Predicting Protein Interactions using Tensor Products

Shawn Martin
Sandia National Laboratories
Albuquerque, NM

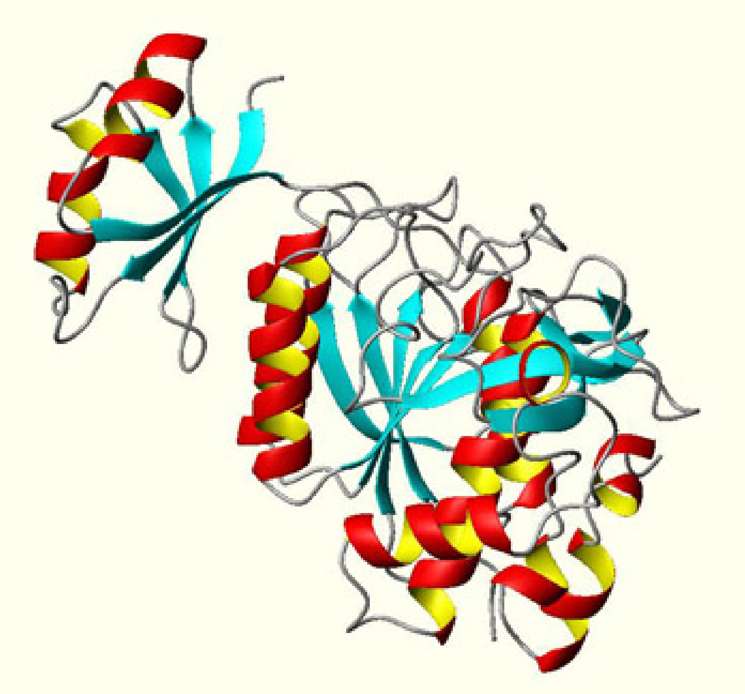
2/5/2008

Proteins

- Proteins participate in most cellular events, such as metabolism, cell signaling, immune response, et cetera.
- A protein is made from a linear sequence of amino acid residues which fold into a 3D structure.
- Many protein sequences are known, most 3D structures are not known.

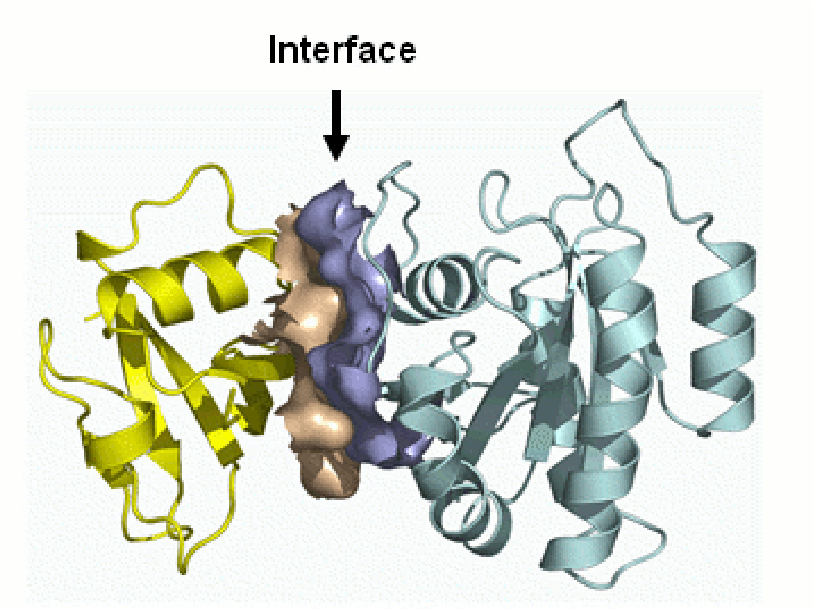
MTTMVL AAAGLR FPGIRP …

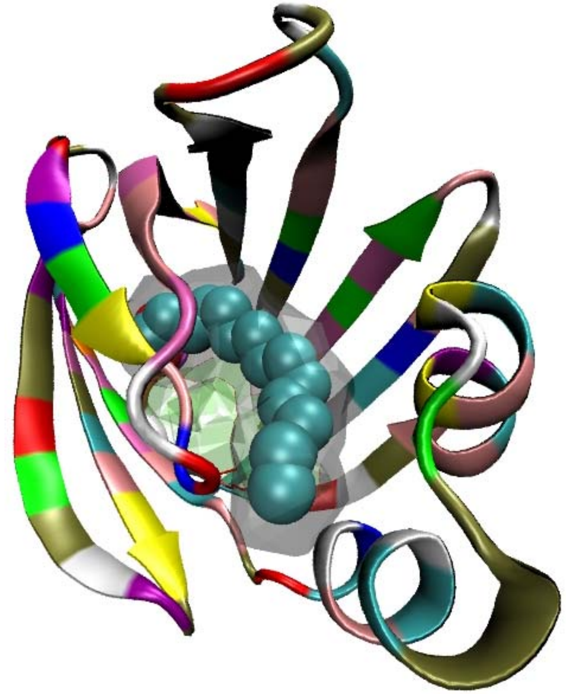
↓



Protein Interactions

- Proteins function by binding with themselves, DNA, and small molecules such as drugs.

