Kernel Space Embedded Graphical Models of Protein Structures

Narges Razavian

Christopher James Langmead

Carnegie Mellon University

NIPS 2012 Workshop: Confluence between Kernel Methods and Graphical Models

December 2012

Application Background

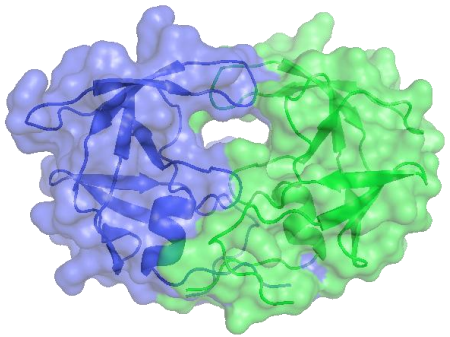
- Proteins: Large dynamic molecules, composed of a sequences of Amino Acids subunits

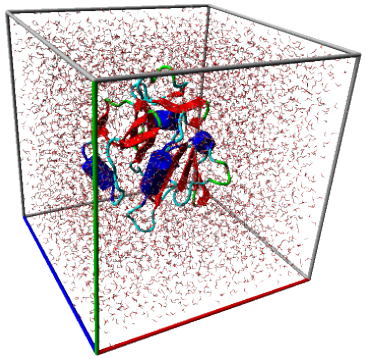
(1) **Crystalize** the protein, get **Snapshot of structure**

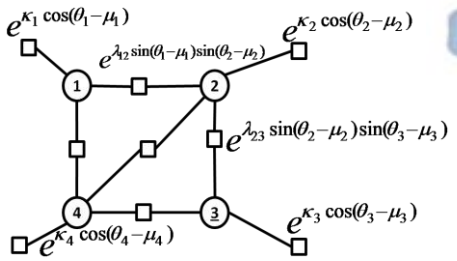
(2) **Generate** dynamics samples via **Molecular Dynamics Simulations**

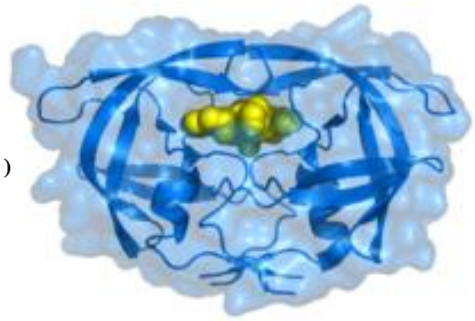
(3) **Learn** probabilistic models of protein structure

(4) **Analyze:** Detect potential targets for drugs, etc.



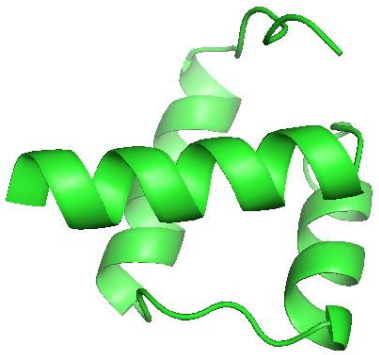




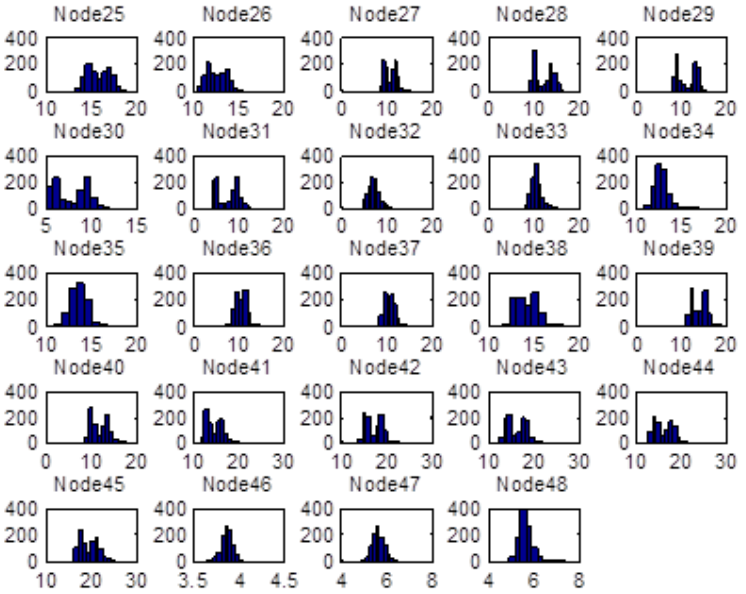


Graphical Models of Protein Conformation

- Data related challenges
  - Large sample set and dimension
  - Protein structure represented as sequences of angles
- Current solutions
  - Discrete Graphical Models
  - Gaussian Graphical Models
  - von-Mises Graphical Models



