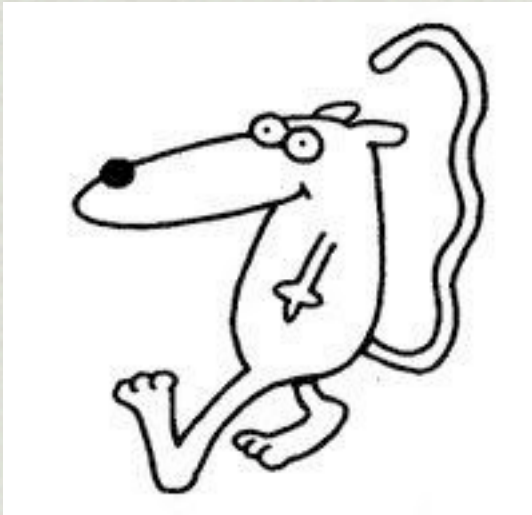


Predicting Gene-disease Association By Markov Random Field (MRF) Modeling

Lab Meeting

3.29.2011

Image Restoration: Introduction to MRF



$$\begin{array}{ll} \text{close} & \sum_{i=1}^N (x_i - d_i)^2 \\ \text{smooth} & \sum_{(i,j) \in \mathcal{N}} (x_i - x_j)^2 \end{array}$$

$$H(\mathbf{x}, \mathbf{d}) = \beta \sum_{i=1}^N (x_i - d_i)^2 + \sum_{(i,j) \in \mathcal{N}} (x_i - x_j)^2$$



$$P(\mathbf{x} \mid \mathbf{d}) \propto P(\mathbf{x}, \mathbf{d}) = \frac{1}{Z} \exp\{-H(\mathbf{x}, \mathbf{d})\}$$

- Prohibitively large number of configurations
- Vague concepts for most applications

Gibbs Distribution

Definition: Let G be a neighborhood system for S , and let C be a subset of S . The set C is a clique in $\{S, G\}$ either if it is a one-element set, or $|C| > 1$ and any two of its elements are neighbors. We denote the set of all cliques in $\{S, G\}$ by $\mathcal{C}$.

Definition: A Gibbs probability measure π on Ω relative to $\{S, G\}$, is defined by

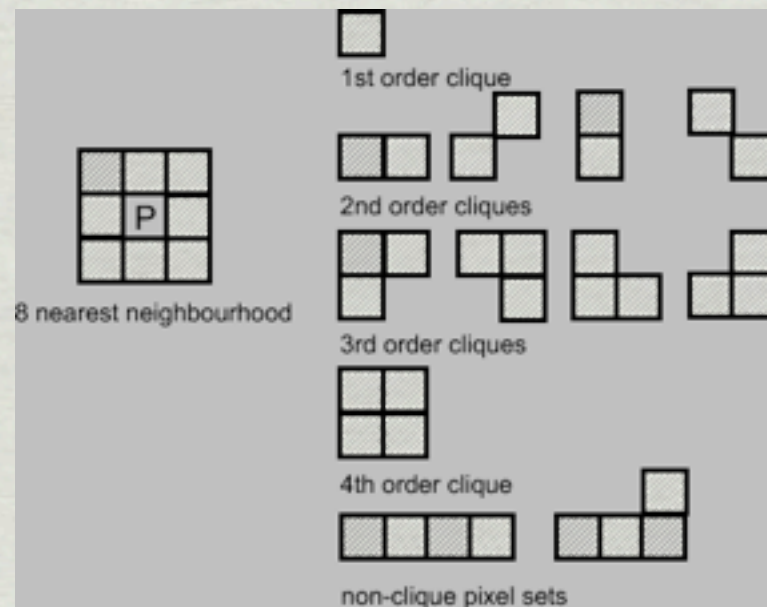
$$\pi(\omega) = \frac{1}{Z} \exp \{-U(\omega)/T\},$$

such that

- for any configuration $\omega \in \Omega$

$$U(\omega) = \sum_{c \in \mathcal{C}} V_c(\omega).$$

- Z is a normalized constant and T is a positive constant.