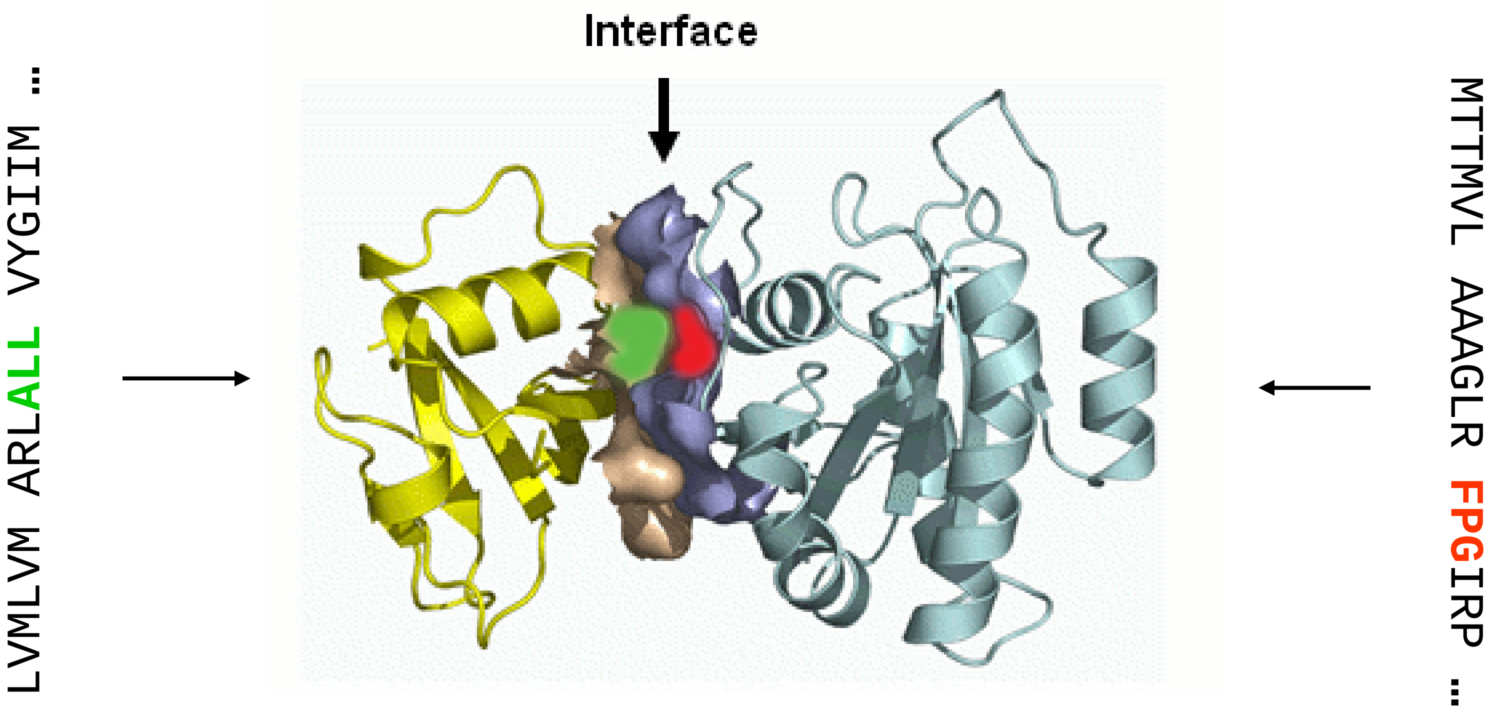




- Protein interactions are predicted using
  - *ab initio* approaches using structure (small scale)
  - *a priori* genomic approaches (large scale)
  - *empirical* approaches based on high-throughput data (large scale)

How do Proteins Interact?

- Current theory is that proteins interact via short sub-sequences (*l*-mers) of amino acid residues in binding pockets.



LVMLVM ARLALL VYGIIM ...

MTTMVL AAAGLR FPGIRP ...

- Our method correlates occurrences of *l*-mer pairs in protein sequences with probability of interaction using experimental data.

Step 1. Count occurrences of *l*-mers in a single protein sequence.

Define $\Phi_s{}^l$: {finite length amino acid strings} → $Z_{\geq 0}^{N_l}$ by

$$\Phi_s^l(P_i) = \sum\nolimits_j \sigma_j \mathbf{z}_j,$$

where

- P_i is the protein sequence.
- $\mathbf{z}_j$ are basis vectors for Z^{N_l} corresponding to *l*-mers.
- σ_j counts the number of occurrences of *l*-mer corresponding to $\mathbf{z}_j$.
- N_l is number of possible *l*-mers.

$$\Phi_s^3(\mathrm{LVMLVM}) = \sum\nolimits_j \sigma_j \mathbf{z}_j = 2\underset{\substack{\updownarrow \\ \mathrm{LVM}}}{\begin{pmatrix}1\\0\\0\\0\\0\\\vdots\end{pmatrix}} + 1\underset{\substack{\updownarrow \\ \mathrm{VML}}}{\begin{pmatrix}0\\1\\0\\0\\0\\\vdots\end{pmatrix}} + 1\underset{\substack{\updownarrow \\ \mathrm{MLV}}}{\begin{pmatrix}0\\0\\1\\0\\0\\\vdots\end{pmatrix}} + 0\underset{\substack{\updownarrow \\ \mathrm{VMT}}}{\begin{pmatrix}0\\0\\0\\1\\0\\\vdots\end{pmatrix}} + \cdots$$

Step 2. Count occurrences of *l*-mer pairs between protein pairs.

Define $\Phi_{s\otimes s}^{l_1\otimes l_2} : \{\text{pairs of amino acid sequences}\} \to Z_{\geq 0}^{N_{l_1}N_{l_2}}$ by

$$\Phi_{s\otimes s}^{l_1\otimes l_2}(P_i, P_j) = \Phi_s^{l_1}(P_i) \otimes \Phi_s^{l_2}(P_j)$$

$$\begin{aligned}\Phi_s^3(\text{LVMLVM}, \text{MTTMVL}) &= (2,1,1,0,0)^T \otimes (1,0,0,2,1)^T \\ &= (2,1,1,0,0)^T (1,0,0,2,1) \\ &= \begin{pmatrix} 2 & 0 & 0 & 4 & 2 \\ 1 & 0 & 0 & 2 & 1 \\ 1 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}\end{aligned}$$

Notes:

- We normally write this matrix as a vector.
- If $l_1 = l_2$ we use $\Phi_s{}^l$ to denote $\Phi_{s\otimes s}^{l_1\otimes l_2}$.

Step 3. Compute similarity between two protein pairs.

We define the similarity between two protein pairs using

$$k_{s\otimes s}^{l_1\otimes l_2}((P_{i_1},P_{i_2}),(P_{j_1},P_{j_2}))=\Phi_{s\otimes s}^{l_1\otimes l_2}(P_{i_1},P_{i_2})^T\,\Phi_{s\otimes s}^{l_1\otimes l_2}(P_{j_1},P_{j_2})$$

$$\Phi_s^3(\mathrm{LVMLVM},\mathrm{MTTMVL})=\begin{pmatrix}2&0&0&4&2\\1&0&0&2&1\\1&0&0&2&1\\0&0&0&0&0\\0&0&0&0&0\end{pmatrix}$$

$$\Phi_s^3(\mathrm{VLMVLM},\mathrm{TTMVLM})=\begin{pmatrix}1&0&1&1&1\\1&0&1&1&1\\2&0&2&2&2\\0&0&0&0&0\\0&0&0&0&0\end{pmatrix}$$

k_s^3((LVMLVM, MTTMVL), (VLMVLM, TTMVLM)) = 20

Steps 1-3. Observations.