True marginal distributions



Reproducing Kernel Hilbert Space Embedded Graphical Models

- Kernel space embedding of all components of the graphical model
  - Conditional probabilities
  - Kernel Belief Propagation
    - Initial messages from observed nodes
    - Combine incoming messages
    - Compute beliefs at the root node
- Assumes structure known a priori.
  - Unreasonable for protein structure models

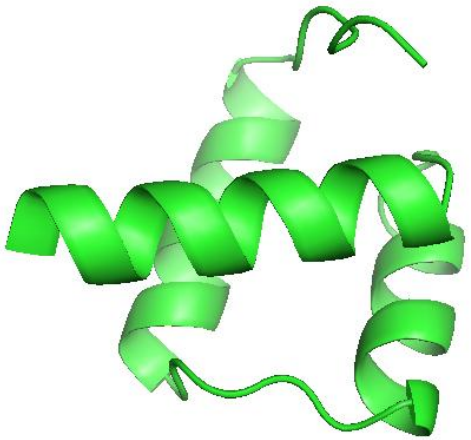
Structure Learning for RKHS embedded graphical models

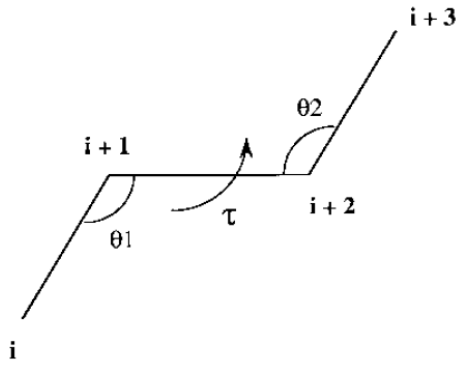
- Nonparametric latent tree structure learning in RKHS [Song et.al. 2011]
  - Nonparametric *tree metric* based on nonparametric correlation coefficient

- Neighborhood selection[Meinhausen et.al. 2010]
  - Parallel Lasso Regression models for each node

Experiments.

- Dataset: MD simulation of Engrailed Homeodomain
  - 54 residue, DNA binding domain
  - Ultra-fast folding (15 μ sec)
  - Representation: Sequence of Theta and Tau angles
  - 500K samples at 350 Kelvin: Used 2 small subsampled data, due to our model's scalability limitations.