Markov-Gibbs Equivalence

Hammersley-Clifford Theorem: a random field is a Markov random field if and only if the corresponding joint probability distribution is a Gibbs distribution.

Conditional Probability

$$\begin{aligned} P(X_i \mid X_{[-i]}) &= \frac{\exp\{-U(\omega)/T\}}{\sum_{x_i} \exp\{-U(\omega)/T\}} \\ &= \frac{\exp\{-\sum_{i \in c} V_c(\omega)/T\}}{\sum_{x_i} \exp\{-\sum_{i \in c} V_c(\omega)/T\}} \end{aligned}$$

Ising Model

$$U(\omega) = \alpha \sum x_{ij} + \beta \left(\sum x_{ij} x_{i(j+1)} + \sum x_{ij} x_{(i+1)j} \right)$$



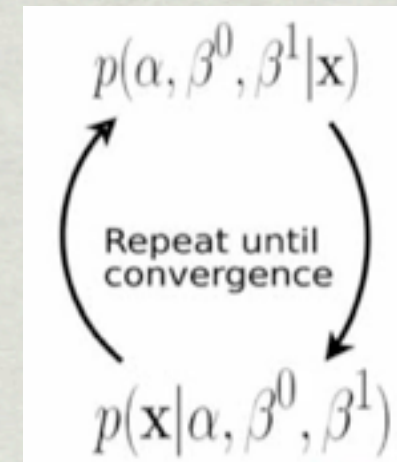
Prediction of protein function using protein-protein interaction data

Deng et al. (2003)

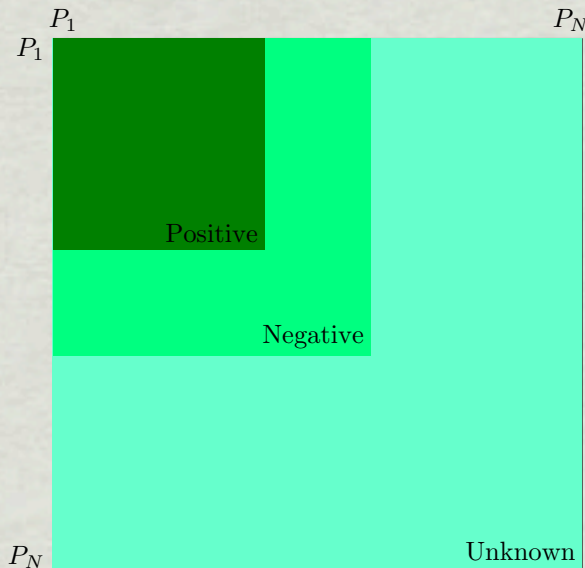
$$\log \frac{P(X_i = 1 \mid X_{[-i]}, \theta)}{1 - P(X_i = 1 \mid X_{[-i]}, \theta)} = \alpha + \beta^{(0)} M_0^{(i)} + \beta^{(1)} M_1^{(i)}$$

Bayesian Markov Random Field analysis for protein function prediction based on network data

Kourmpetis *et al.* (2010)



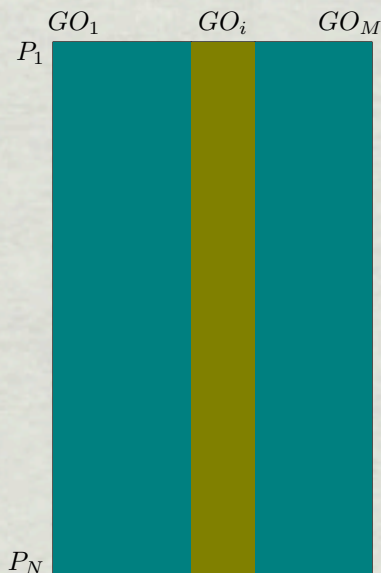
MRF-Deng Model



Prior probability

$$\prod_{i=1}^N \pi^{x_i} (1 - \pi)^{1-x_i}$$

A pair of interacting proteins tend to have the same function rather than different functions.



Conditional probability

$$P(\mathbf{d} \mid \mathbf{x}) \propto \exp\{\beta N_{01} + \gamma N_{11} + N_{00}\}$$

The goal is to find the posterior probability for each unknown protein.