- Advantages:
  - By comparing a protein pair **P** = (P_1, P_2) with pairs $\{\mathbf{P}_i = (P_{i_1}, P_{i_2})\}$ known to interact we can predict if **P** is an interacting pair.
  - Ignores position of *l*-mer in protein sequence.
  - Allows arbitrary sequence lengths.
- Disadvantages:
  - Produces very high-dimensional vectors in $Z_{\geq 0}^{20^{l+l}}$.
  - Not symmetric with respect to sequence order.
  - Not symmetric with respect to protein pair order.
  - Not normalized with respect to sequence length.

Step 4. Computational simplification to alleviate high dimensionality.

- To avoid explicit computation of tensor products, we use the following identity:

$$
\begin{aligned}
k_s^l((P_{i_1},P_{j_1}),(P_{i_2},P_{j_2})) &= (\Phi_s^l(P_{i_1})\otimes\Phi_s^l(P_{j_1}))^T(\Phi_s^l(P_{i_2})\otimes\Phi_s^l(P_{j_2})) \\
&= \mathrm{trace}\,((\Phi_s^l(P_{i_1})\Phi_s^l(P_{j_1})^T)(\Phi_s^l(P_{i_2})\Phi_s^l(P_{j_2})^T)^T) \\
&= \cdots \\
&= \Phi_s^l(P_{j_1})^T\Phi_s^l(P_{j_2})\Phi_s^l(P_{i_1})^T\Phi_s^l(P_{i_2}) \\
&= k_s^l(P_{i_1},P_{i_2})k_s^l(P_{j_1},P_{j_2})
\end{aligned}
$$

- Now we can compute similarities between protein pairs by computing similarities between proteins.

$k_s{}^3$((LVMLVM, MTTMVL), (VLMVLM, TTMVLM)) =
$k_s{}^3$(LVMLVM, VLMVLM)×$k_s{}^3$(MTTMVL, TTMVLM) = 5×4 = 20

Step 5. Additional modifications.

- Symmetry in sequence order is accomplished by replacing *l*-mers with odd length "signatures," where middle letter is first and strings on either side are alphabetized:

LVM → VLM   MLV → LMV

VML → MLV   VMT → MTV

- Symmetry in protein comparison order is accomplished by using a symmetric sum:

$$\Phi_s^l(P_i, P_j) = \Phi_s^l(P_i) \otimes \Phi_s^l(P_j) + \Phi_s^l(P_j) \otimes \Phi_s^l(P_i)$$

- Normalization according to protein length is accomplished by using a generic normalized similarity:

$$k(P_1, P_2) \Big/ \sqrt{k(P_1, P_1) k(P_2, P_2)}$$

Step 6. Use Support Vector Machine (SVM) function approximation to correlate occurrences of *l*-mer pairs with probability of interaction.

A protein interaction SVM is given by

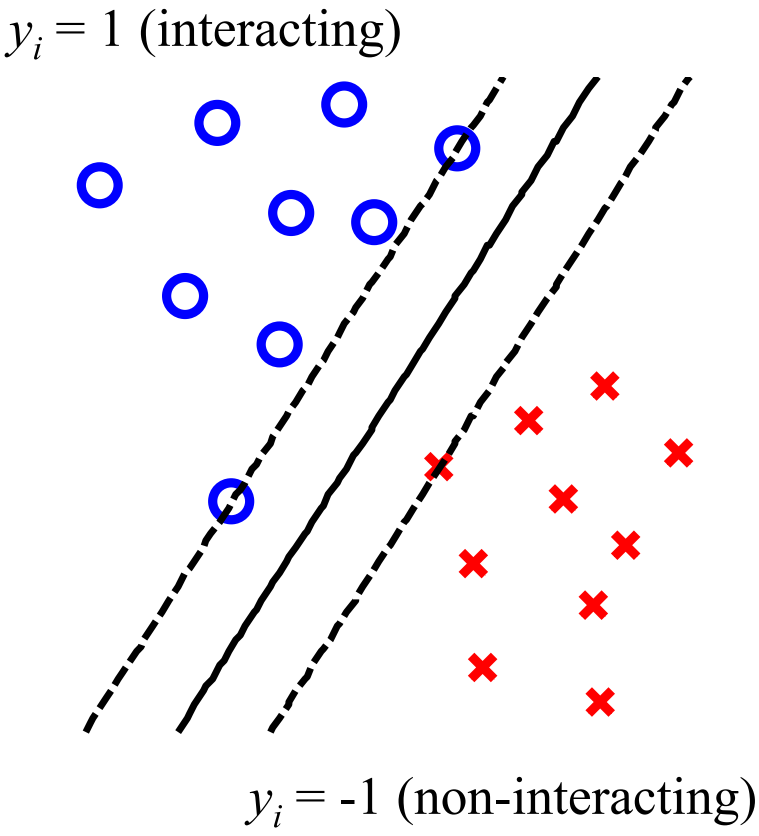
$$f(\mathbf{P}) = \sum_i y_i \alpha_i k(\mathbf{P}, \mathbf{P}_i) + b$$

where we obtain α_i by solving the quadratic programming problem

$$\max_\alpha \quad \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j k(\mathbf{P}_i, \mathbf{P}_j) - \sum_i \alpha_i$$

$$\text{s.t.} \quad \sum_i y_i \alpha_i = 0, \qquad 0 \le \alpha_i \le C$$

(b is obtained implictly.)



Solving this optimization problem is known as “training” the SVM.

Application 1. Protein-Protein Interactions.

- We first benchmarked our method on Yeast and *H. pylori* datasets.
  - 709 Yeast SH3 domain-ligand pairs (Tong *et al.*, 2002).
  - 2082 Yeast protein pairs (Sprinzak & Margalit, 2001).
  - 1458 *H. pylori* protein pairs (Rain *et al*., 2001).
  - 7714 Yeast "gold standard" protein pairs (Jansen *et al*., 2003).
  - Non-interacting pairs were chosen at random.

