPLVLEWRQFEQSKQLTENG
AESVLQVFREAKAEGCDITI

Structure

Function

ligand

From Sequence to Function and Back…



Structure

Sequence

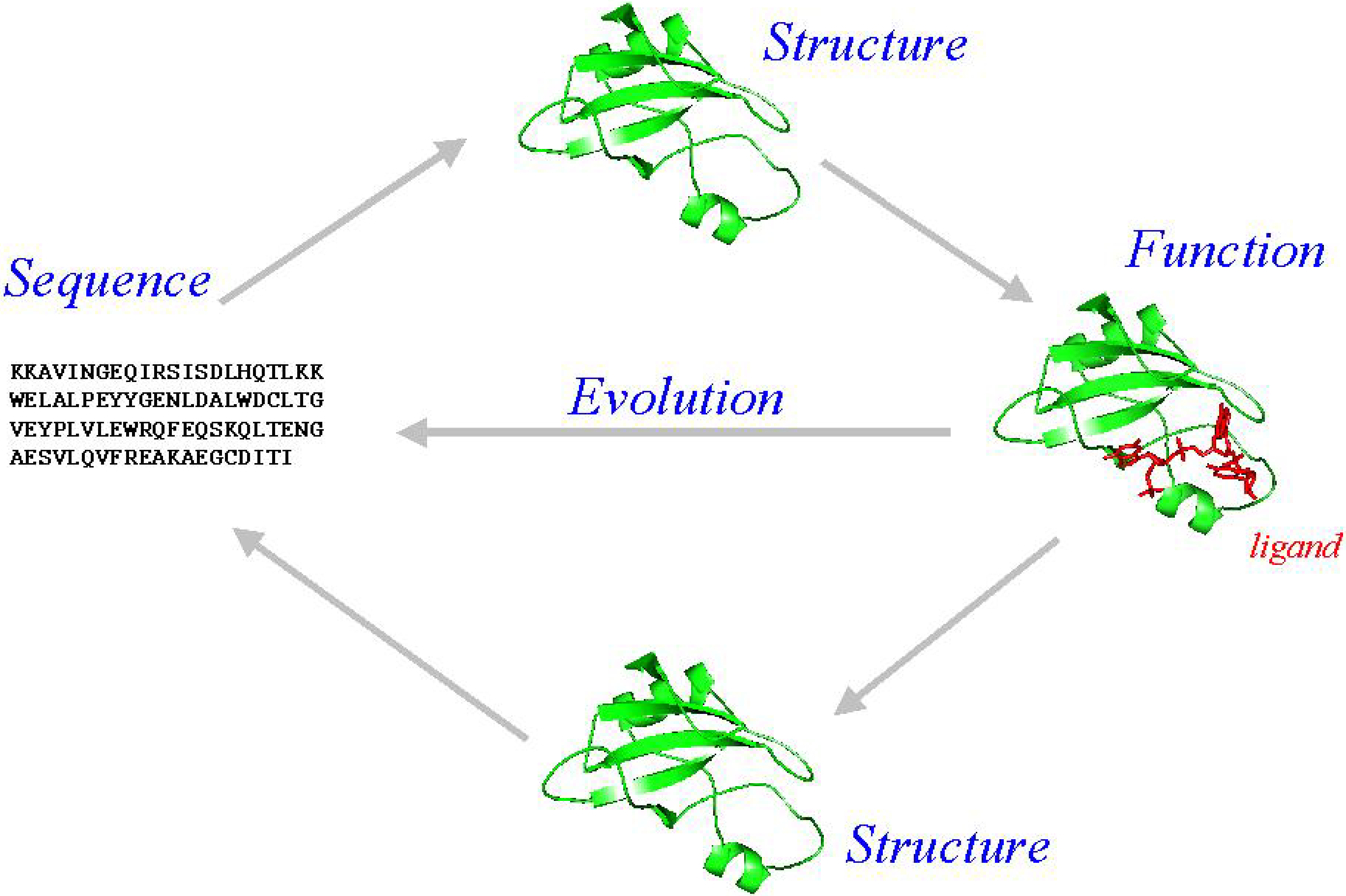
```
KKAVINGEQIRSISDLHQTLKK
WELALPEYYGENLDALWDCLTG
VEYPLVLEWRQFEQSKQLTENG
AESVLQVFREAKAEGCDITI
```

Function

Evolution

ligand

From Sequence to Function and Back…



Structure

Sequence

```
KKAVINGEQIRSISDLHQTLKK
WELALPEYYGENLDALWDCLTG
VEYPLVLEWRQFEQSKQLTENG
AESVLQVFREAKAEGCDITI
```

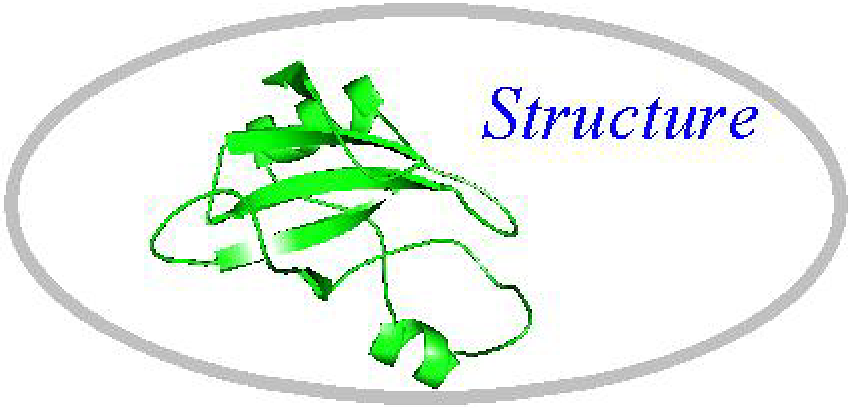
Function

ligand

Evolution

Structure

From Sequence to Function and Back…



Sequence

KQLAVIMGKQIRSISDLHQILKK
WETLRLDEYYCGEHLDALWDCLTG
VRYPLVLRWSQFEQSIQLTENG
AESVLQVFREAERAEGCDITI

Function

Outline

- Introduction

  What is a Protein?

- Protein Energy Functions

  Computational Geometry Tools

- Classifying Proteins

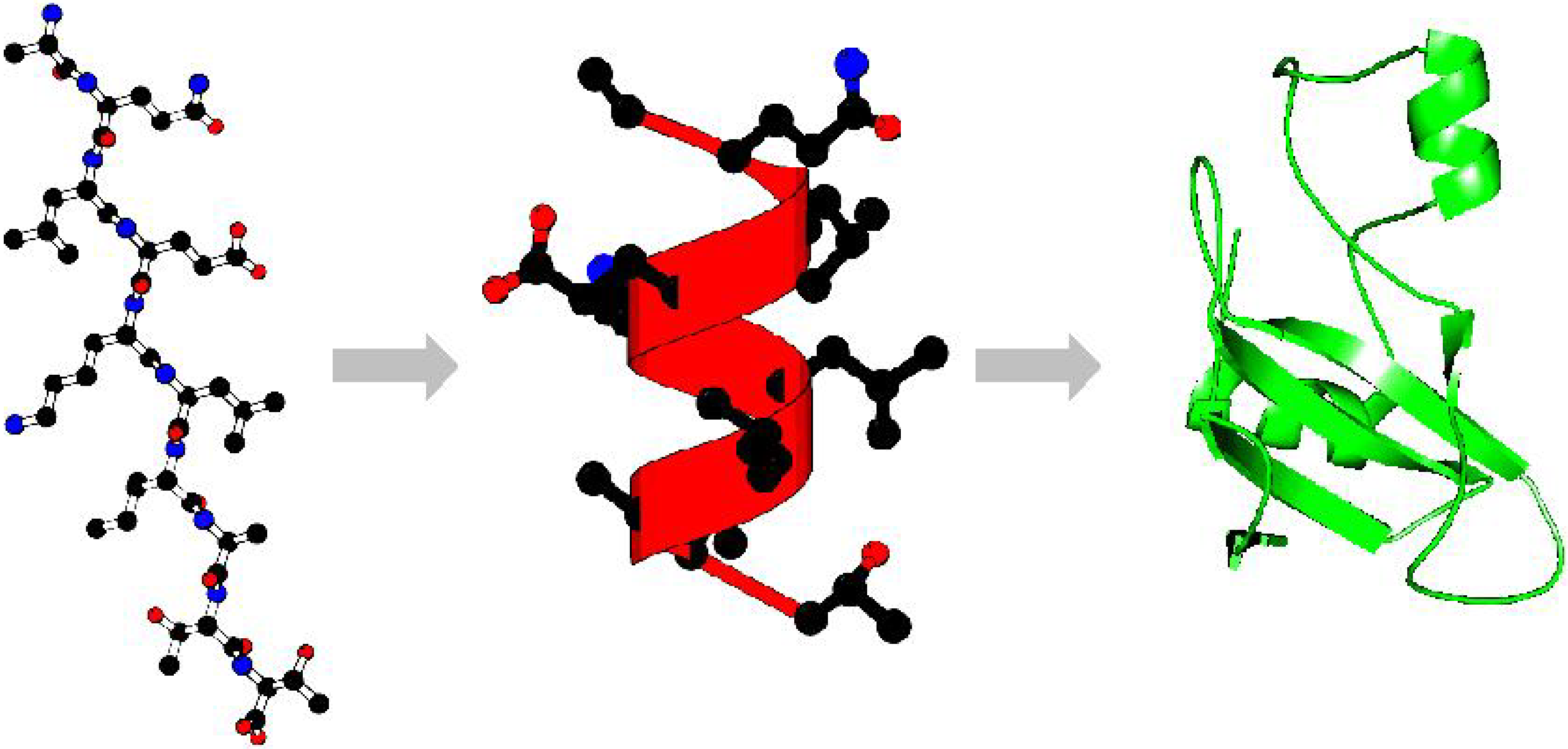
  The Shapes of Protein Structures

Outline

- Introduction
  - *What is a Protein?*
- Protein Energy Functions
  - *Computational Geometry Tools*
- Classifying Proteins
  - *The Shapes of Protein Structures*

What is a Protein ?

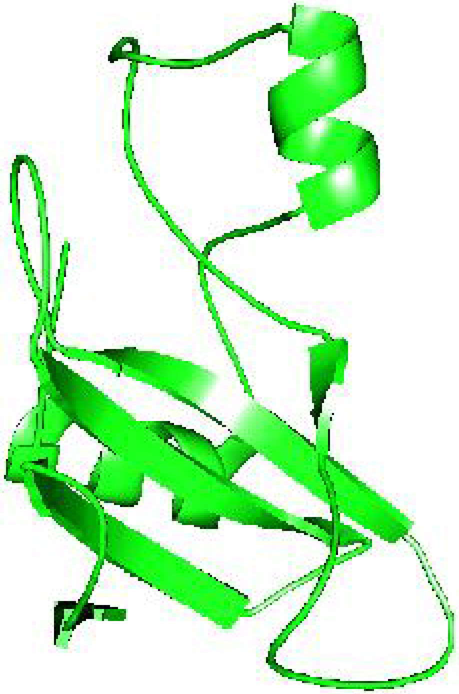
Primary Structure *Secondary Structure* *Tertiary Structure*



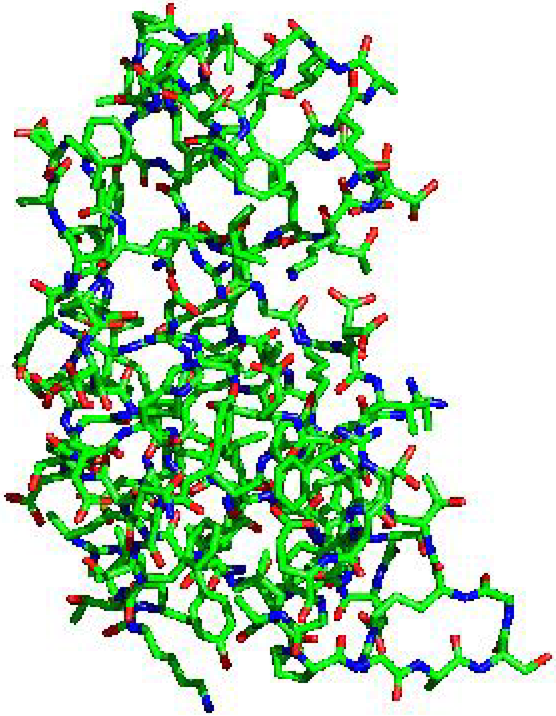
Sequence of amino acids

Native protein

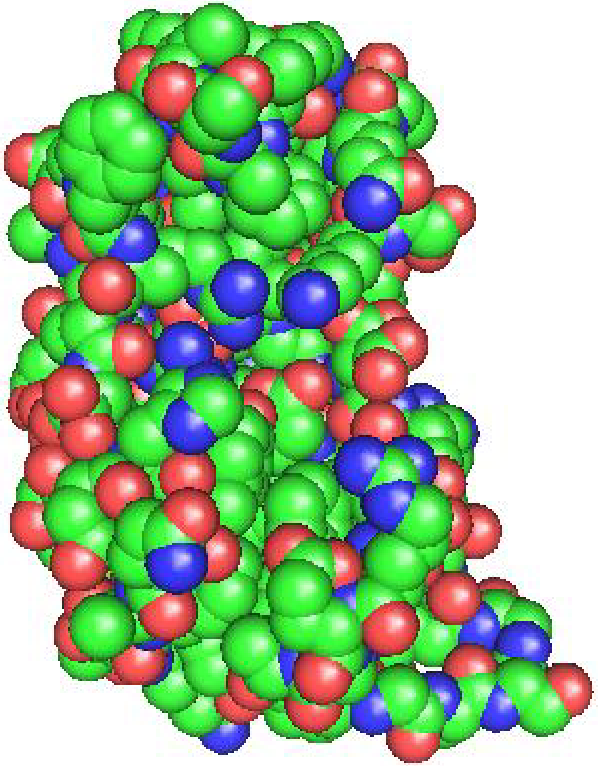
Protein Representations



Cartoon



Stick



Space-filling Model

Outline

- Introduction
  - *What is a Protein?*
- Protein Energy Functions
  - *Computational Geometry Tools*
- Classifying Proteins
  - *The Shapes of Protein Structures*

