

New Kernels for Protein Structural Motif Discovery and Function Classification

Chang Wang
Stephen D. Scott

University of Massachusetts
University of Nebraska

ICML 2005

Three-dimensional structural analysis of proteins

- Very useful in determining function
- Rather than making global comparisons, we make numerous local comparisons, seeking small conserved substructures

Our results:

- New kernels for 3D protein structure analysis
 - More sensitive to Trx fold and faster than DALI and CE
- Application of kernels to motif discovery
- Application of kernels to function classification
- Families: oxidoreductases and cytochrome P450 reductases

New Kernels
for Protein
Structural
Motif
Discovery and
Function
Classification

Chang Wang
Stephen D.
Scott

Introduction

Outline

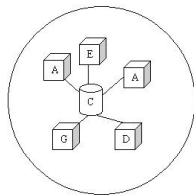
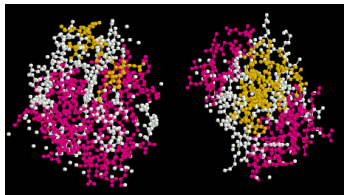
Our New
Kernels

Using our
Kernels

Experimental
Results

Conclusions

- Our new kernels: $K_{Pattern_Sim}$, K_{Redox_Func} , K_{3Dball}
- Using our kernels in motif discovery and function classification
- Experimental results
- Conclusions



We view each amino acid (*central AA*) as the center of a fixed-radius substructure in 3D space

All amino acids that lie in the ball (*outer AAs*) are also included in this substructure

New Kernels
for Protein
Structural
Motif
Discovery and
Function
Classification

Chang Wang
Stephen D.
Scott

Introduction

Outline

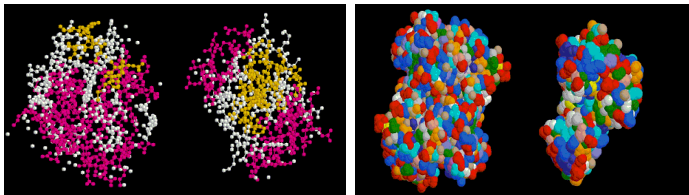
Our New
Kernels

K_{3Dball}
 $K_{Pattern_Sim}$
 K_{Redox_Fun}

Using our
Kernels

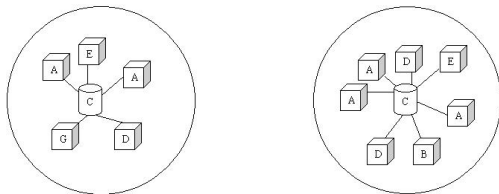
Experimental
Results

Conclusions



Now a protein is a 3D space filled with 3D balls.

Similarity of two proteins is defined by summing the similarities of their constituent balls



If two substructures share the same central amino acid, then their similarity is the number of pairs of outer amino acids shared by the two substructures, otherwise 0

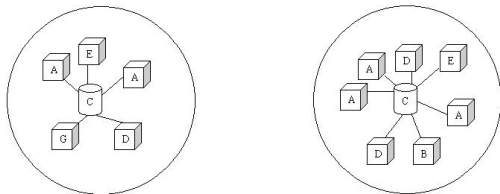
These two sub-structures share the same central amino acid (C), and four outer amino acids: A, A, D, E \Rightarrow similarity = 4

Theorem

K_{3Dball} is a true (positive semidefinite) kernel.

Intuition:

- K_{3Dball} captures both the overall shape info and the active site info
- If two proteins have similar shape and constitution, they should share many substructures (3D balls)
- Substructures corresponding to the active site should be very similar

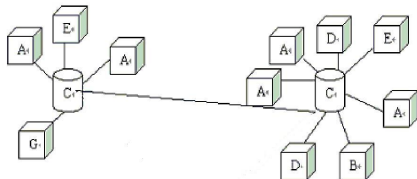


Similar to K_{3Dball} , but considers the distance from each outer amino acid to the central amino acid via an RBF kernel:

$$\exp\left(-\frac{\|dist(S_{outer}, S_{central}) - dist(T_{outer}, T_{central})\|^2}{\delta}\right)$$

For a set of proteins with similar functions, the most conserved 3D ball can be seen as a candidate for the active site

Not necessarily positive semidefinite, but works well in practice



Variation of $K_{Pattern_Sim}$ that models a complex active site

A 3D ball can have multiple central amino acids. The kernel considers the type of the outer amino acids, distances from outer amino acids to central amino acids, and distances between central amino acids.