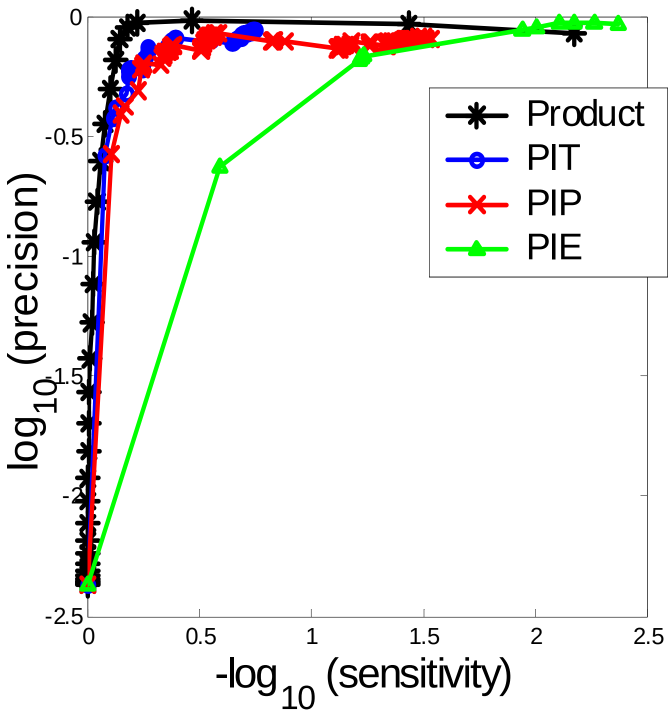
- We compared against other methods by using 10-fold cross validation and computing accuracy, precision, and sensitivity.

Comparisons with Other Methods

Yeast SH3	*Accuracy*	*Precision*	*Sensitivity*
Ligand-Only	73. 7	75. 5	63. 1
Product	80. 7	81. 4	75. 2
PSSM	75. 4	68. 8	81. 3

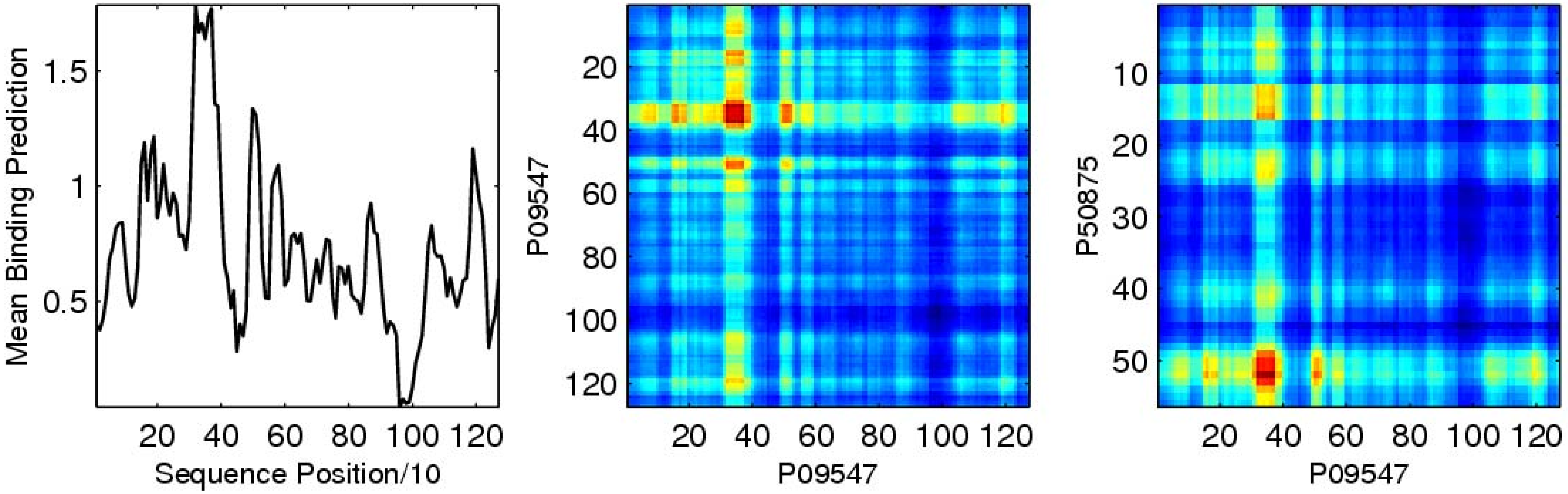
Full Yeast	*Accuracy*	*Precision*	*Sensitivity*
Product	69. 0	71. 5	63. 2
InterPro	70. 8	86. 5	49. 2
Sprinzak	68. 8	79. 8	50. 0

H. pylori	*Accuracy*	*Precision*	*Sensitivity*
Product	83. 4	85. 7	79. 9
Bock&Gough	75. 8	80. 2	69. 8



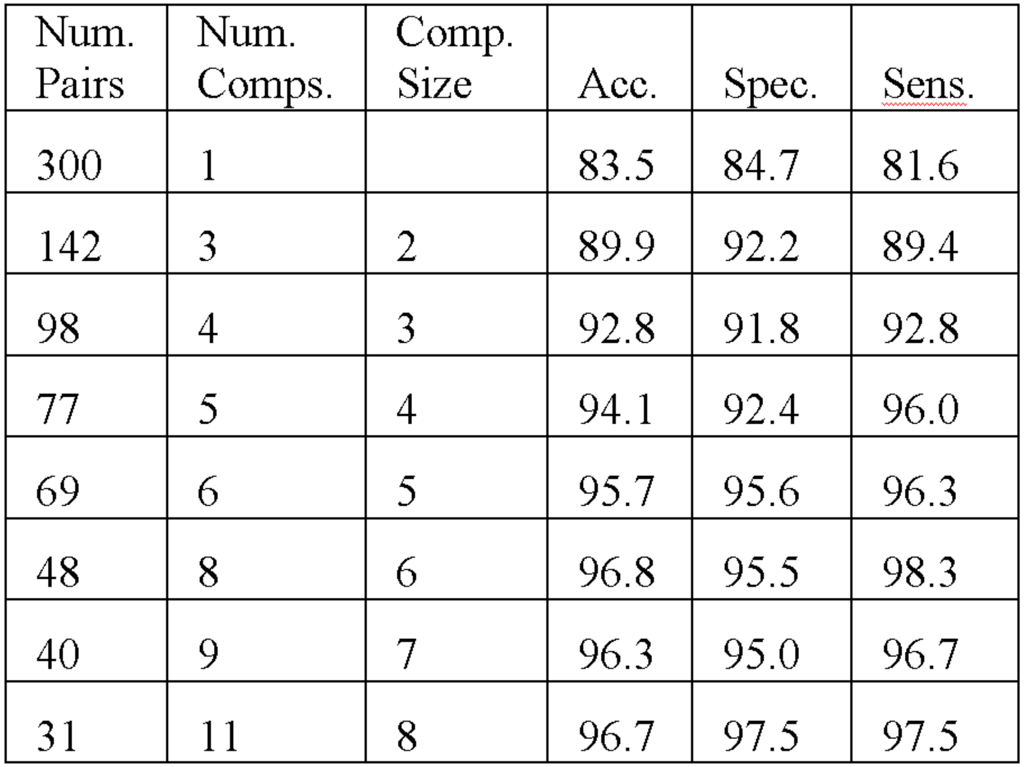
Locating Protein Domains

- We also tested the ability of our algorithm to locate protein domains.
  - Domains are evolutionarily conserved subsequences thought to be good candidate binding sites.
- We used a sliding window of 50 amino acid residues in Yeast proteins.