Energy of a Protein

Bonded Interactions

Bonds, Angles, Dihedral angles

Non Bonded Interactions

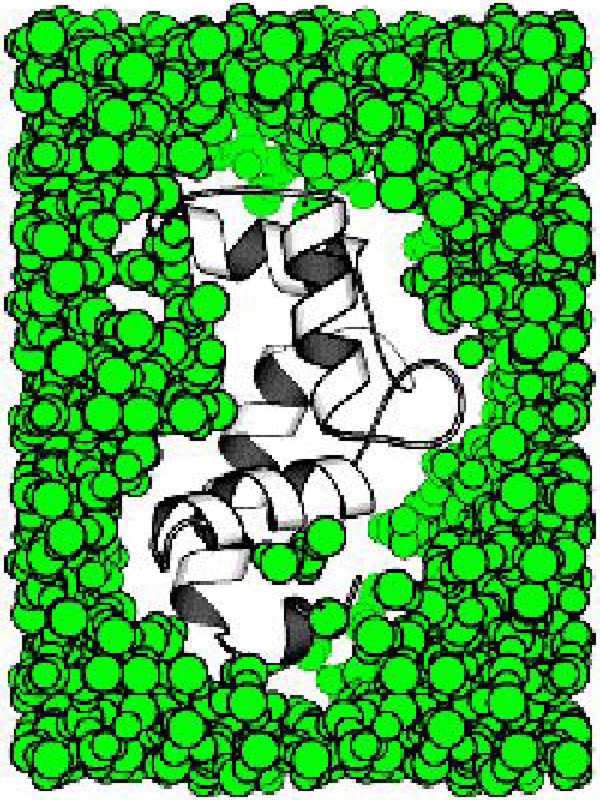
van der Waals interactions, Electrostatics

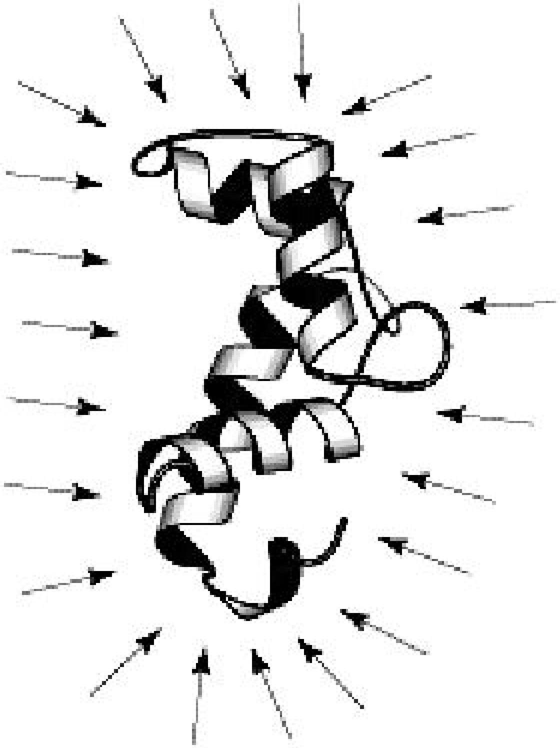
Solvent

Most difficult

Solvent

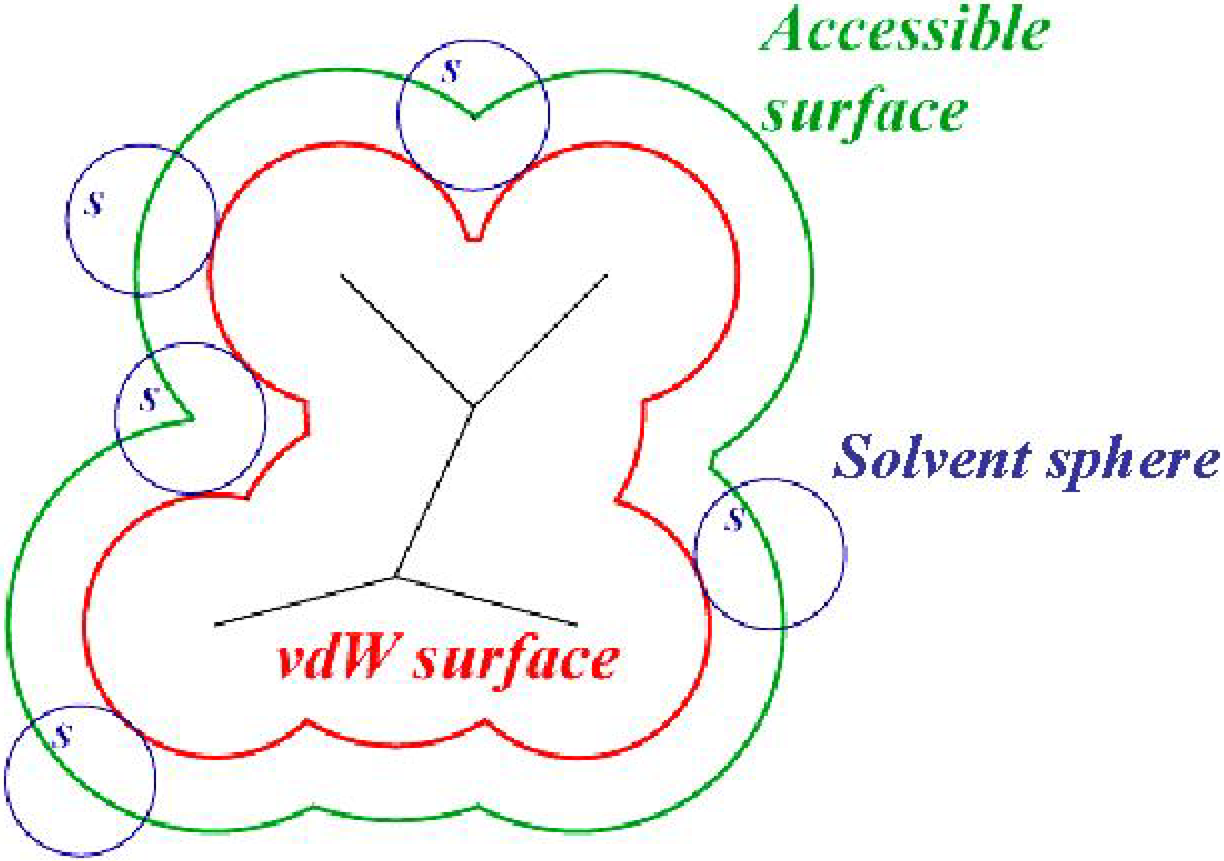
Explicit or Implicit ?





$$G_{sol} = G_{pol} + G_{cav} + G_{vdW}$$

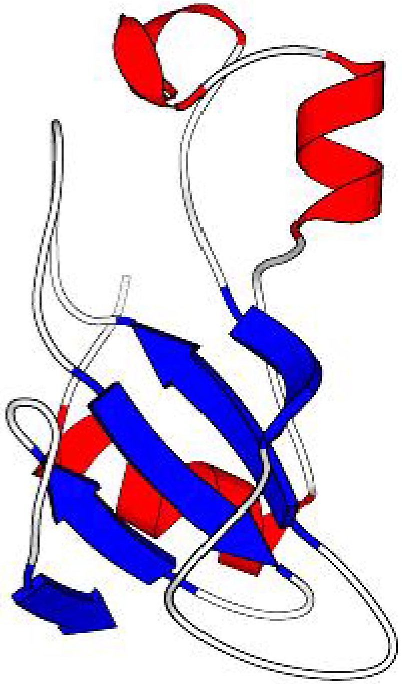
Solvation Potential



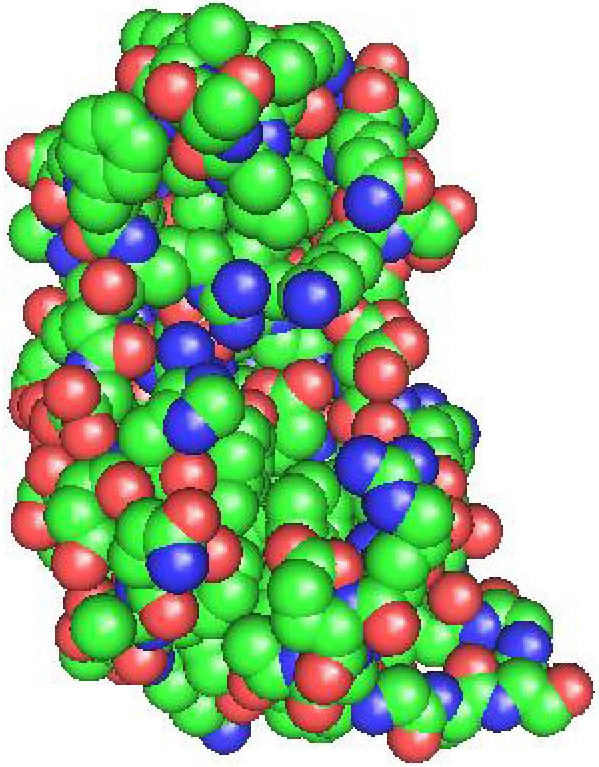
$$G_{cav} + G_{vdW} = \sum_{k=1}^{N} \sigma_k SA_k$$

Need Surface and Volume

Geometry of Protein Structure

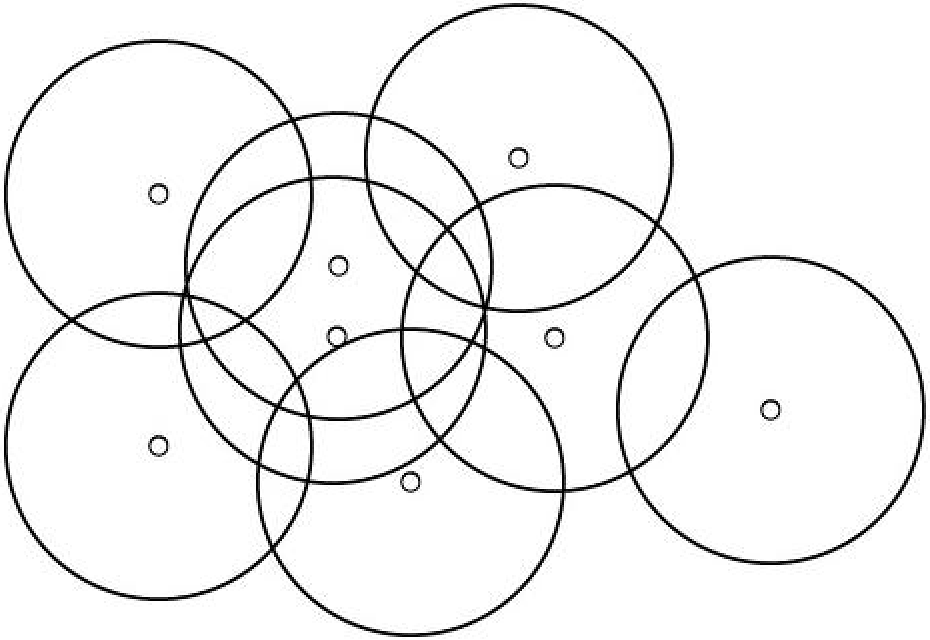


Cartoon



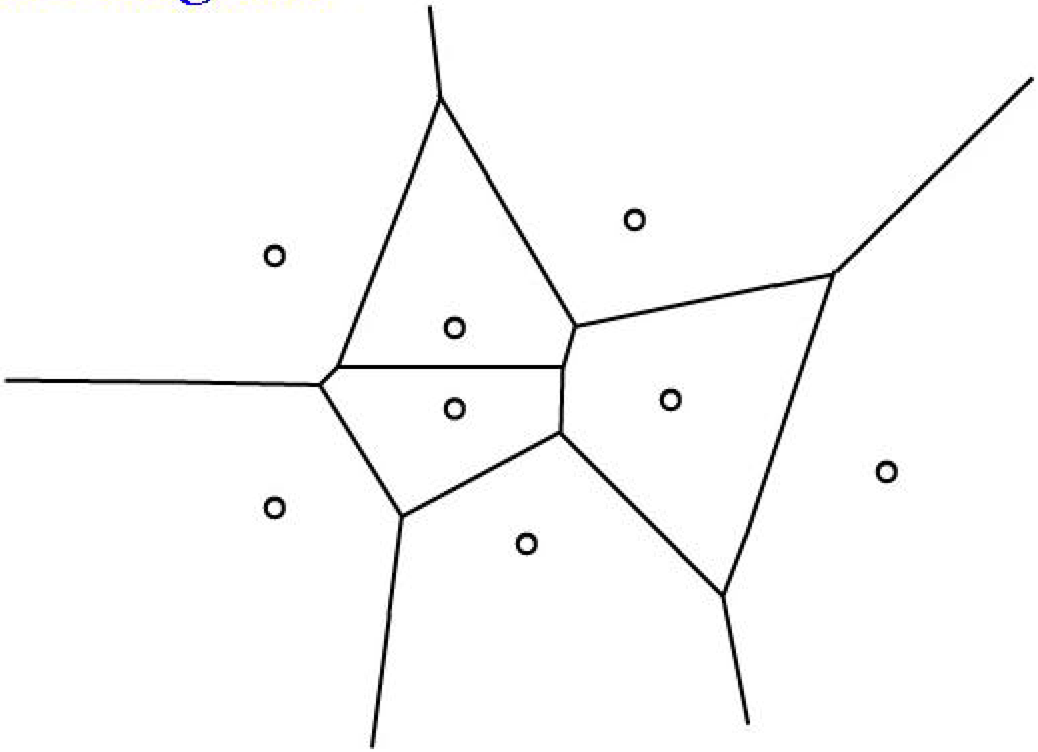
Space-filling Model

Computing the Surface Area and Volume of a Union of Balls



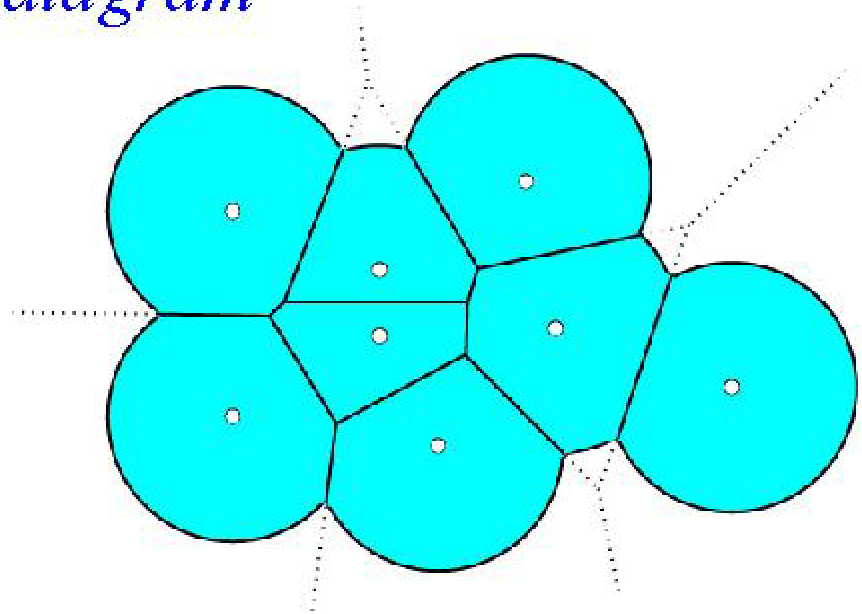
Computing the Surface Area and Volume of a Union of Balls