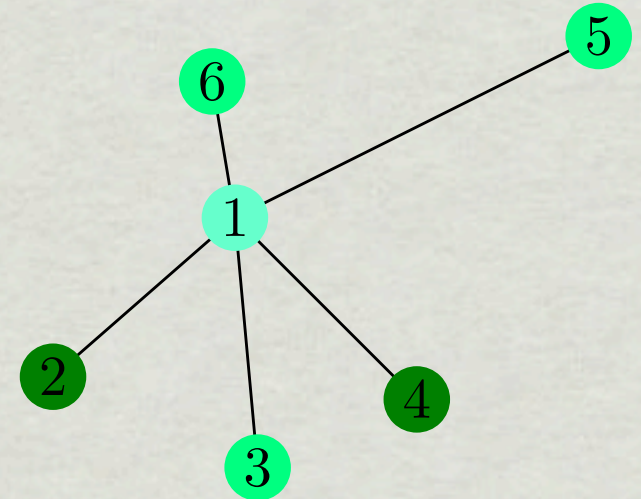
MRF-Deng Model

Gibbs sampler

$$P(X_i = 1 \mid X_{[-i]}, \theta) = \frac{\exp(\alpha + \beta^{(0)} M_0^{(i)} + \beta^{(1)} M_1^{(i)})}{1 + \exp(\alpha + \beta^{(0)} M_0^{(i)} + \beta^{(1)} M_1^{(i)})}$$

Parameter estimation

$$\log \frac{P(X_i = 1 \mid X_{[-i]}, \theta)}{1 - P(X_i = 1 \mid X_{[-i]}, \theta)} = \alpha + \beta^{(0)} M_0^{(i)} + \beta^{(1)} M_1^{(i)}$$



Functional labels of the proteins in the network are not independent.

BMRF Model

- Initial parameters are estimated by MRF-Deng method but using brglm instead of standard logistic method.
- Further parameters are updated by DEMC (Differential Evolution Markov Chain)
 - Propose new parameters

$$\theta^* = \theta + \gamma(Z_{R1} - Z_{R2}) + \mathbf{e}$$

- Compute the Pseudo-Likelihood function

$$\text{PLF}(\mathbf{x} \mid, \theta) = \prod_{i=1}^N P(x_i \mid \mathbf{x}_{[-i]}, \theta)$$

- Metropolis step

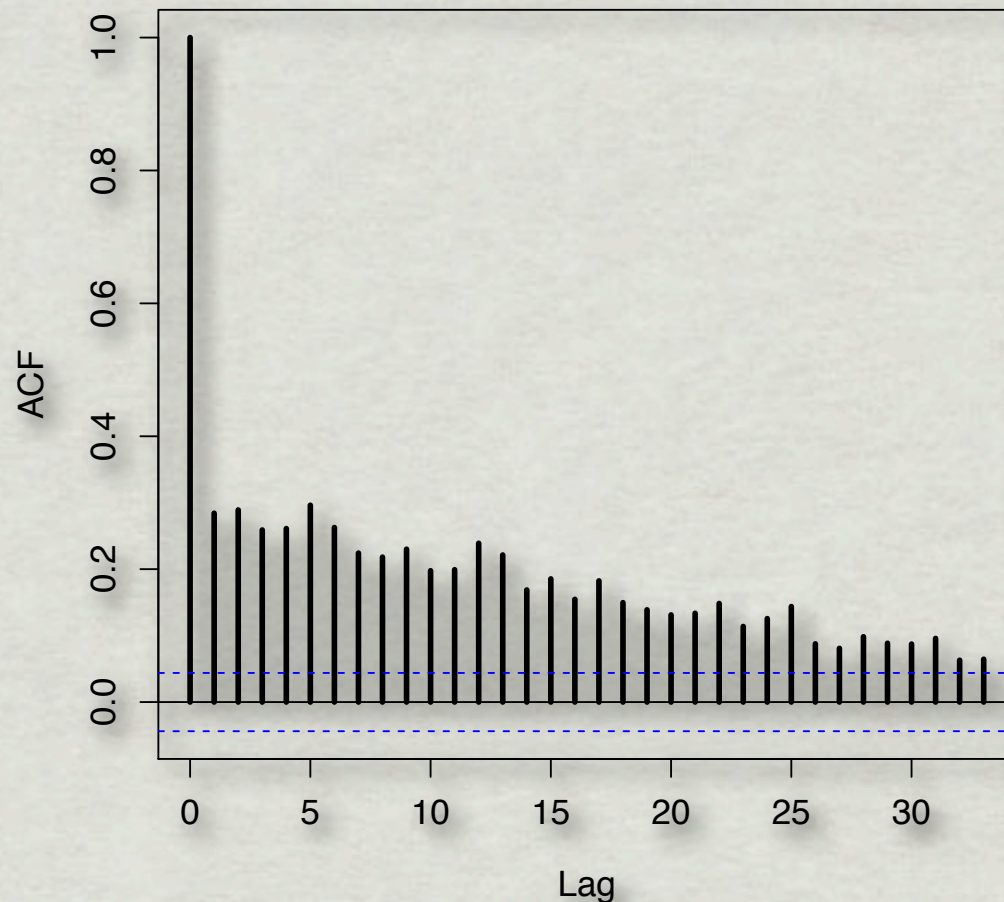
$$r = \min \left(1, \frac{\text{PLF}(\mathbf{x}^{(t)} \mid \theta^*)}{\text{PLF}(\mathbf{x}^{(t)} \mid \theta)} \right)$$