Using Protein Complexes

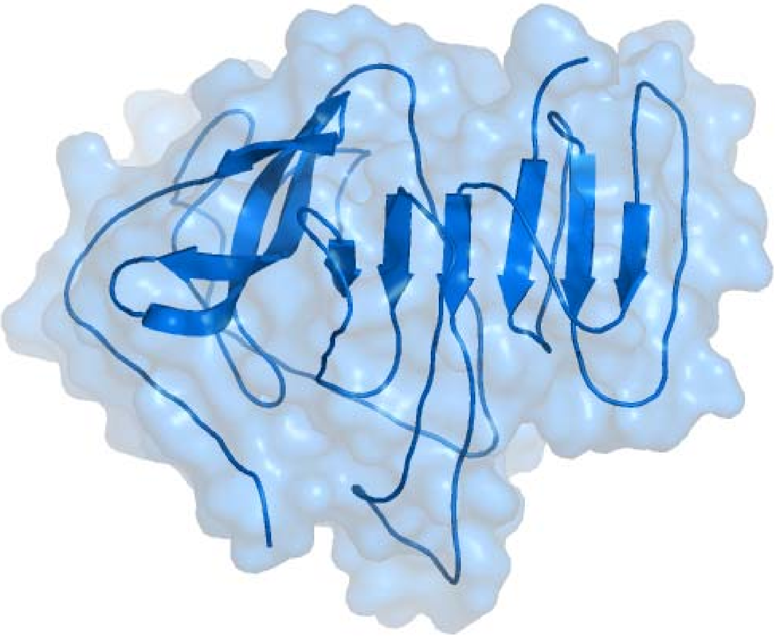
- In a collaboration with S. Rasheed’s group at USC (Viral Oncology and Proteomics Research), we used protein complex data infer a feline protein network.
  - Proteins were given in experimentally determined functional groups.
  - Protein pairs belonging to multiple groups were more likely to interact.



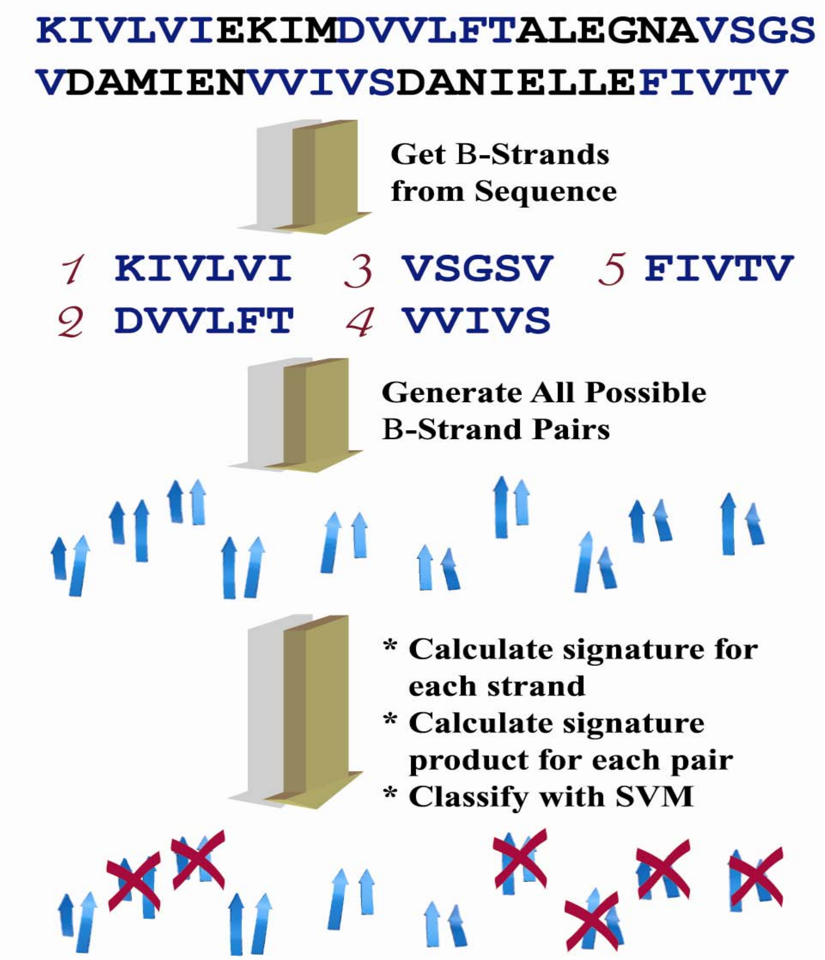
Num. Pairs	Num. Comps.	Comp. Size	Acc.	Spec.	Sens.
300	1		83.5	84.7	81.6
142	3	2	89.9	92.2	89.4
98	4	3	92.8	91.8	92.8
77	5	4	94.1	92.4	96.0
69	6	5	95.7	95.6	96.3
48	8	6	96.8	95.5	98.3
40	9	7	96.3	95.0	96.7
31	11	8	96.7	97.5	97.5

Application 2. *β*-Strand Ordering.