Power Diagram

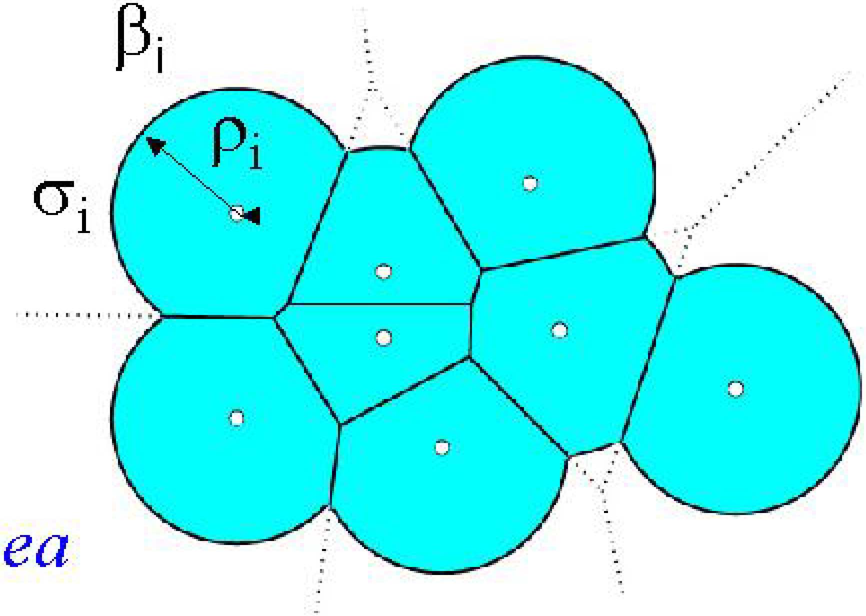


Computing the Surface Area and Volume of a Union of Balls

Decomposition of the Space-filling diagram



Computing the Surface Area and Volume of a Union of Balls



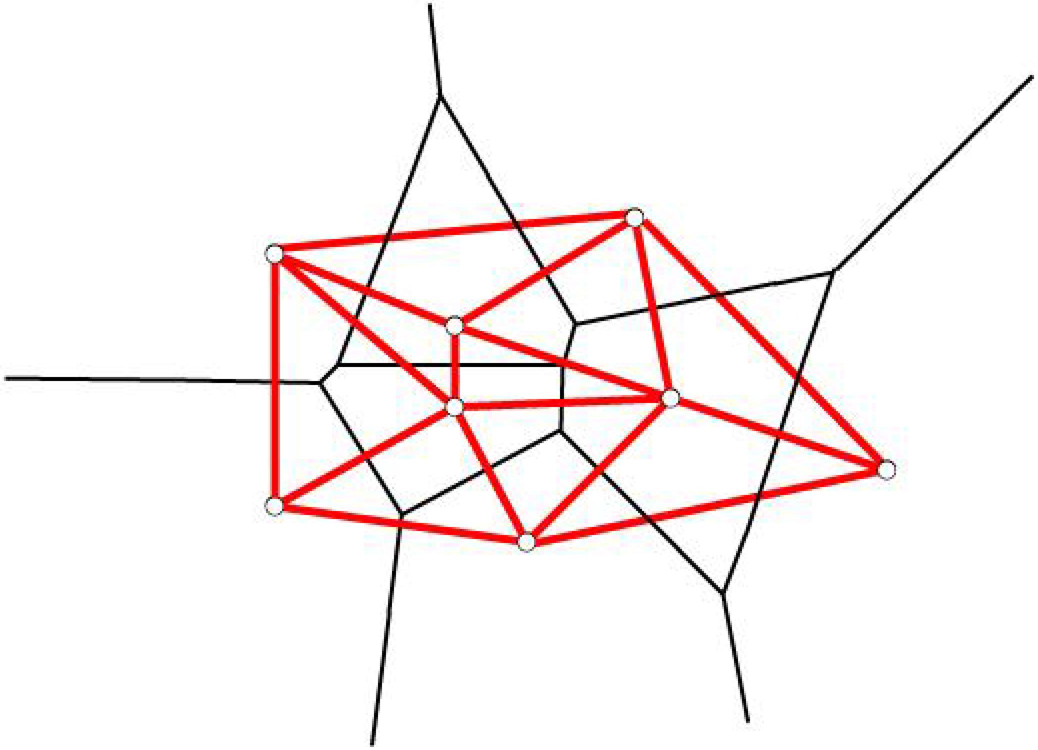
Surface Area

$$A = 4\pi \sum_{i=1}^{N} \rho_i^2 \sigma_i$$

Volume

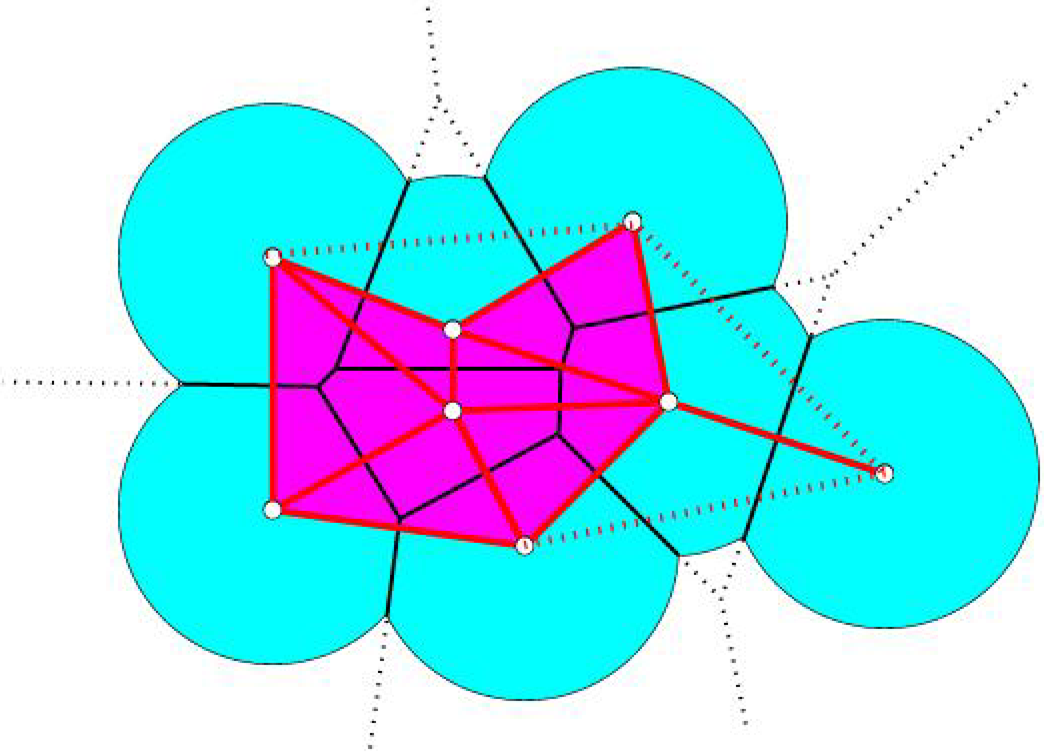
$$V = \frac{4\pi}{3} \sum_{i=1}^{N} \rho_i^3 \beta_i$$

Computing the Surface Area and Volume of a Union of Balls



The weighted Delaunay triangulation is the dual of the power diagram

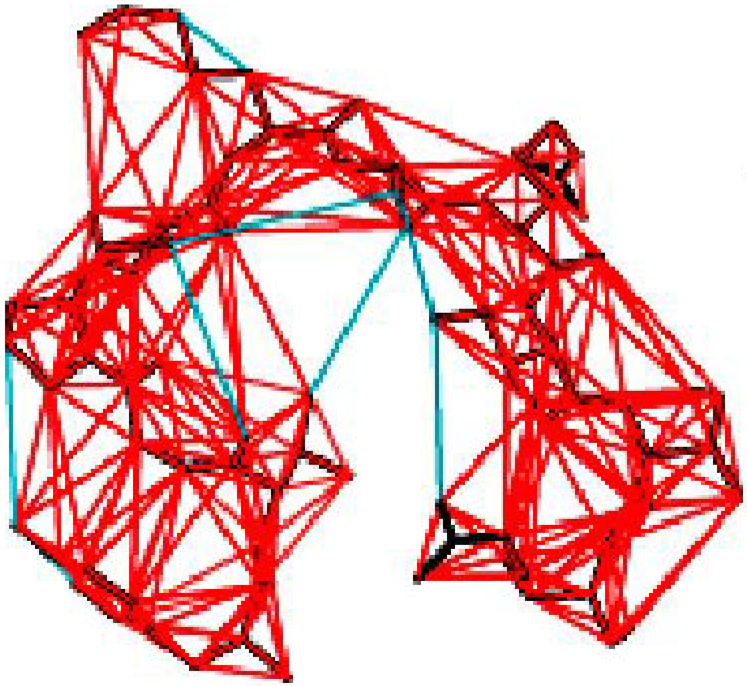
Computing the Surface Area and Volume of a Union of Balls



The dual complex K is the dual of the decomposition of the space-filling diagram

Computing the Surface Area and Volume of a Union of Balls

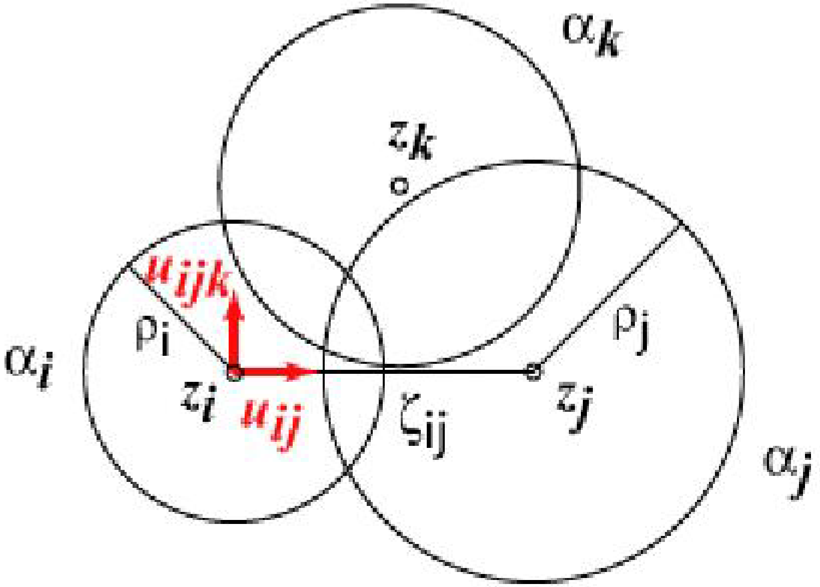
Inclusion-exclusion formulas for surface and volume



$$Area\left(\mathrm{Y}\,B\right) = \sum_{X \in K} (-1)^{\dim X}\, Area\left(\mathrm{I}\;\; X\right)$$

$$Vol\left(\mathrm{Y}\,B\right) = \sum_{X \in K} (-1)^{\dim X}\, Vol\left(\mathrm{I}\;\; X\right)$$

Weighted Area Derivative

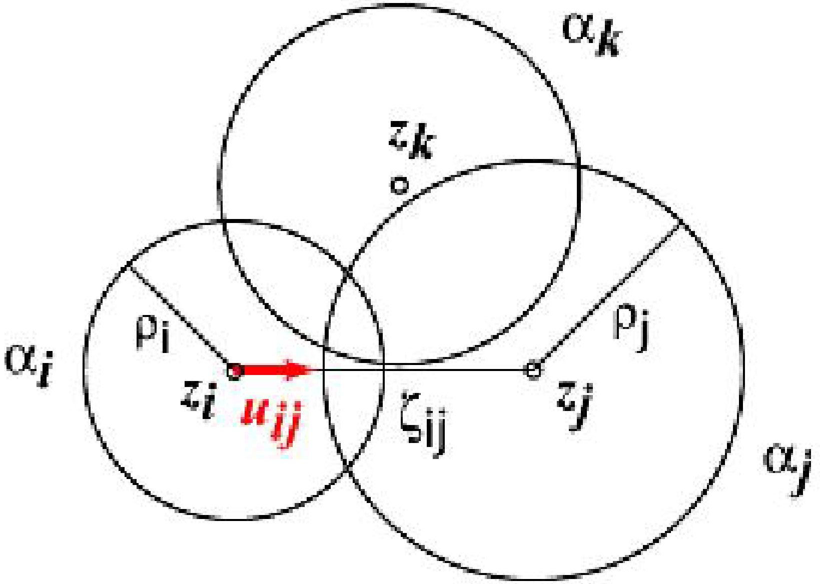


$$\begin{bmatrix} e_{3i-2} \\ e_{3i-1} \\ e_{3i} \end{bmatrix} = \sum_j \left(\sigma_{ij} \bullet e_{ij} + \sum_k \beta_{ijk} \bullet e_{ijk} \right)$$

$$e_{ij} = \pi \left[\left(\alpha_i \rho_i + \alpha_j \rho_j \right) - \left(\alpha_i \rho_i - \alpha_j \rho_j \right) \frac{\rho_i^2 - \rho_j^2}{\zeta_{ij}^2} \right] u_{ij}$$

$$e_{ijk} = 2 \rho_{ijk} \frac{\alpha_i \rho_i - \alpha_j \rho_j}{\zeta_{ij}} u_{ijk}$$