Can be adapted to other known structural motifs

- 1 Select random set $\{P_1, \dots, P_n\}$ of proteins
- 2 Let $S_j^{P_i}$ be the radius- r ball centered at P_i 's j th AA
 - Represent P_i as $\{S_1^{P_i}, \dots, S_{m_i}^{P_i}\}$
- 3 For each $i' \neq i$, find substruc in $P_{i'}$ most similar to $S_j^{P_i}$
 - $S_j^{P_i}$'s *fitness* is sum of these:

$$fitness \left(S_j^{P_i} \right) = \sum_{i'=1, i' \neq i}^n \max_{1 \leq j' \leq m_{i'}} K_{PatternSim} \left(S_j^{P_i}, S_{j'}^{P_{i'}} \right)$$

- 4 Relatively high fitness \Rightarrow highly conserved
across $\{P_1, \dots, P_n\} \Rightarrow$ candidate structural motif

Used two machine learning techniques with our kernels:

- 1 Support vector machines (SVMs)
- 2 Variant of k nearest neighbor (k NN):
 - Given new (unlabeled) protein S to classify, compute similarities between S and all positive training proteins
 - Take the mean of the similarities of the top $k\%$ positive proteins most similar to S (use the same process for negative proteins)
 - If mean similarity between S and the positives is larger than that for the negatives, predict S to be positive, otherwise negative

New Kernels
for Protein
Structural
Motif
Discovery and
Function
Classification

Chang Wang
Stephen D.
Scott

Introduction

Outline

Our New
Kernels

Using our
Kernels

Experimental
Results

Structural Motif
Discovery

Structural
Classification

Function
Classification:
Leave One Out
Test

PDB Search

Evaluated:

- Structural motif discovery
- Structural classification
- Function classification
 - Leave-one-out test
 - PDB search

Used $K_{pattern_sim}$

Tested our active site identification method on thiol/disulfide oxidoreductases and P450 reductases

Used each AA in each protein as central AA of a substructure

Radius = 6 Å

Results:

- Successfully found the known active site in oxidoreductases: two cysteines and a proline
 - No false positive motifs found
- Found a possible active site in P450 reductases
 - No active site is known for this family
 - Confirmed that the active site we found is related to binding surface

K_{3Dball} + SVM and modified k nearest neighbor

10 families were retrieved from CATH (protein fold) database

Each leave one out test used 20 positives from one family +
100 random negatives

Results:

- k NN: 85% avg TP, 5% avg FP
 - SVM: 72.5% avg TP, 2.5% avg FP
- ⇒ K_{3Dball} is general and can work for many families (i.e. identifies many different folds)

$$\{K_{3Dball}, K_{patternsim}, K_{redoxfunc}\} + \{SVM, kNN\}$$

Positive set was 21 thiol/disulfide oxidoreductases: 17 with thioredoxin (Trx) fold, 4 without

Negative set was 100 random sequences from PDB

Compared to HMMs, DALI [Holm & Sander 93] and CE [Shindyalov & Bourne 98]

Experimental Results

Function Classification: Leave One Out Test (Results)

	<i>TP</i> for redox protein with fold	<i>TP</i> for redox protein without fold	<i>TN</i>
HMM (primary structure)	70.6%	0%	98%
DALI(entire structure)	100.00%	0%	97%
CE (entire structure)	100.00%	0%	98%
$K_{Pattern.Sim} + kNN$	88.23%	50%	98%
$K_{Pattern.Sim} + SVM$	82.35%	50%	100%
$K_{Redox.Func} + kNN$	100.00%	75%	99%
$K_{Redox.Func} + SVM$	94.12%	50%	98%
$K_{3Dball} + kNN$	94.12%	50%	95%
$K_{3Dball} + SVM$	70.6%	50%	99%

DALI and CE could not find proteins without Trx fold, but our methods did

Our methods were > 100 times faster than DALI and CE

From nearly 30,000 protein 3D structures:

- Hit > 280 oxidoreductase candidates (> 90% known ones found)
- Hit 66 P450 reductase candidates (100% known ones found), including NADPH ferredoxin reductase (precursor to Cyt. P450)
- Several candidates to be further examined

New kernels for comparisons of 3D protein structures

- Combine multiple local comparisons into a global comparison
- Very fast to compute ($> 100\times$ faster than DALI and CE)
- K_{3Dball} and $K_{Pattern_Sim}$ require no prior information
 - ⇒ Generally applicable

Presented methods to use our kernels to find structural motifs and perform function classification

- Found known structural motif for oxidoreductases and candidate one for Cytochrome P450
- Able to identify oxidoreductases lacking Trx fold
- Found candidate oxidoreductases and P450s



New Kernels
for Protein
Structural
Motif
Discovery and
Function
Classification

Chang Wang
Stephen D.
Scott

Introduction

Outline

Our New
Kernels

Using our
Kernels

Experimental
Results

Conclusions