- In a collaboration with C. Strauss at Los Alamos National Laboratory Bioscience Division, we tested our methods ability to predict protein secondary structure.
  - Protein amino acid subsequences interact to form secondary structures, such as *α*-helices and *β*-sheets.
  - *Can we use our method to predict β-strand ordering in β-sheets?*



β-Strand Ordering Prediction



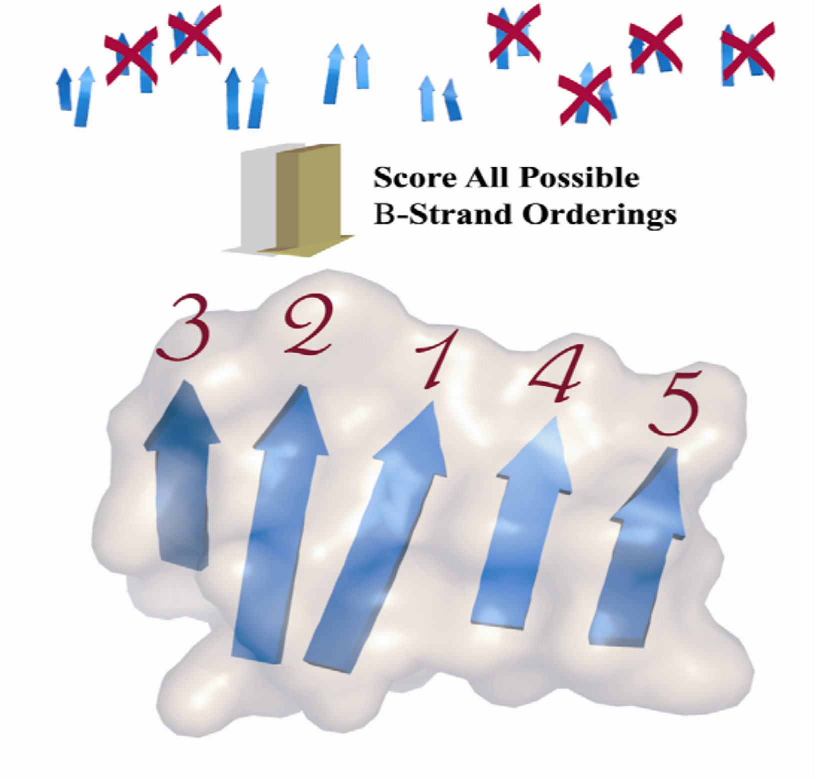
KIVLVIEKIMDVVLFTALEGNAVSGS
VDAMIENVVIVSDANIELLEFIVTV

Get B-Strands
from Sequence

1 KIVLVI
2 DVVLFT
3 VSGSV
4 VVIVS
5 FIVTV

Generate All Possible
B-Strand Pairs

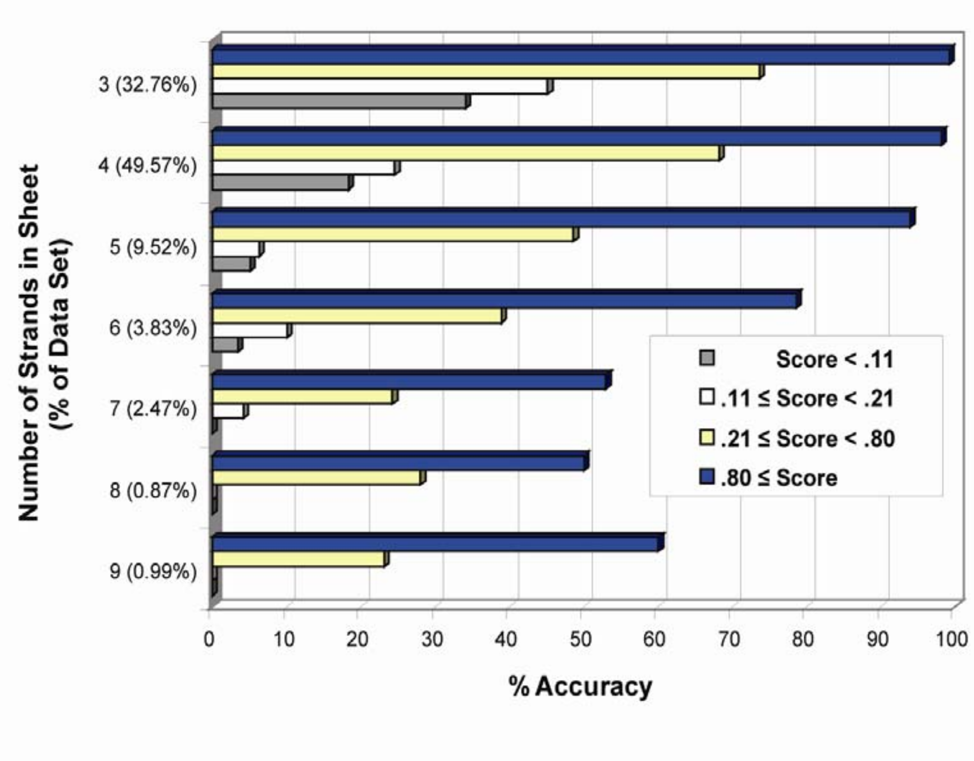
* Calculate signature for each strand
* Calculate signature product for each pair
* Classify with SVM