Weighted Volume Derivative



$$\begin{bmatrix} w_{3i-2} \\ w_{3i-1} \\ w_{3i} \end{bmatrix} = \sum_j S_{ij} \left(w_{ij} \bullet u_{ij} + x_{ij} \bullet v_{ij} \right)$$

$$w_{ij} = \left[\frac{\alpha_i + \alpha_j}{2} + \frac{(\alpha_i - \alpha_j)(\rho_i^2 - \rho_j^2)}{2\zeta_{ij}^2} \right]$$

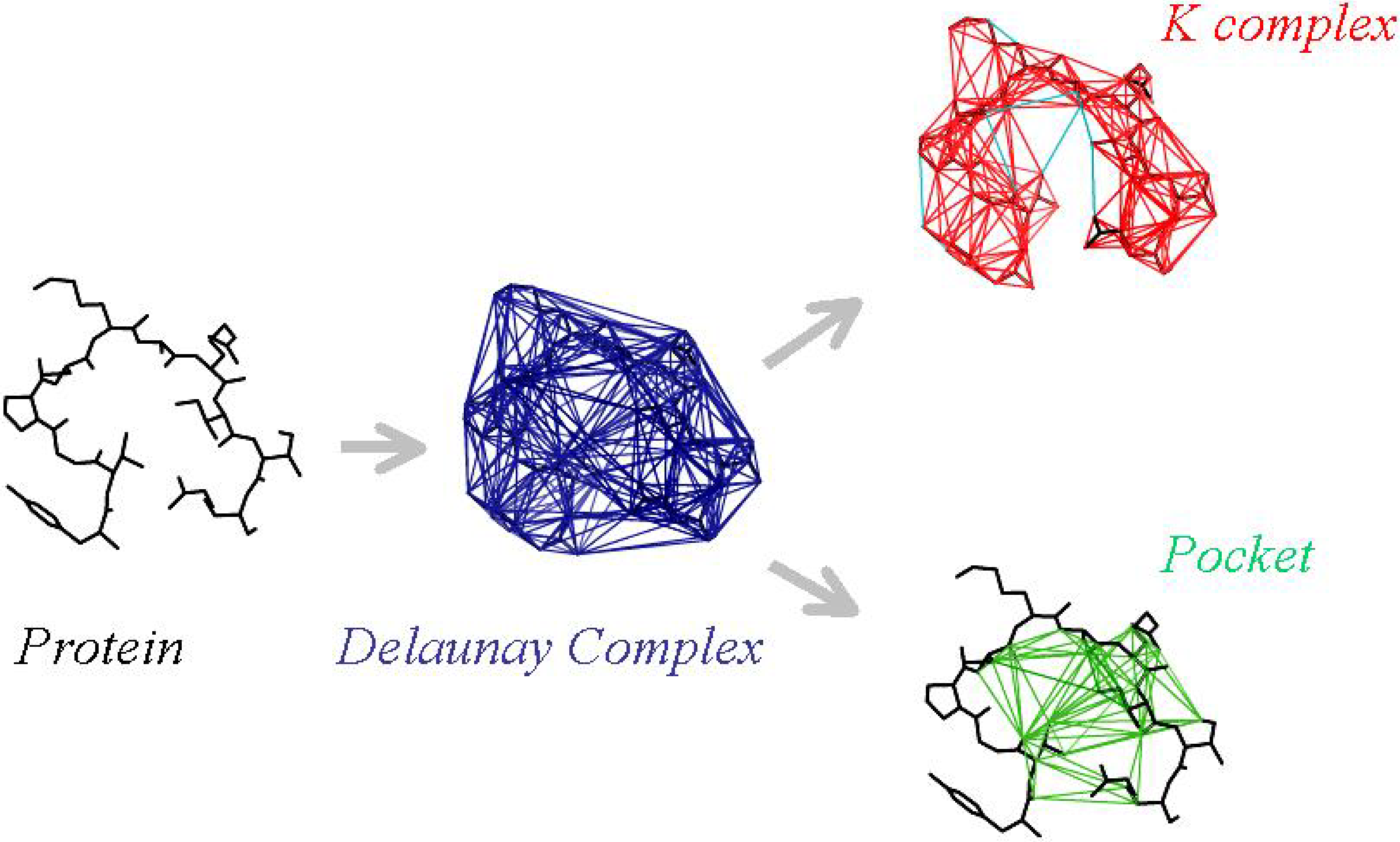
$$x_{ij} = \frac{2(\alpha_i - \alpha_j)}{3\zeta_{ij}}$$

S

S_{ij}: Area of facet (i,j)

V_{ij}: Average vector from center to boundary of facet (i,j)

Computing the Surface Area and Volume of a Protein



K complex

Protein

Delaunay Complex

Pocket

http://csb.stanford.edu/koehl/ProShape/download.php

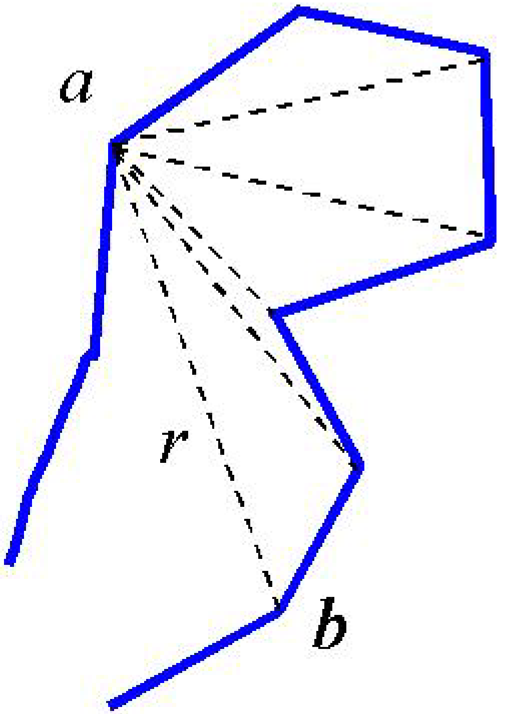
Software: ProGeom

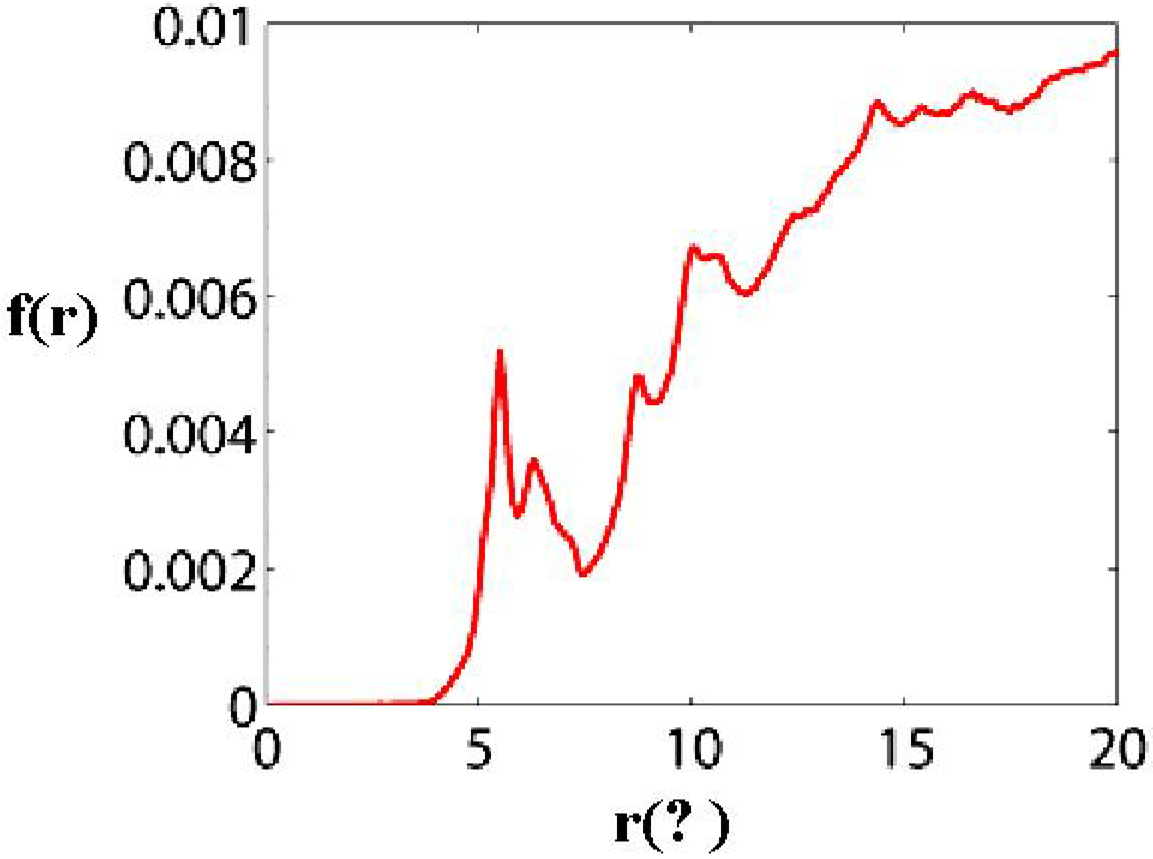
	Regular Triangulation	*Dual Complex*	*Inclusion-Exclusion*
Surface Area	0.19	0.06	0.05
Surface Area, Derivatives	0.19	0.06	0.08
Volume	0.19	0.06	0.18
Volume, Derivatives	0.19	0.06	0.29

System: 3740 balls (corresponding to a protein with 492 residues)
Computing time is seconds, on a Athlon 1.8 Ghz PC computer)

(http://csb.stanford.edu/koehl/ProShape/download.php)

