Journal of Proteomics & Bioinformatics

JPB/Vol.1/August 2008

# Pathway Modeling: New face of Graphical Probabilistic Analysis

**Somnath Tagore[1], Virendra S. Gomase[1*] and Rajat K. De[2]**

[1]Department of Bioinformatics, Padmashree Dr. D.Y. Patil University, Plot No-50, Sector-15, CBD Belapur, Navi Mumbai 400614, India

[2]Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India

*Corresponding author: Virendra S. Gomase, Department of Bioinformatics, Padmashree Dr. D.Y. Patil University, Plot No-50, Sector-15, CBD Belapur, Navi Mumbai 400614, India, E-mail: virusgene1@yahoo.co.in

**Citation:** Somnath T, Virendra SG, Rajat KD (2008) Pathway Modeling: New face of Graphical Probabilistic Analysis. J Proteomics Bioinform 1: 281-286.

## Abstract

**Pathway analysis is one of the most interesting aspects of Systems Biology. Modeling biological pathways is interesting as well as difficult to optimize. Various modeling problems of diseases can be successfully analyzed using this simulation approach. Graphical probabilistic approaches are one of the unique methodologies that are used for designing and analyzing pathways. We have discussed the various graphical approaches that are actively involved in pathway modeling.**

**Keywords:** Pathway modeling; Pathway analysis; Helmholtz machine; HMM

## Introduction

Biological pathways are modeled for analyzing and visualizing various sub-steps of the network, study gene expression profiles and predicting outcome of various alterations made to the cells. A major challenge in developing these models is to choose the correct abstraction. Due to the large and diverse nature of biological networks, it is essential to balance computational complexity against model fidelity and to move between models of different levels of detail, using different meaning ways. Here, graphical probabilistic models are discussed for modeling biochemical pathways. Biological pathways are categorized into Metabolic Pathways, Signal Transduction Pathways and Gene regulatory Networks. Here, we have tried to look into all these aspects of biological pathway modeling.

## Graphical probabilistic models

Graphical Probabilistic Models represent multivariate probability densities. These multivariate probability densities are represented by a product of terms that involves few variables. Furthermore, the products are represented by graph theoretical approach. This graph relates the variables that are represented by a common term. The common types of graphical models are discussed here (Agarwal et al., 2000; Hall et al., 1999).

## Types of graphical probabilistic models

### Bayesian networks

Bayesian Networks are used for predicting relationship within variables. It is a directed acyclic graph whose nodes represent random variables; arcs represent statistical de-

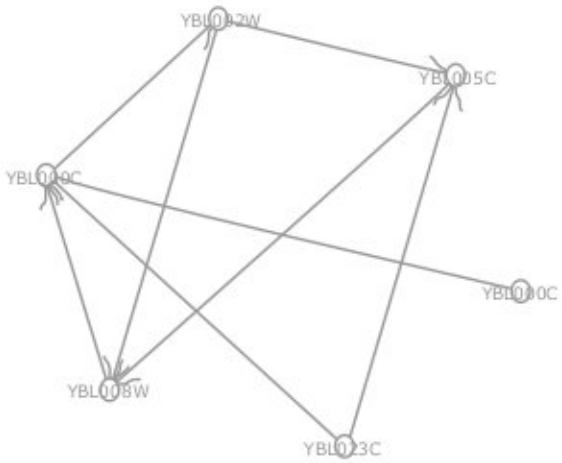


**Figure 1:** Figure showing a gene regulatory network explained using Bayesian statistics.

pendence relations among the variables and local probability distributions for each variable given values of its parents (Levitsky et al., 2007; Marashi et al., 2007).