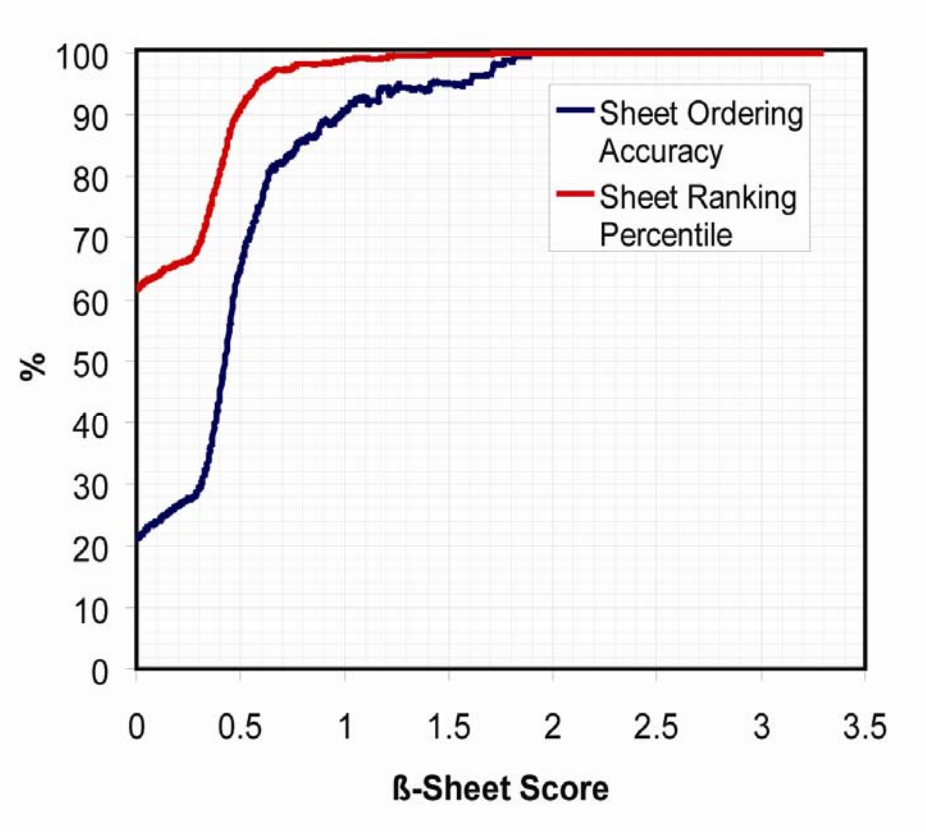


Score All Possible
B-Strand Orderings

β-Strand Ordering Results

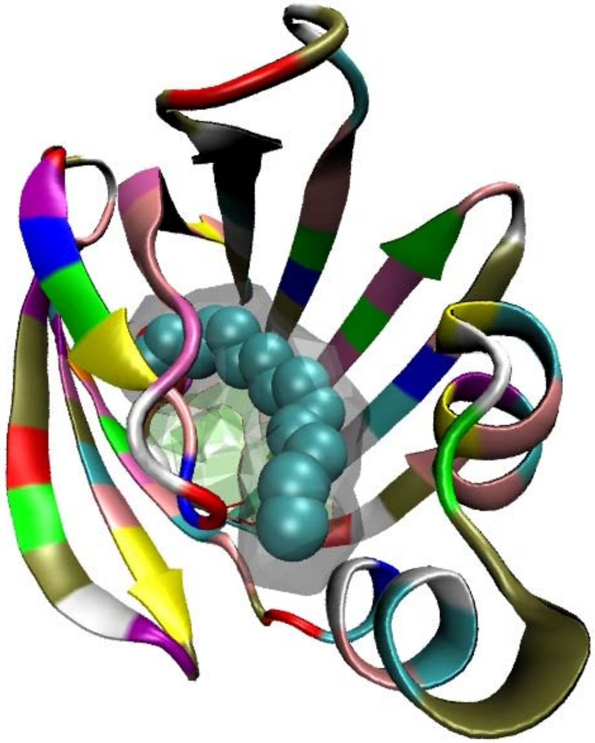
(using 27,196 Strands from Protein Data Bank)



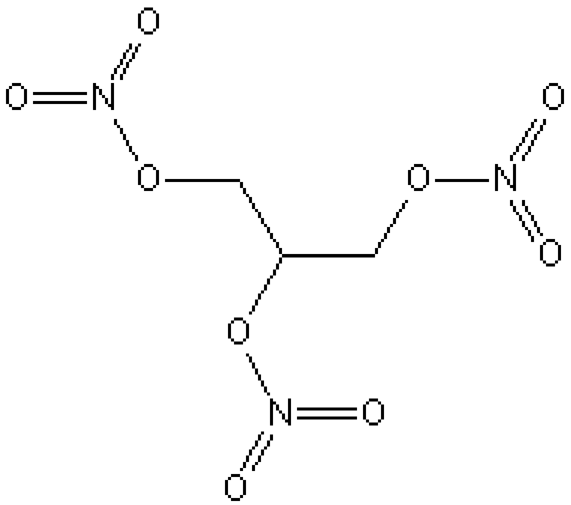


Application 3. Protein-Chemical Interactions.

- Protein-chemical interaction prediction is useful in drug design.
- Almost all interaction prediction is done at a small (but accurate) scale.
- *Can we use our method to do large scale empirical predictions?*



Describing Chemicals



Define $\Phi_g{}^h$: {chemical graphs} → Z^{N_h} by

$$\Phi_g^h(C_i) = \sum_j \sigma_j \mathbf{z}_j,$$

where

- C_i is a labeled graph describing a chemical
- $\mathbf{z}_j$ are basis vectors for Z^{N_h} corresponding to depth h subgraphs.
- σ_j counts the number of occurrences of depth h subgraph corresponding to $\mathbf{z}_j$.
- N_h is the number of depth h subgraphs.

$$3\ \mathrm{O(NC)} \leftrightarrow \mathbf{z}_1$$

$$6\ \mathrm{O(=N)} \leftrightarrow \mathbf{z}_2$$

$$3\ \mathrm{N(O=O=O)} \leftrightarrow \mathbf{z}_3$$

$$5\ \mathrm{H(C)} \leftrightarrow \mathbf{z}_4$$

$$2\ \mathrm{C(OHHC)} \leftrightarrow \mathbf{z}_5$$

$$1\ \mathrm{C(OHCC)} \leftrightarrow \mathbf{z}_6$$

$$\updownarrow$$

$$(3,6,3,5,2,1)^T$$

Comparing Protein-Chemical Pairs

- In order to predict protein-chemical interactions we again define a similarity measure for protein-chemical pairs.