Statistical Potentials

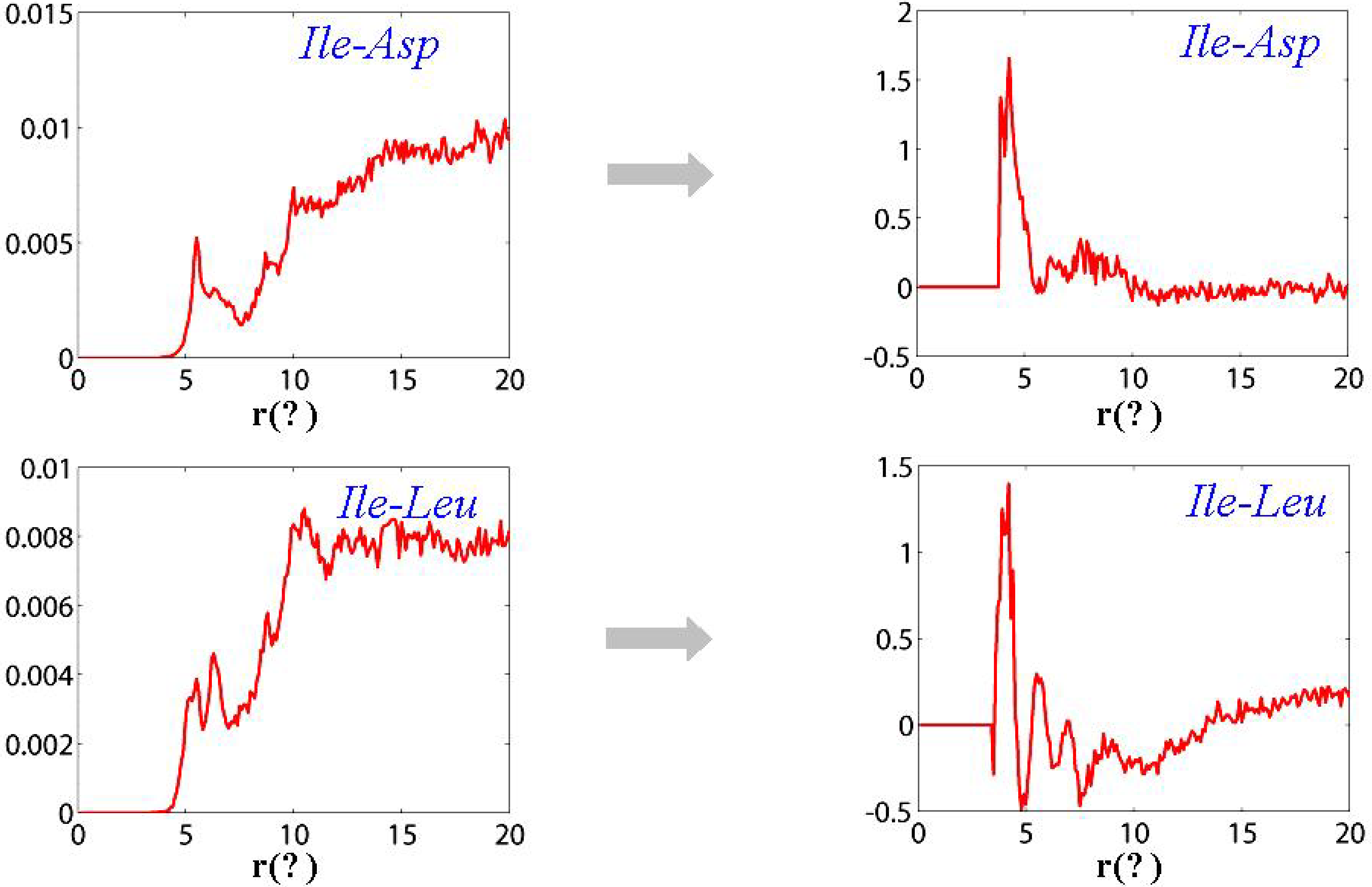




$$E(a,b,r) = -\ln\left(\frac{P_{(a,b)}(r)}{P(r)}\right)$$

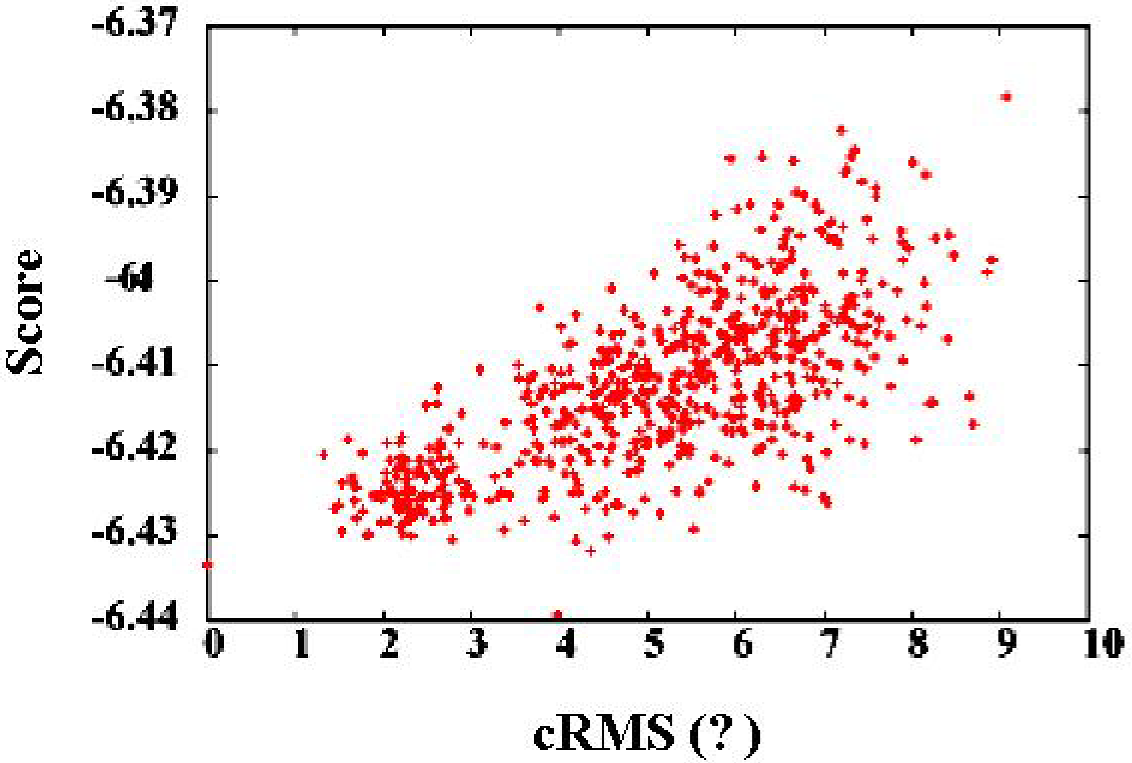
Counts

Energy



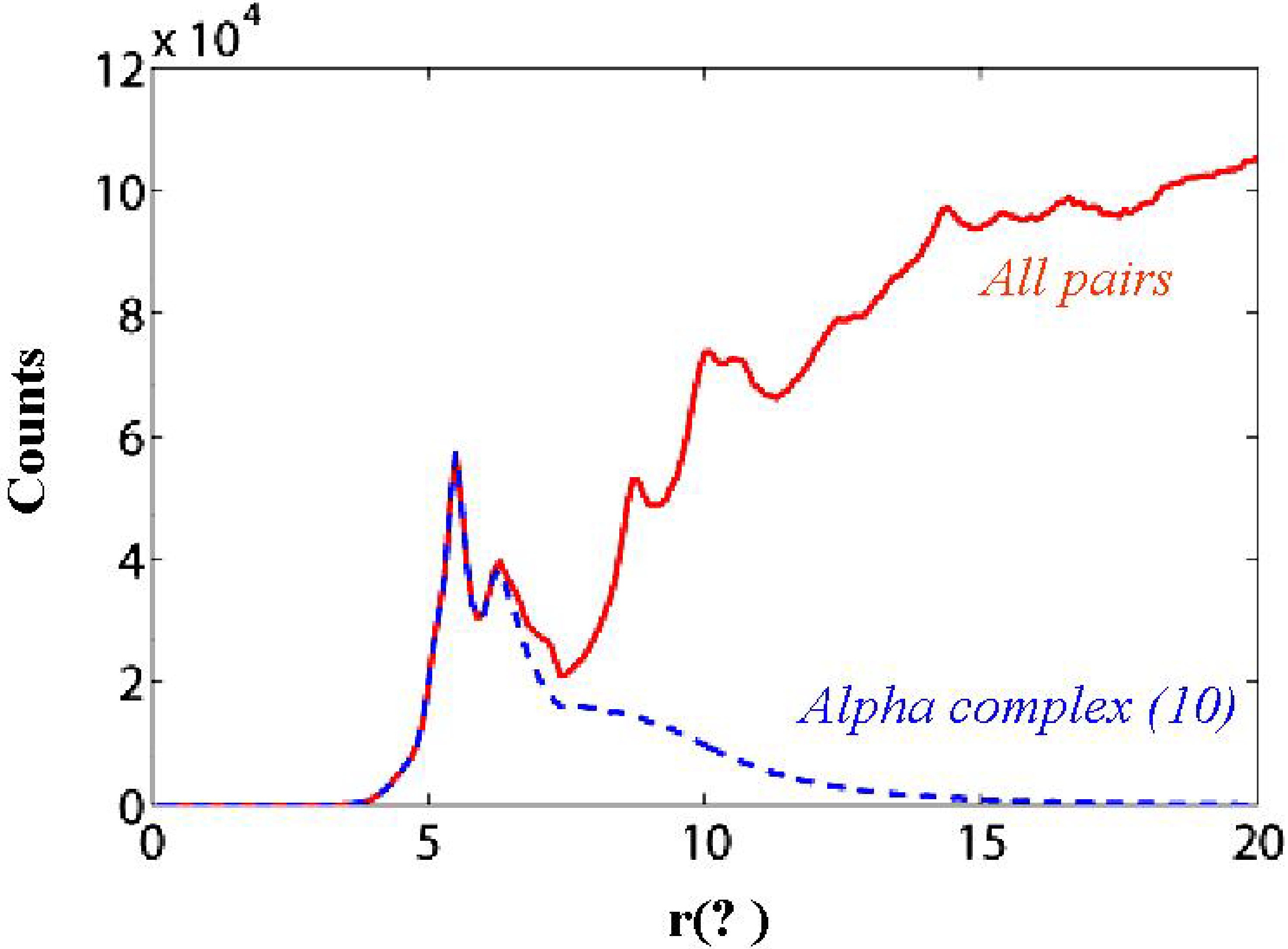
The Decoy Game
Finding near native conformations

1CTF

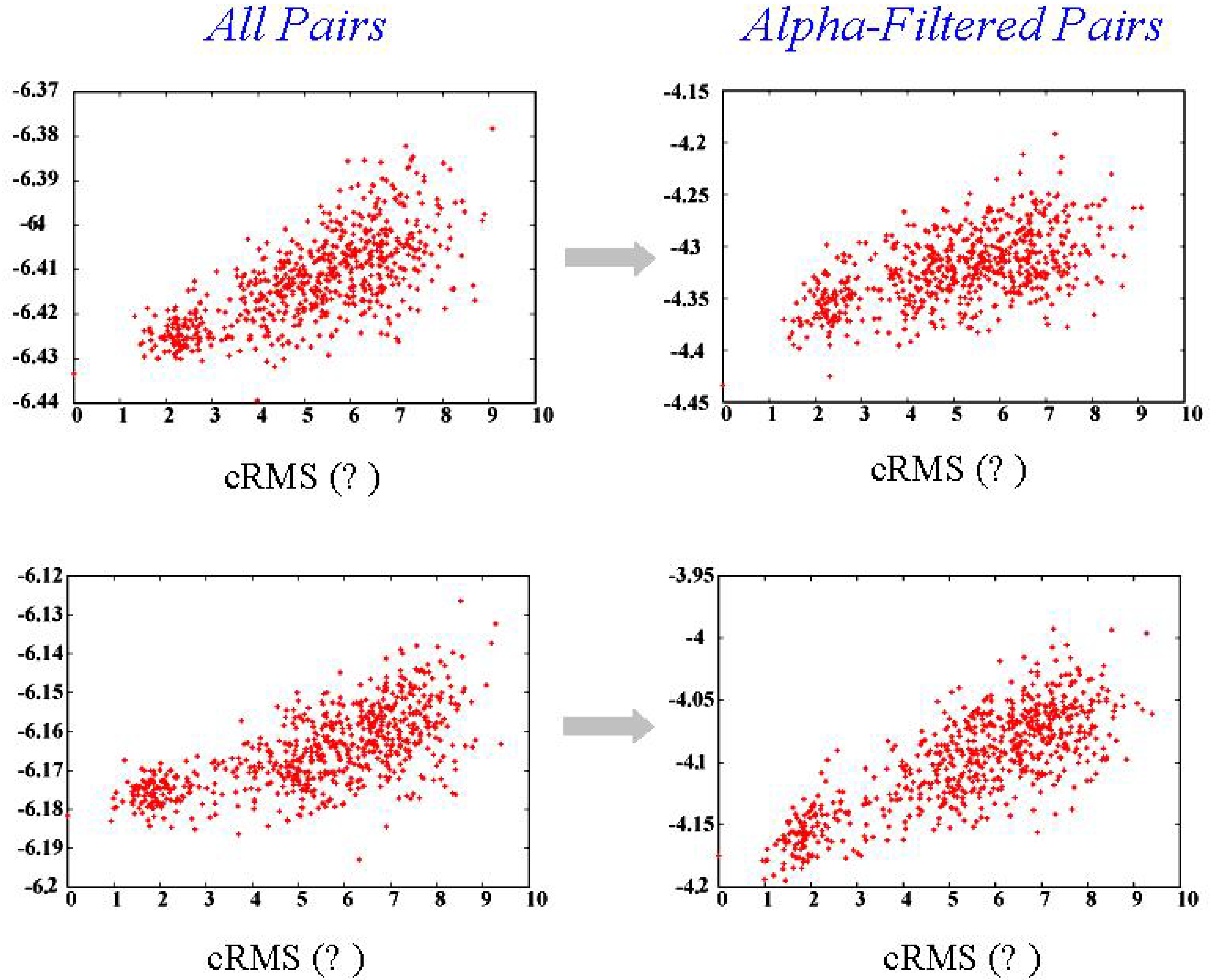


$$E = \sum_{i<j} E(i,j) = -\sum_{i<j} \ln\left(\frac{P(a_i,a_j,r_{ij})}{P(r_{ij})}\right)$$

Geometric Filtering of the Residue Pairs



Filtering Does Not Reduce Performance of PMF



Outline

- Introduction

  What is a Protein?