i	θ^* / θ	$\rightarrow$	x
1	$\theta^* / \theta^{(1)}$	$\rightarrow$	$x^{(1)}$
2	$\theta^* / \theta^{(2)}$	$\rightarrow$	$x^{(2)}$
$\vdots$	$\vdots$		$\vdots$
s	$\theta^* / \theta^{(s)}$	$\rightarrow$	$x^{(s)}$

Application: Yeast PPI Data

1,622 proteins; 90 GO terms; 5% unknown proteins

Auto-Correlation Function Estimation

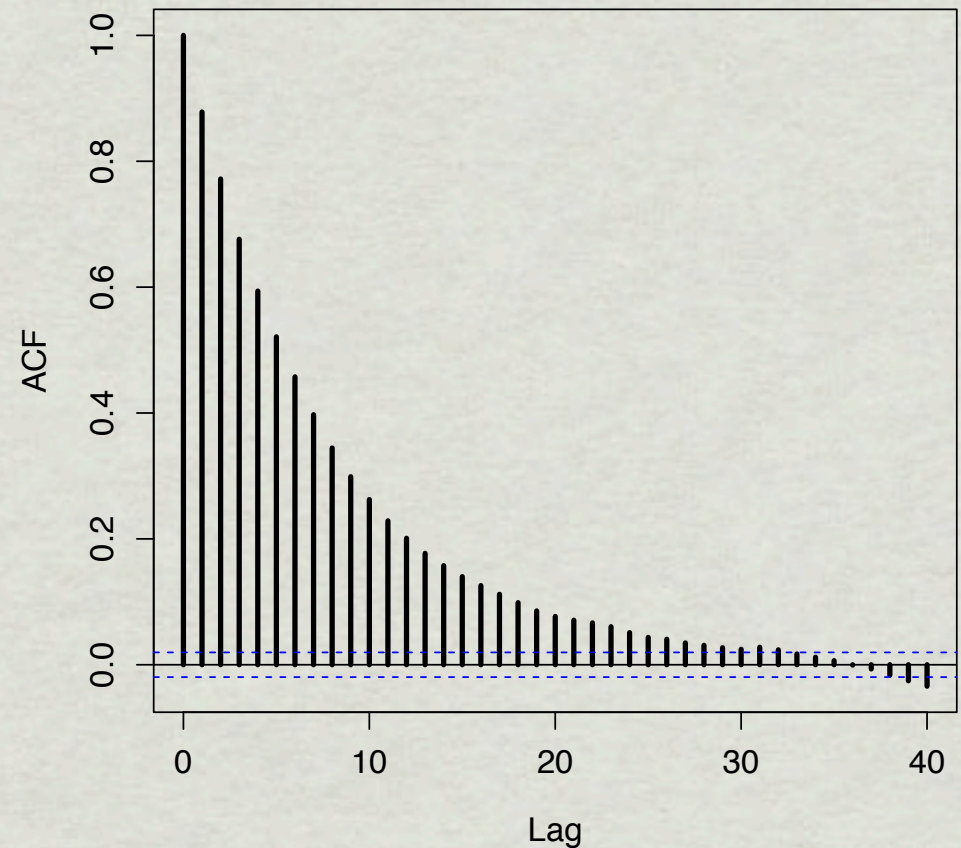
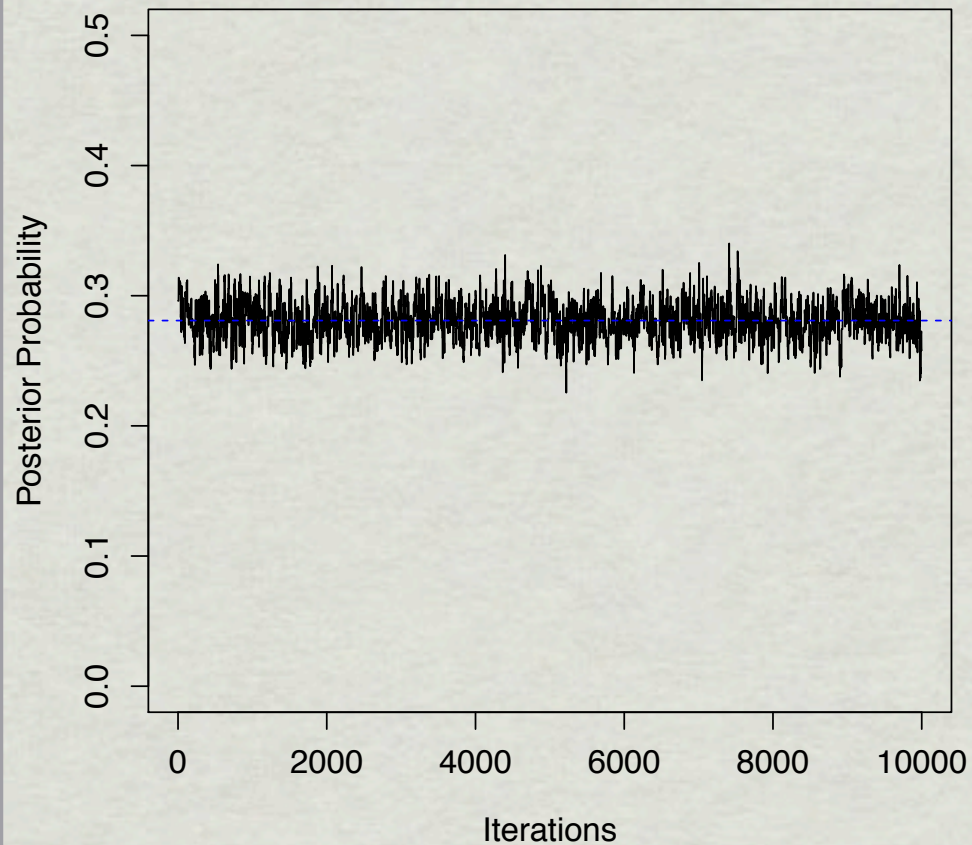


Sensitivity	Specificity	Accuracy	AUC
1.000	0.870	0.872	0.935

Application: PhenoPred Data

9,141 proteins; 294 DO terms; 83% unknown proteins

Auto-Correlation Function Estimation



Application: PhenoPred Data

test set: 33 positives, 122 negatives.

Data set	Iterations	Sensitivity	Specificity	Accuracy	AUC
Complete	2,000	0.576	0.852	0.794	0.721
Complete	50,000	0.788	0.623	0.658	0.757
Partial	2,000	0.788	0.615	0.652	0.742
Partial	10,000	0.848	0.656	0.697	0.795