$$k_g^h(C_i, C_j) = \Phi_g^h(C_i)^T \Phi_g^h(C_j)$$

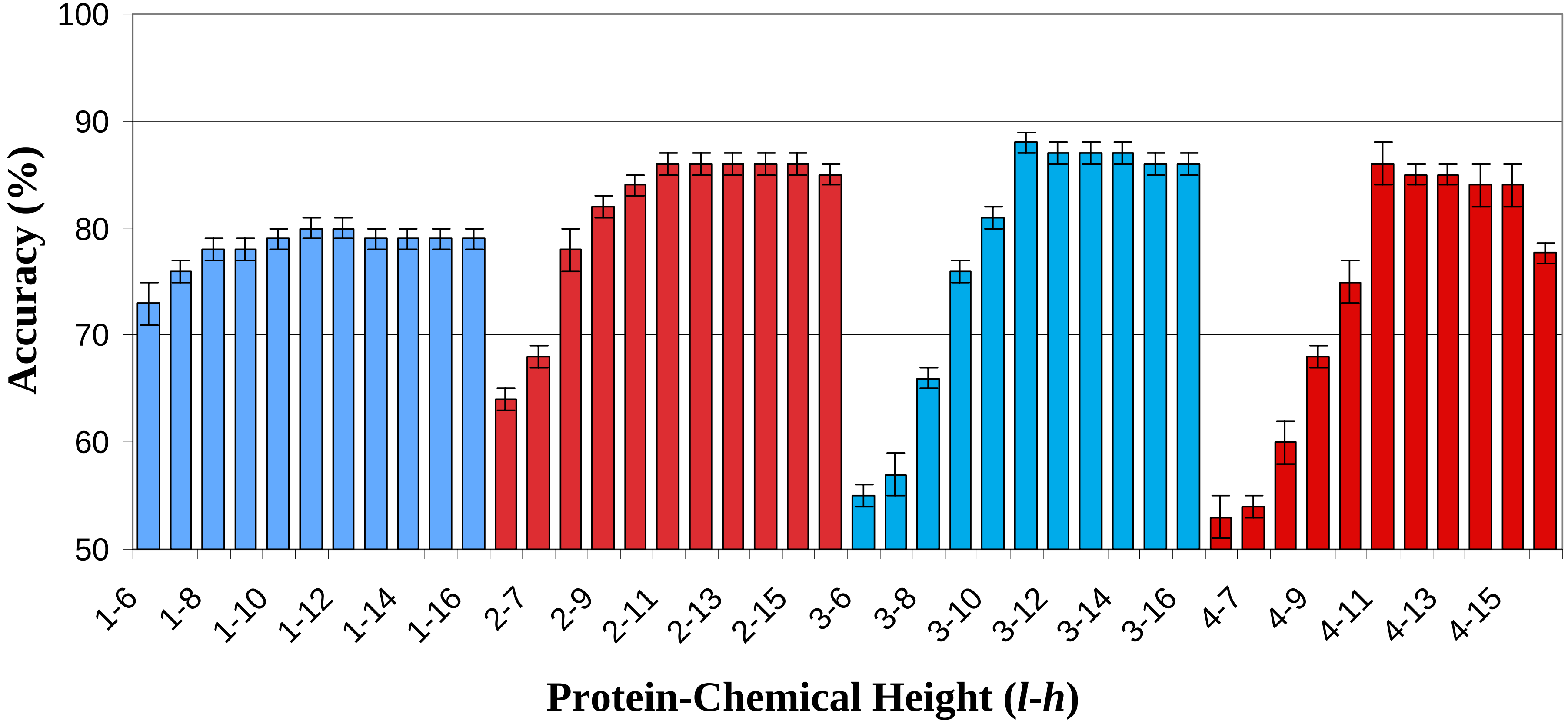
$$\Phi_{s\otimes g}^{l\otimes h}(P_i, C_i) = \Phi_s^l(P_i) \otimes \Phi_g^h(C_i)$$

$$k_{s\otimes g}^{l\otimes h}((P_i, C_i), (P_j, C_j)) = \Phi_{s\otimes g}^{l\otimes h}(P_i, C_i)^T \Phi_{s\otimes g}^{l\otimes h}(P_j, C_j)$$

$$k_{s\otimes g}^{l\otimes h}((P_i, C_i), (P_j, C_j)) = k_s^l(P_i, P_j) k_g^h(C_i, C_j)$$

Drug-Target Prediction Results

(using 873 pairs from KEGG)



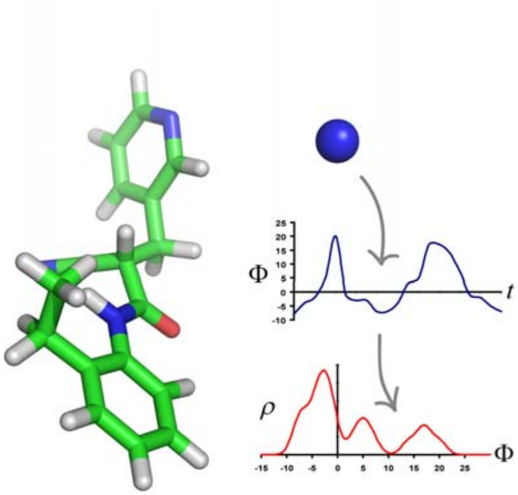
Conclusions

Structure Based Methods

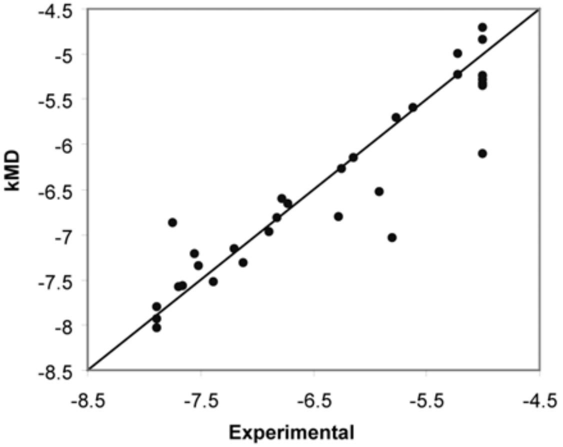
- Accurate
- Slow
- Small Scale
- Often completely *ab initio*

Sequence Based Methods

- Less accurate
- Fast
- Large scale
- Usually completely empirical



$$\xrightarrow{\int}$$



- *Future work: hybrid structure/statistical method.*

References

- Invited Book Chapter:
  - S. Martin, W. M. Brown, and J.-L. Faulon (in press, 2007), “Predicting Protein Interactions using Product Kernels,” *Advances in Biochemical Engineering/Biotechnology*, Springer-Verlag.

- Sandia National Laboratory Projects:
  - J.-L. Faulon, M. Misra, S. Martin, K. Sale, and R. Sapra (in press, 2007), “Genome Scale Enzyme-Metabolite and Drug-Target Interaction Predictions using the Signature Molecular Descriptor,” *Bioinformatics*.
  - S. Martin, D. Roe, and J.-L. Faulon (2005), “Predicting Protein-Protein Interactions using Signature Products,” *Bioinformatics* 21(2):218-226.

- Collaboration with USC Laboratory of Viral Oncology and Proteomics:
  - S. Martin, Z. Mao, L. S. Chan, and S. Rasheed (2007), “Inferring Protein-Protein Interaction Networks using Protein Complex Data,” *International Journal of Bioinformatics Research and Applications* 3(4):480-492.
  - S. Martin, Z. Mao, L. S. Chan, S. Rasheed (2006), “Protein Interaction Extrapiolated from Feline Protein Complexes,” Proceedings of the 3rd Biotechnology and Bioinformatics Symposium (BIOT):45-52.

- Collaboration with Los Alamos National Laboratory Bioscience Division:
  - W. M. Brown, S. Martin, J. Chabarek, C. Strauss, and J.-L. Faulon (2006), “Prediction of Beta-Strand Packing Interactions using the Signature Product,” *Journal of Molecular Modeling* 12(3):355-361.