- Protein Energy Functions

  Computational Geometry Tools

- Classifying Proteins

  The Shapes of Protein Structures

Protein Structure Space

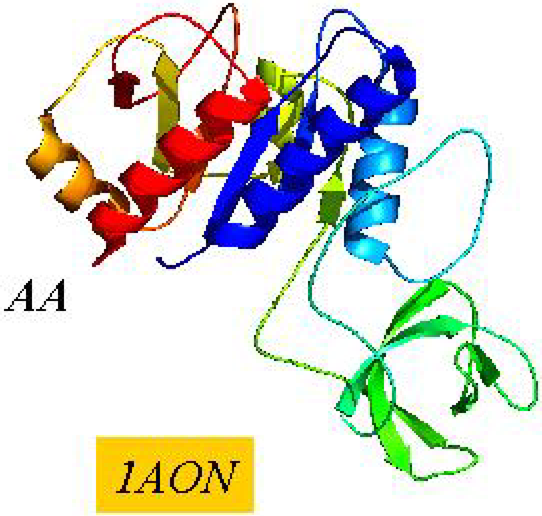
1CTF

68 AA

1TIM

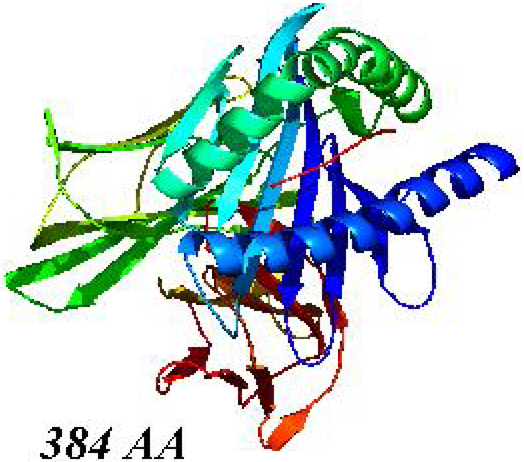
247 AA

1K3R



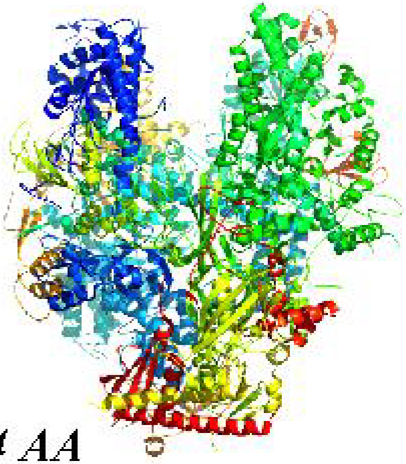
268 AA

1A1O



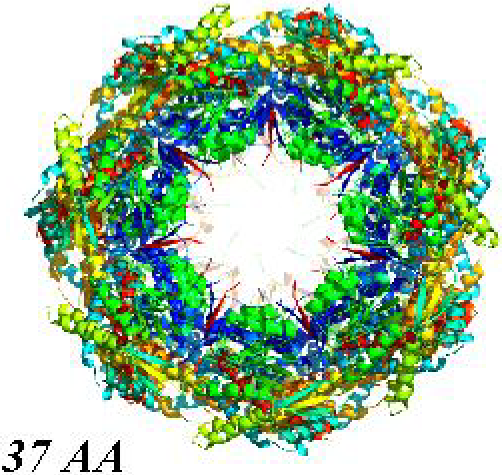
384 AA

1NIK



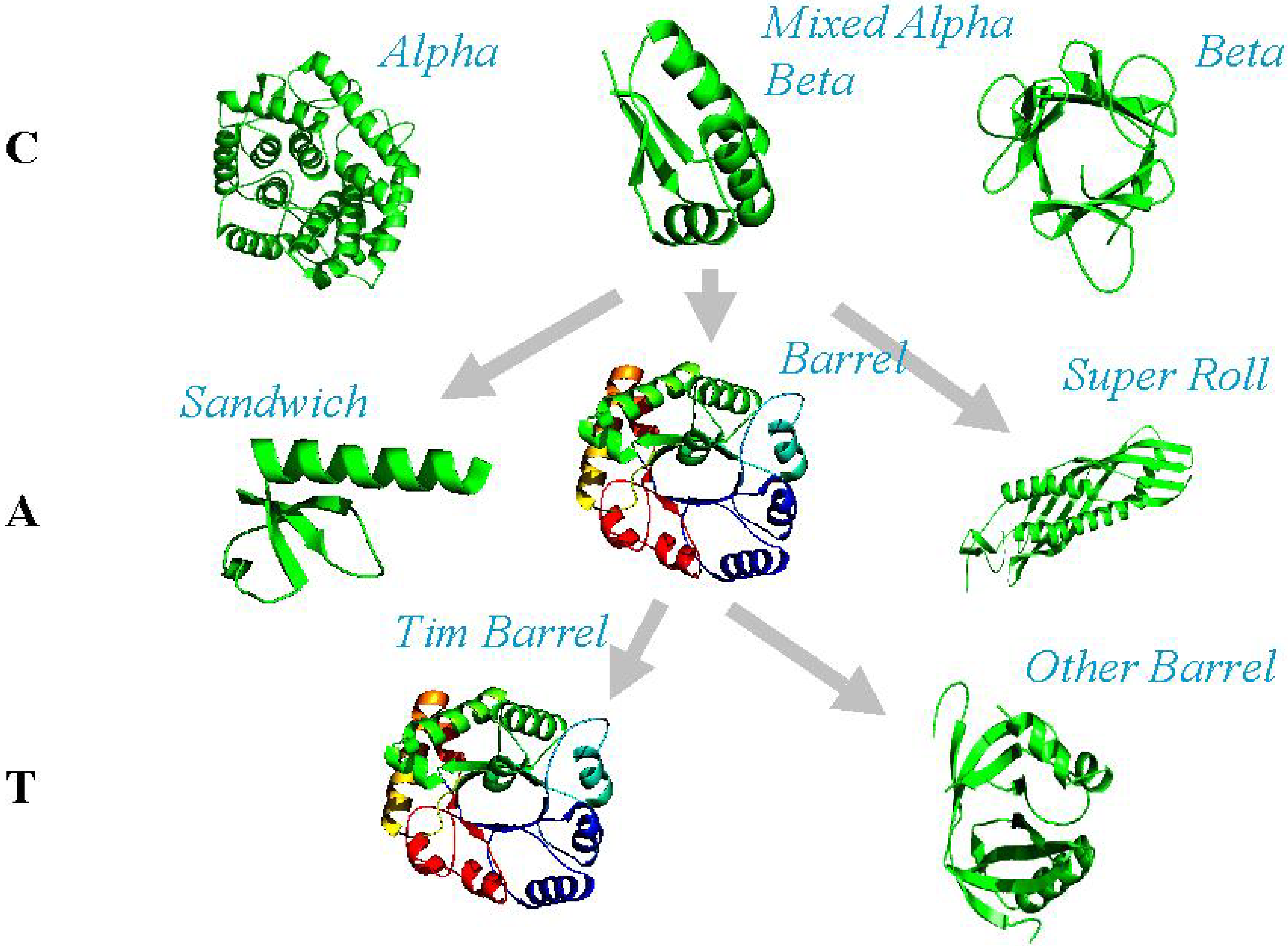
4504 AA

1AON



8337 AA

Classification of Protein Structure: CATH



C

Alpha

Mixed Alpha Beta

Beta

A

Sandwich

Barrel

Super Roll

T

Tim Barrel

Other Barrel

Protein Structure Similarity

Test set

2,930 proteins out of 23,000 proteins in PDB

No sequence similarity (Fasta E-value < e-4)

Reference structural similarity defined from CATH

769 folds

104,000 pairs of similar structures out of 4,600,000 pairs

Projecting Protein Structure Space

$$D = \begin{bmatrix} 0 & \dots & d_{1N} \\ \dots & 0 & \dots \\ d_{N1} & \dots & 0 \end{bmatrix} \longrightarrow G = X^T X \longrightarrow X$$

Distance Matrix

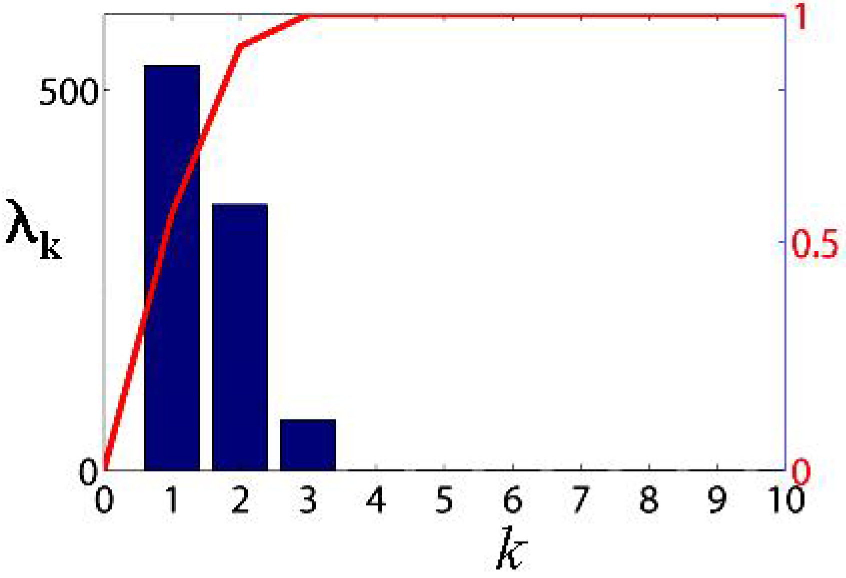
Metric Matrix

Points in Space

Projecting Protein Structure Space

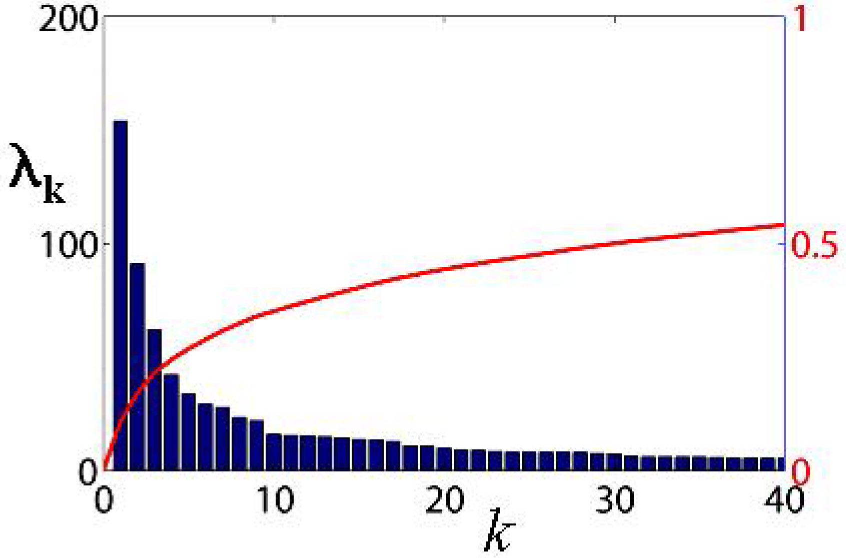
Class

$$D = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & \dots \\ 1 & \dots & 0 \end{bmatrix}$$



Fold

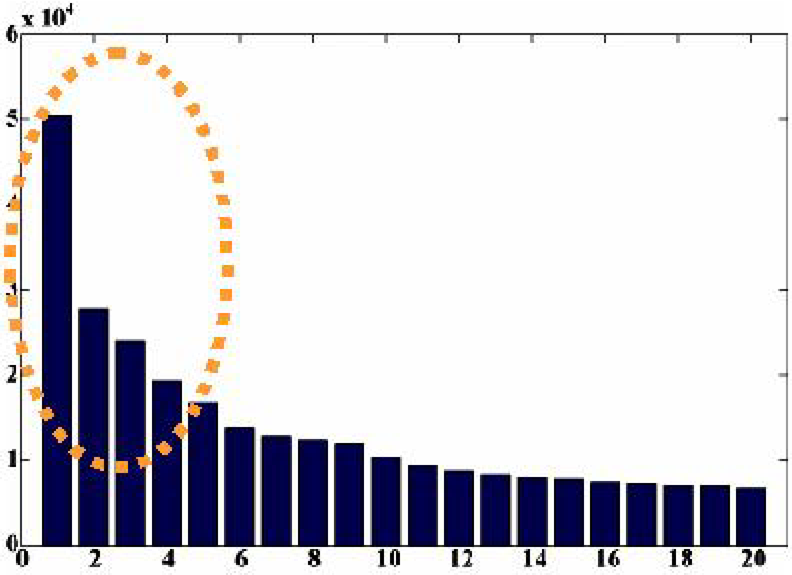
$$D = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & \dots \\ 0 & \dots & 0 \end{bmatrix}$$



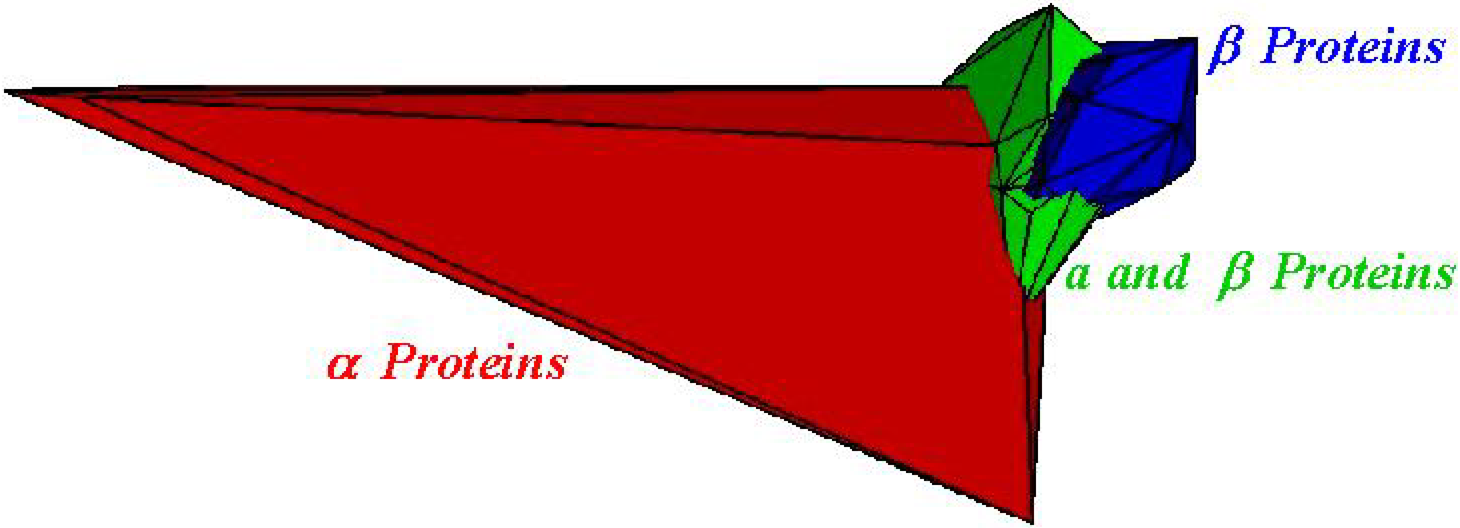
Protein Structure Classes

Measure of Structure Similarity: cRMS after Optimal Superposition (Structal)

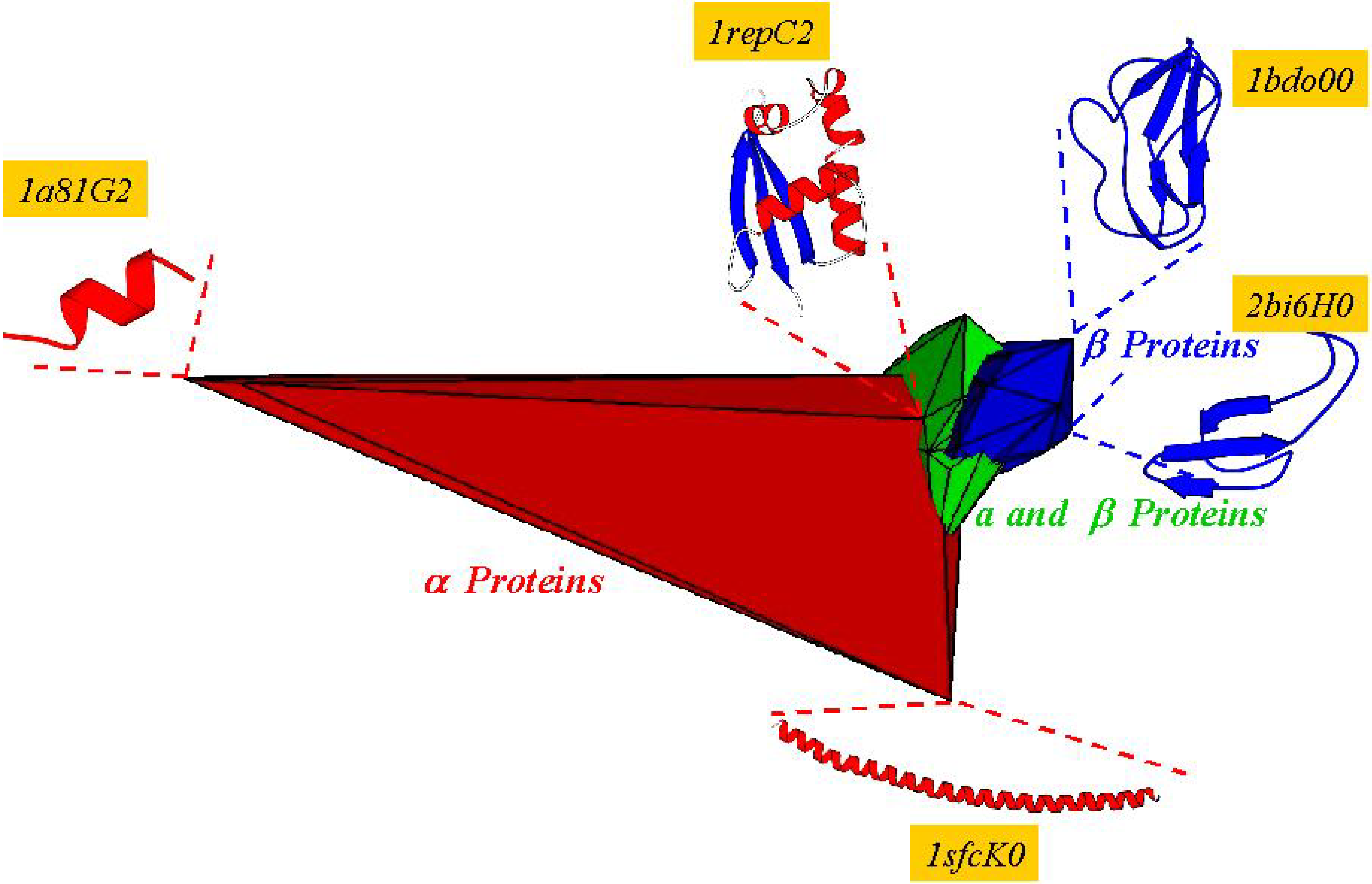
Eigenvalues of the Metric Matrix:



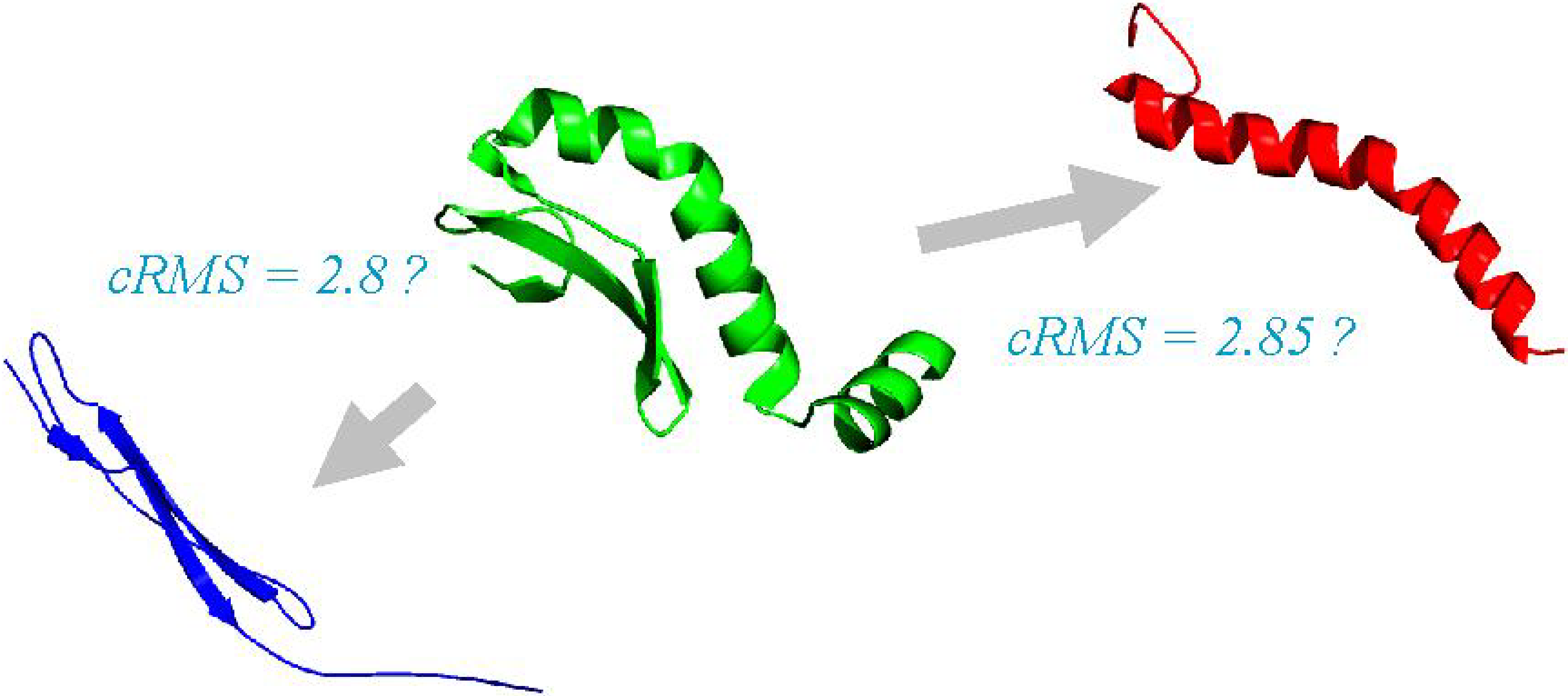
A Picture of the Protein Structure Space



A Picture of the Protein Structure Space



cRMS is not a Metric

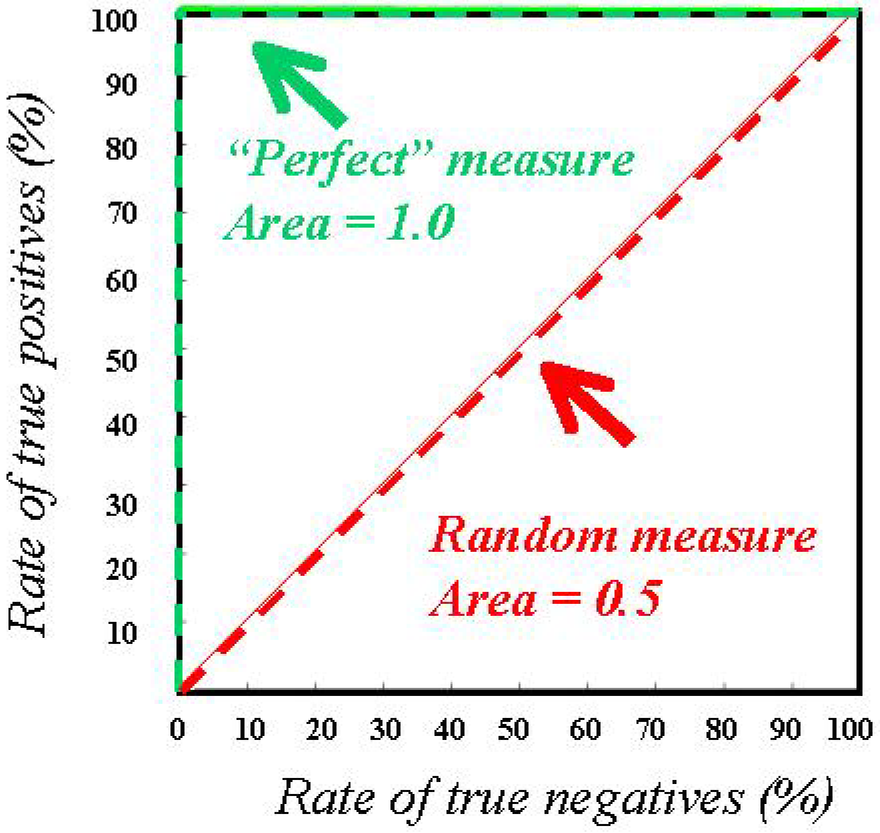


cRMS = 2.8 ?

cRMS = 2.85 ?

Protein Fold Space
ROC Analysis

(Receiver Operating Characteristic)



Protein Fold Space
ROC Analysis

(Receiver Operating Characteristic)

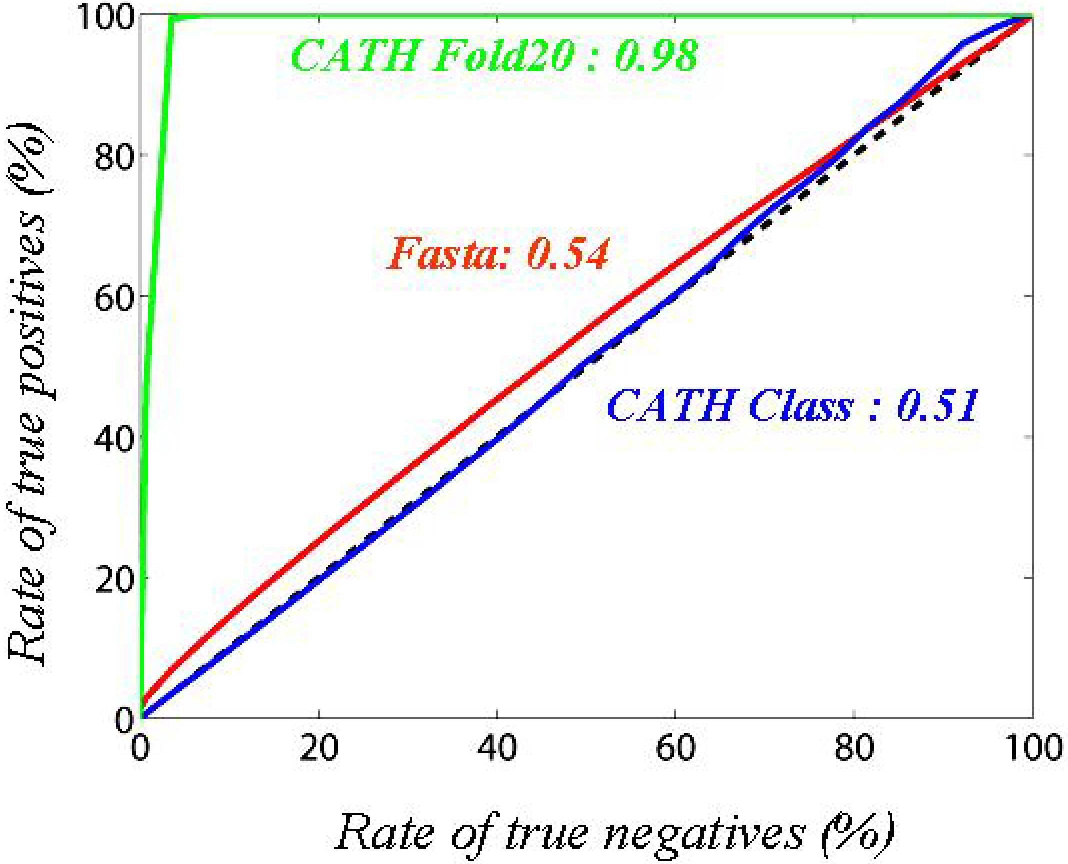
True positives

pairs of proteins that belong to the same T class of CATH

True negatives

pairs of proteins that belong to the same C class, but not the same T class.

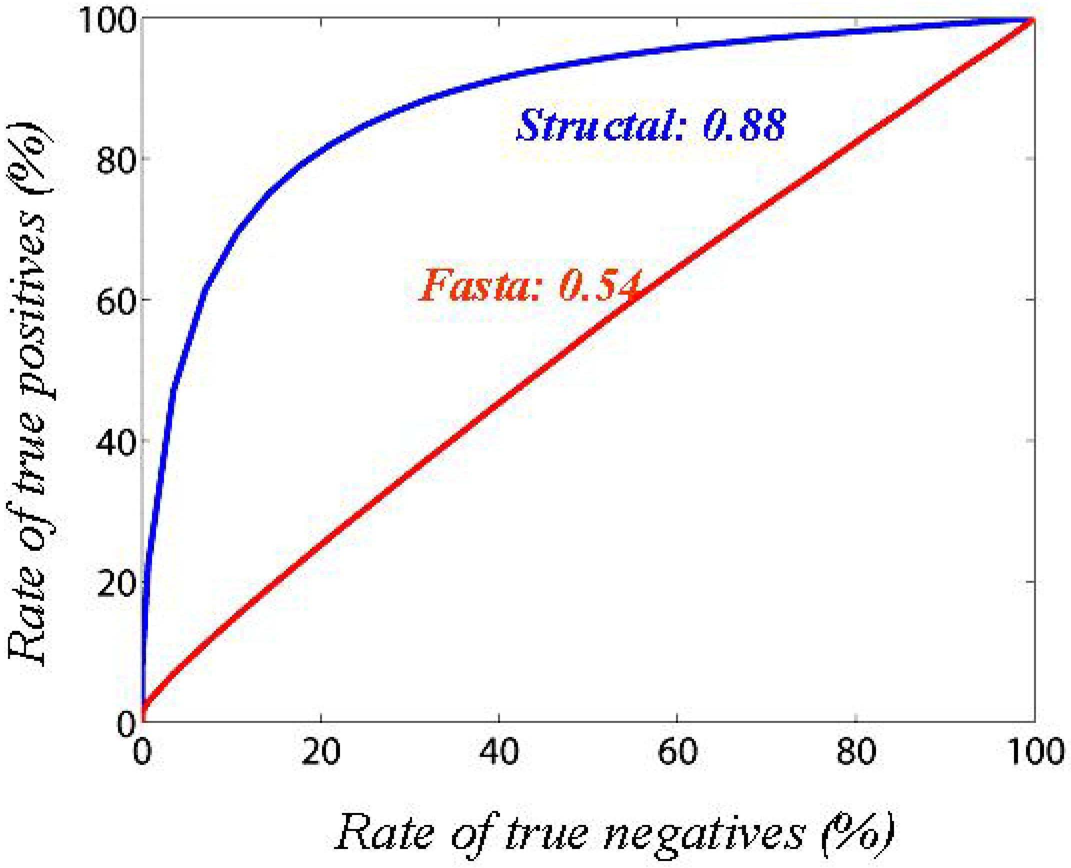
Protein Fold Space



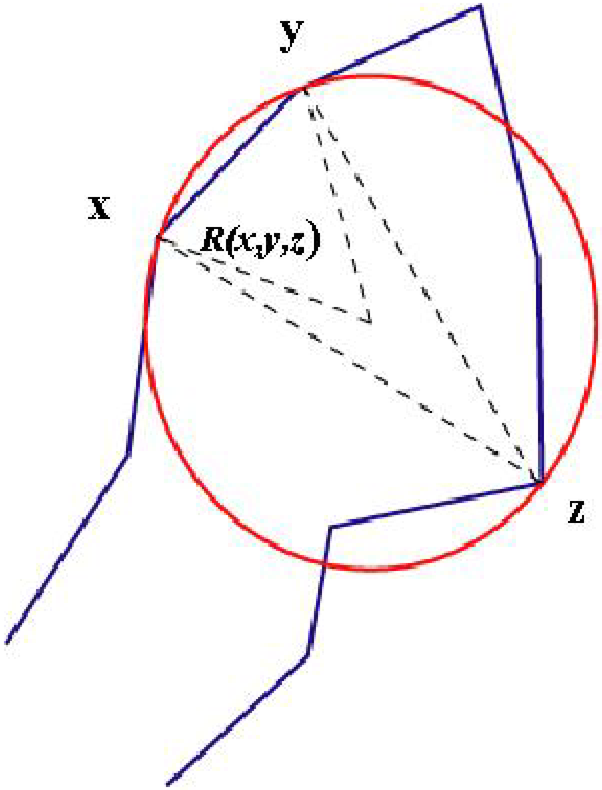
Fold20: first 20 coordinates derived from the CATH fold matrix