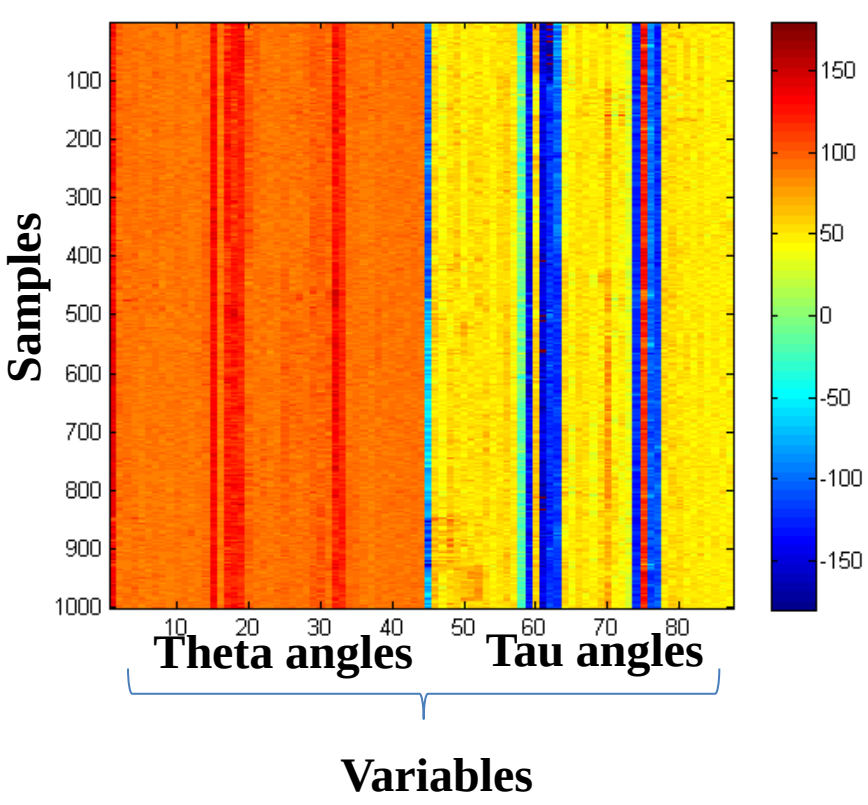




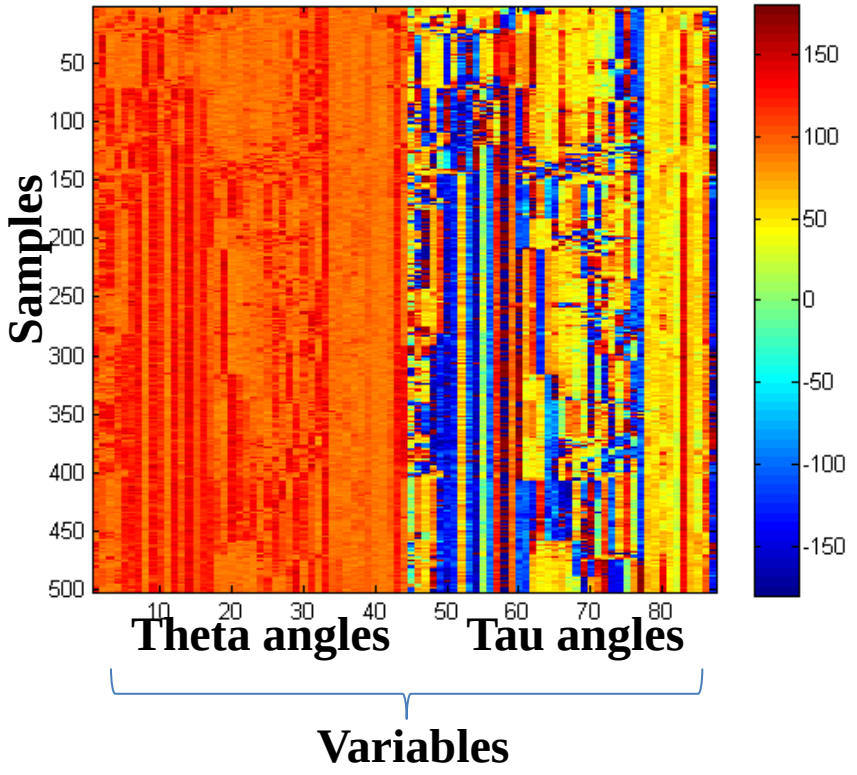
- Leave-one-out cross validation
- Baseline: Non-paranormal, and Gaussian Graphical models

Visual depiction of two subsampled datasets

First 1000 samples



Uniformly sampled 1000 samples



Results for Structure Learning and Nonparametric Inference

Model and Inference	RMSE for first1000 data	RMSE for Uniform1k
kernel 1:Gaussian (structure by: neighborhood selection)	8.42	54.76
Kernel 2:Triangonometric (structure by: neighborhood selection)	7.30	51.34
Non-paranormal Graphical Model	8.43	63
Gaussian Graphical Model	8.46	59.4

* All experiments achieved Wilcoxon Rank test p-value under 7.5e-7

$$Kernel\ 1 : K(x,y) = exp\{-\lambda\ ||x-y||^{2}\}$$

$$Kernel\ 2 : K(x,y) = exp\{-\lambda\ sin(||x-y||)^{2}\}$$

Results: Neighborhood Selection vs. nonparametric Tree learning

Structure Learning method	RMSE(in degrees) First 1000 dataset
nonparametric Tree metric (using kernel 2)	7.41
Neighborhood selection via Linear regression (using kernel 2)	7.30

* Wilcoxon Rank test p-value 5.1e-13

$$Kernel\ 1 : K(x,y) = exp\{-\lambda\ ||x-y||^{2}\}$$

$$Kernel\ 2 : K(x,y) = exp\{-\lambda\ sin(||x-y||)^{2}\}$$

Conclusions and Future Work

- Nonparametric graphical models outperform baselines of Gaussian Graphical Models and non-paranormal graphical models for protein structure modeling
  - Only if the appropriate kernel is used
- Neighborhood Selection performs better than nonparametric tree structure learning
- Future work:
  - Investigate nonparametric regression models for Neighborhood selection
  - Improve scalability