Thus, for each variable $X_i$,

$$i \bullet \{1, \ldots, N\} \quad (1)$$

the set of parent variables is denoted by parents ($X_i$), then the joint distribution of the variables is product of the local distributions.

$$\Pr(X_1, \ldots, X_{n)} = \text{Đ} \Pr(X_r \mid \text{parent}(X_i)) \quad (2)$$

## Gaussian Networks

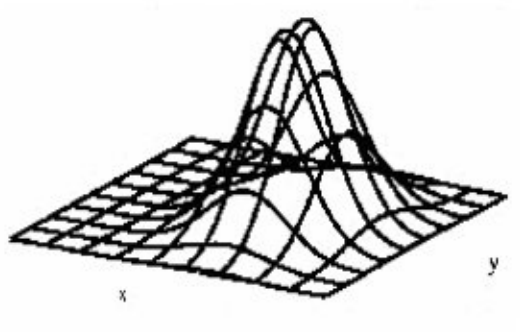


**Figure 2:** Figure showing the Gaussian network

The normal distribution is univariate in nature. But, there is a difficulty working with univariate distribution as the covariance matrix must be positive definite in nature. But with gaussian networks, this constraint needs not to be considered (McKinney et al., 2006).

### Maximum likelihood

Maximum Likelihood Estimation begins with writing a mathematical expression called the Likelihood Function of the sample data. It is the probability of obtaining that particular set of data, given the chosen probability distribution model. This expression contains the unknown model parameters. The values of these parameters that maximize the sample likelihood are known as the Maximum Likelihood Estimators (MLE's) (Hu, 2004; Jin et al., 2008).

Thus, Given a family M{i} of probability distributions parameterized by 'i' associated with a known probability function fn{i}, we may draw a sample x{1} to x{n} of 'n' values from this distribution and then using fn{i} we may compute the probability density (Justenhoven, et al., 2008).

$$[fn\{i\}(x\{1\} \text{ to } x\{n\}) \mid i] \quad [3]$$

In this case, the likelihood function is given by,

$$[L(i)=[fn\{i\}(x\{1\} \text{ to } x\{n\}) \mid i] \quad [4]$$

### Density estimation

Density Estimation is the construction of an estimate based on an un-observed data. This is again based upon an un-observed probability density function (Estivill-Castro and Houle, 2001).

### Helmholtz machine (HM)

Helmholtz Machines are neural networks that learn the hidden structure of a set of data one being trained to create a generative model, producing the original set of data. Thus, by learning the various representations of the data, the underlying structure of the generative model approximates the hidden structure of the data set (Estivill-Castro et al., 2001; Estivill-Castro et al., 2001). These are categorized as Autoencoders, Deterministic HM and Stochastic HM. Autoencoders reconstructs its best guess of the input on the basis of the code that it sees, whereas Deterministic HM is inspired by mean-field methods and Stochastic HM captures the correlation between the activities in different hidden layers (Han and Kamber, 2000).

### Latent variable models (LVM)

Latent Variable Models relates a set of manifest variables to set of latent variables, which are grouped according to whether the manifest and latent variables are categorical or continuous. It provides a means to parse out measurement error by combining across observed variables and allow for the estimation of complex

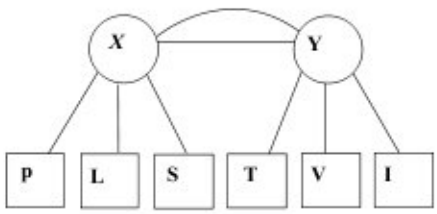


**Figure 3:** Figure shows LVM.

causal models. Furthermore, these are well developed for metric and discrete observed variables. Also, these account for clustering random effects (Tonella, 2001).

### Generative topographic mapping (GTM)

In Generative Topographic Mapping (GTM), the training data is assumed to arise by first picking a point probabilistically in a low-dimensional space, then mapping the point to the high-dimensional input space that is observed. This is done by a smooth function and then adding noise in the high dimensional input space. The Expectation-Maximization (EM) algorithm is used to make a training set that can be used to train the parameters of the low-dimensional probability distribution (Cormen et al., 2000).