CATH class: first 3 coordinates derived from the CATH class matrix

Protein Fold Space



Protein Structure Features



$$R(x,y,z) = \frac{d(x,y)}{2\left|\sin\left(\hat{z}\right)\right|}$$

Global radius of curvature:

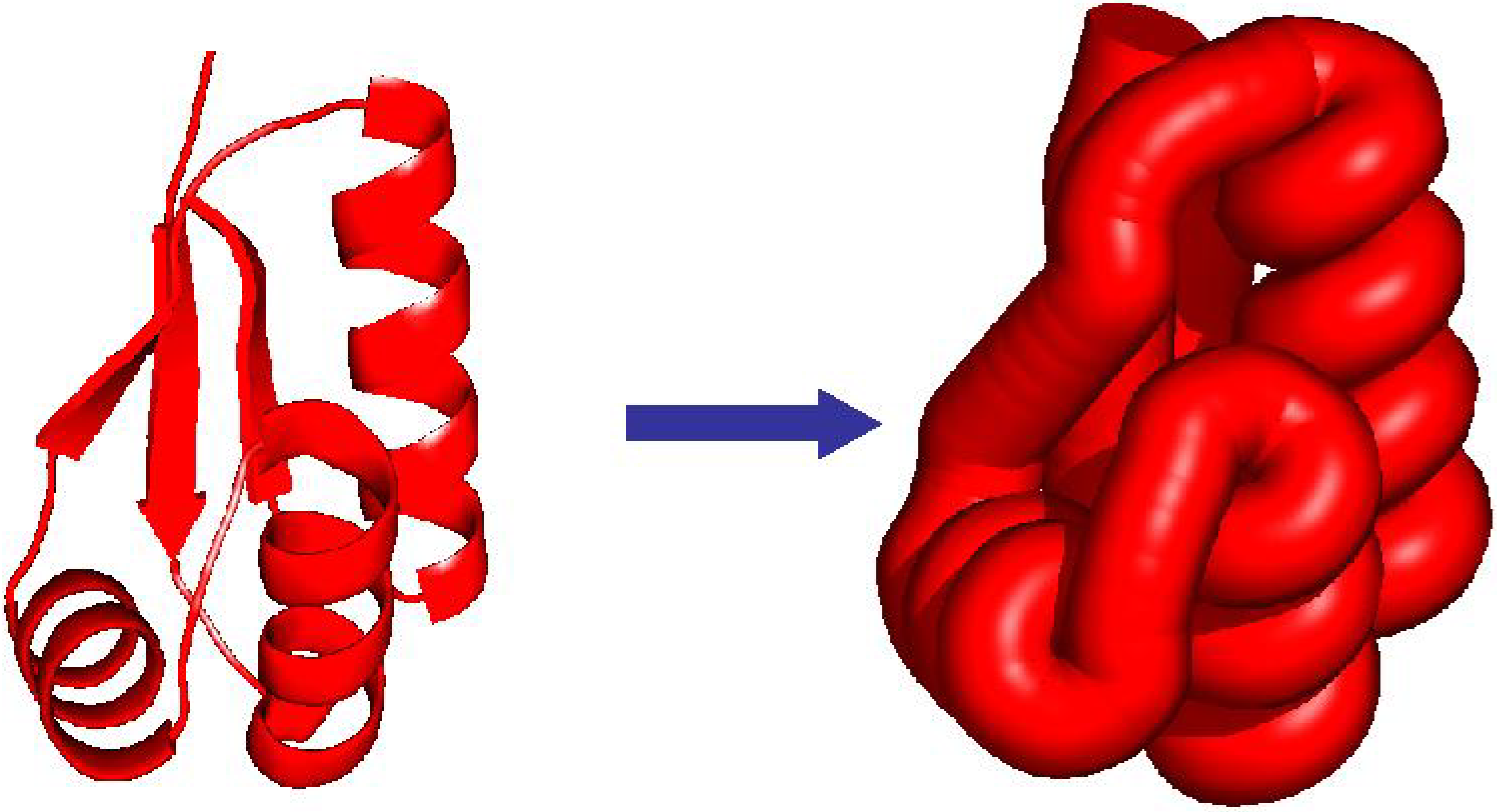
$$\boxed{\rho(x) = \min_{(y,z)}\left\{R(x,y,z)\right\}}$$

Thickness:

$$\boxed{\Delta = \min_{x}\left\{\rho(x)\right\}}$$

(Gonzalez & Maddocks, PNAS, 1999, 96:4769)

Thickness of a protein structure



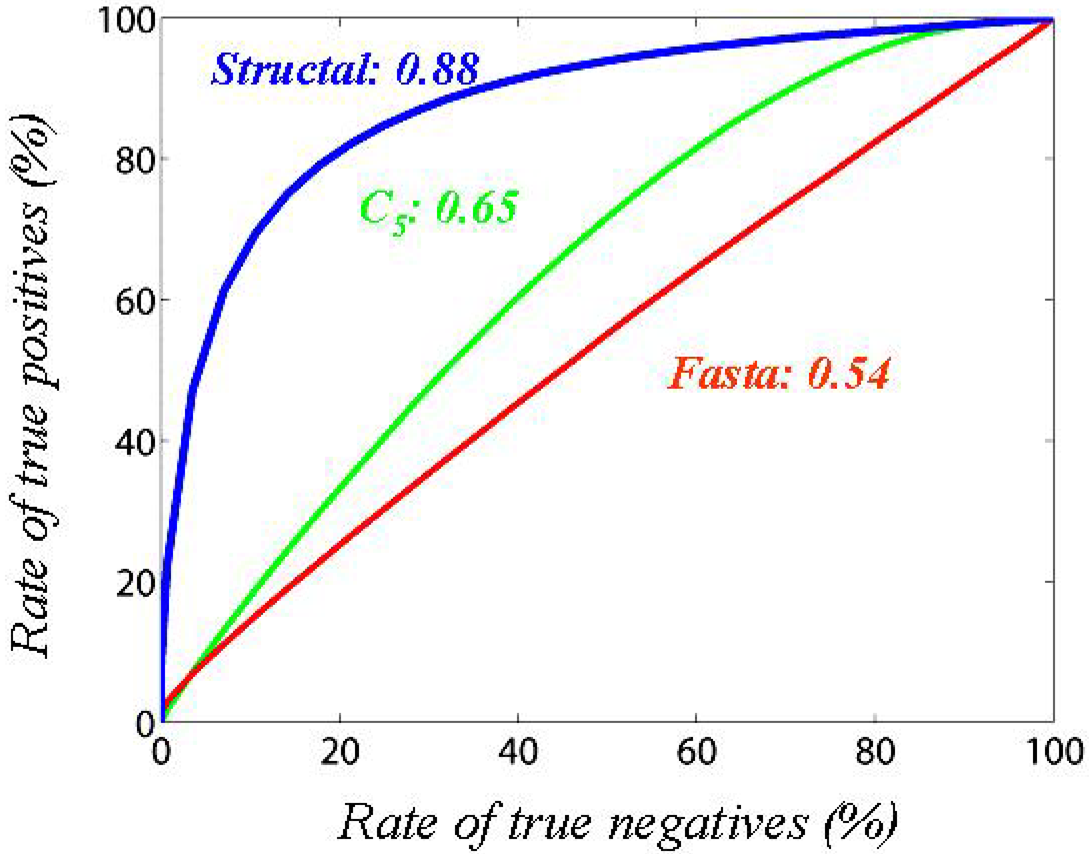
Δ= 2.60 ?

Curvature Feature Vector

$$U_p = \left(\iiint \frac{1}{R(x, y, z)^p} dC_x dC_y dC_z \right)^{1/p}$$

$$C_5 = \begin{bmatrix} U_1 & U_2 & U_3 & U_4 & U_5 \end{bmatrix}$$

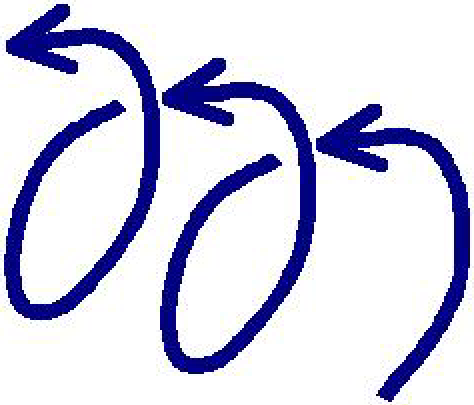
Performance of the Curvature Feature Vector



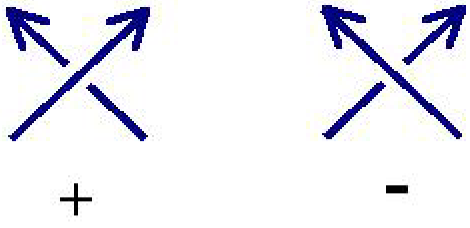
Curvature vector performs better than fasta.

Needs more features to match Structal.

Protein Structure Features: Writhing



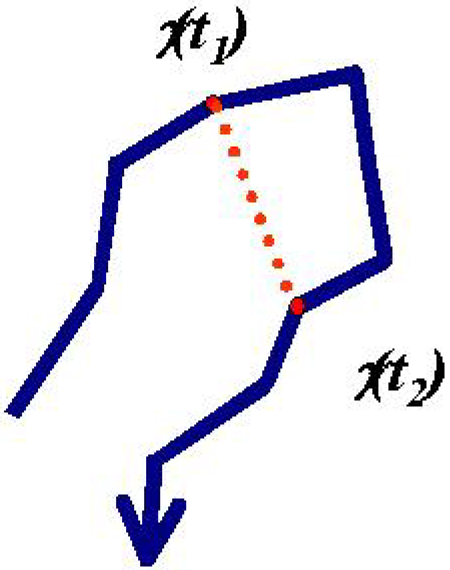
Sign of Crossing



Writhing Number

$$Wr_1 = \frac{1}{4\pi} \iint_{\Delta^2} \omega(t_1, t_2)\, dt_1 dt_2$$

$$\omega(t_1, t_2) = \frac{\det(\gamma'(t_1), \gamma(t_1) - \gamma(t_2), \gamma'(t_1))}{|\gamma(t_1) - \gamma(t_2)|^3}$$

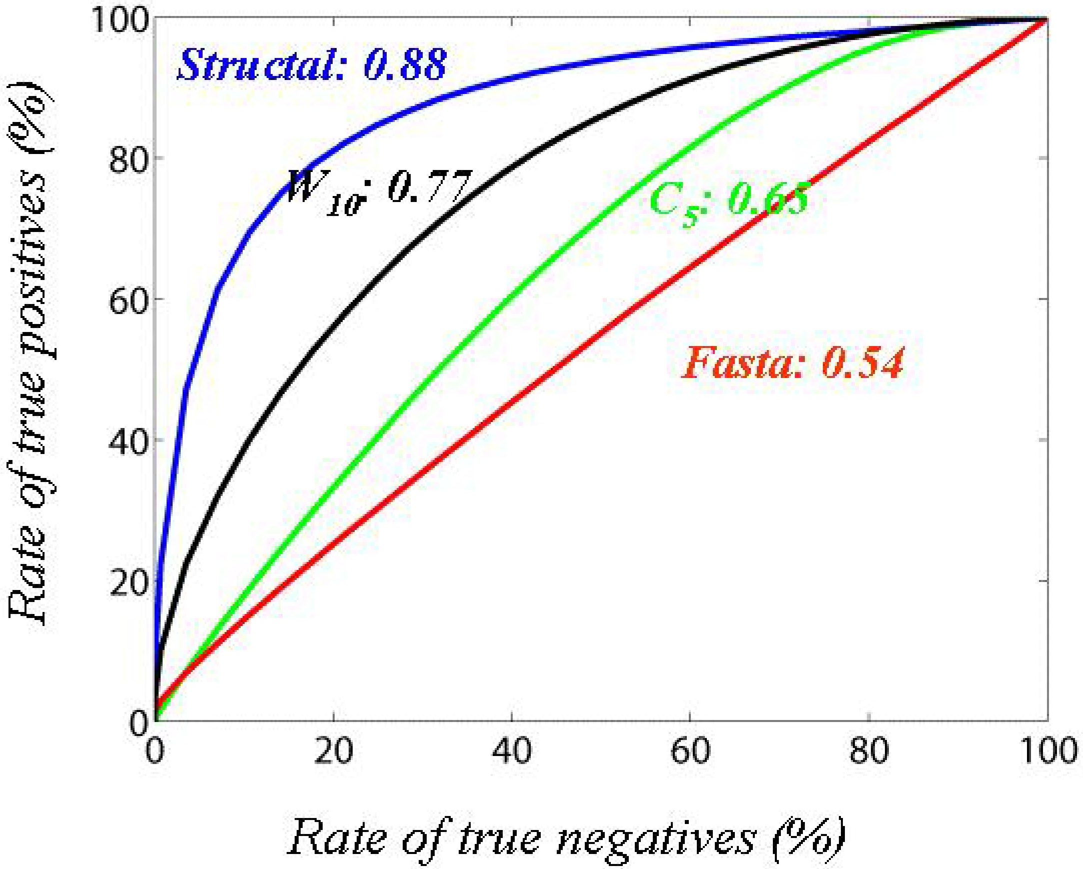


Writhe Feature Vector for Each Protein

$$W_{10} = [Wr_1 \quad |Wr_1| \quad Wr_{12} \quad |Wr_{12}| \qquad]$$

Fain and Røgen, PNAS, 100: 119 (2003)

Protein Structure Features: Writhing



W_{10} Writhe performs better than C_5 Curvature

Collaborators

- Leo Guibas
  Stanford University

- Herbert Edelsbrunner
  Duke University

- Michael Levitt
  Stanford University

- Afra Zomorodian
  Stanford University

- Rachel Kolodny
  Stanford University

Thank You