### Hidden markov model (HMM)

In a Hidden Markov Model, a state is not directly visible, but variables influenced by the state are visible. Each state

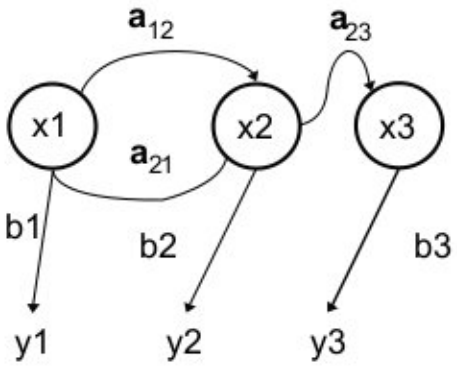


**Figure 4:** Figure showing HMM.

has a probability distribution over the possible output tokens. This model is a finite set of states, each of which is associated with a probability distribution (Demetrescu et al., 2004). Transitions among the states are governed by a set of probabilities called transition probabilities. In a particular state an outcome or observation can be generated, according to the associated probability distribution. The three main problems of HMM include Evaluation Problem, Decoding Problem and Learning Problem (Demetrescu et al., 2003).

## Application of graphical probabilistic models

### Application to metabolic pathway modeling

A machine learning system is introduced for gene functions determination from heterogeneous data sources using a Weighted Naive Bayesian network (WNB). The aim is to infer functions of putative genes or Open Reading Frames (ORFs) from existing databases using computational methods. While integrating evidence from multiple and complementary sources significantly improves the prediction accuracy. The experimental results suggest that the stated hypothesis is valid and provide guidelines for using the WNB system for data collection, training and predictions. Furthermore, the combined training data sets consists results from gene expressions, clustering outputs and sequence homology from public databases. It is also used to analyze the contribution of each source of information toward the prediction performance through the weight training process (Deng et al., 2006).

Searching for peptide hormones that signals via membrane receptors is often hampered by their small size, and lack of sequence similarity. A search tool based on the hidden Markov model is developed that uses various peptide hormone sequence features for estimating the likelihood that a protein contains a processed and secreted peptide of this class. Analysis of the top scoring hypothetical and poorly annotated human proteins identifies two candidate peptide hormones. Their analysis shows that both are localized to secretory granules in a transfected pancreatic cell line. The findings demonstrate the utility of a bioinformatics approach to identify novel biologically active peptides (Mirabeau et al., 2007).

Multivariate methods are used for the analysis of molecular data including genotypic data and clinical phenotypes. These methods include latent variable models and joint multivariate modeling techniques. Thus, given the wide variety in the data considered, the objectives of the analysis and the methods applied, direct comparison of the results are discussed (Beyene et al., 2007).

Major stem cell species are studied using a co-clustering latent variable model (LVM). It helps to explain cell type-specific transcription factors, using expression profiles. The LVM-based study also helps to analyze regulatory modules for each stem cell cluster. Furthermore, the identities of the stem cell clusters are revealed by the constituent genes that are directly targeted by the modules (Joung et al., 2006).

### Application to signal transduction modeling

A primer on the use of Bayesian networks is introduced

for analyzing the connectivity of signaling networks. Bayesian networks are used to derive causal influences among biological signaling molecules. An automatically derive a Bayesian network model is introduced from proteomic data and to interpret the resulting model (Pe'er, 2005).

Stochastic biochemical systems are used for modeling transcriptional regulation in single cells. Transcriptional regulation is easily modeled using a hidden Markov model (HMM). It is used to mathematically and computationally study transcriptional regulation in single cells. Furthermore, analysis by Monte Carlo simulation is computationally laborious. Several simulations are employed based on a transcriptional regulatory system for showing the relative merits and limitations of various approximation techniques (Goutsias, 2006).

Graphical models are very well used for analyzing G-Protein coupled receptors (GPCRs). Most of signaling networks in cells are mediated through the interaction of GPCRs with heterotrimeric GTP-binding proteins (G-proteins). Experimental data suggest that heterotrimeric G-proteins interact with parts of the activated receptor at the transmembrane helix-intracellular loop interface. An exploratory approach is designed to generate a refined library of Hidden Markov Models that predict the coupling preference of GPCRs to heterotrimeric G-proteins. It predicts the coupling preferences of GPCRs to Gs, Gi/o and Gq/11, but not G12/13 subfamilies (Sgourakis et al., 2005).

A Hidden Markov model library is designed for classifying protein kinases into 12 families. This classification is also coupled with a mis-classification rate of zero on the characterized kinomes of H. sapiens, M. musculus, D. melanogaster, C. elegans, S. cerevisiae, D. discoideum, and P. falciparum. This is applied to 38 unclassified kinases of yeast including AGC (5), CAMK (17), CMGC (4), and STE (1). It also facilitates the annotation of kinomes and provides data regarding early evolution and subsequent adaptations of the various protein kinase families (Miranda-Saavedra et al., 2007).

### Application to gene regulatory networks

Gene regulatory networks are modeled using probabilistic Boolean network methods and dynamic Bayesian network methods. These methods are compared using certain biological time-series dataset from the Drosophila Interaction Database for designing Drosophila gene network. Also, a subset of time points and gene samples from the whole dataset is used to evaluate the performance of these two approaches (Li et al., 2007).

A hierarchical hidden Markov regression model is introduced for determination of gene regulatory networks from genomic sequence and gene expression microarray data. A hybrid Monte Carlo methodology is devised to estimate parameters under 2 classes of latent structure. One is arising due to the unobservable state identity of genes and the other is due to the unknown set of covariates influencing the response within a state (Gupta et al., 2007).

A comparative gene predictor, called Conrad is proposed, based on semi-Markov conditional random fields (SMCRFs). It is trained to maximize annotation accuracy. It encodes information as features and treats all features equally in the training and inference algorithms. On Cryptococcus neoformans, configuring Conrad to reproduce the predictions of a two-species phylo-GHMM closely matches the performance of Twinscan. Furthermore, it produces similar results on Aspergillus nidulans comparing Conrad versus Fgenesh (DeCaprio et al., 2007).

Hidden Markov Models are compared with genotyping to determine the transmission characteristics of sporadic vancomycin-resistant enterococci (VRE). For this, a structured continuous-time hidden Markov model (HMM) is developed. Two parameters are estimated, one to quantify the cross-transmission of VRE and the other to quantify the level of VRE colonization from sporadic sources. Some evidence is found, based on model selection criteria that the cross-transmission parameter changed throughout the study period. This model estimates that cross-transmission increases at week 120 and declines after week 135, coinciding with environmental decontamination. HMMs are also applied to serial prevalence data to estimate the characteristics of acquisition of nosocomial pathogens and distinguish between epidemic and sporadic acquisition (McBryde et al., 2007).

## Current Research

Bayesian networks are used for predicting interaction partners using multiple alignments of interacting protein domains sequences without the need for any training examples. This also accurately predicts interaction partners in datasets of polyketide synthases. Also, analysis of the predicted genome-wide two-component signaling networks shows that interacting kinase/regulator pairs, which lie adjacent on the genome and which lie isolated form two relatively independent components of the signaling network in each genome (Burger and van Nimwegen, 2008).

A hidden Markov model is used for predictive modeling of nuclear hormone receptor response elements coupled with chromatin microarray technology explains a binding site in the Type I human hepatic 3alpha-hydroxysteroid dehydrogenase (AKR1C4) promoter for the nuclear hormone receptor liver X receptor alpha. It also suggests that LXRalpha modulate the bile acid biosynthetic pathway at a unique site downstream of CYP7A1 (Stayrook et al., 2008).

The probable state path of three nucleotides sequences of cis-regulatory region of target genes are identified using a Hidden Markov Model (HMM). These regions are key elements in the transcriptional regulation of gene expression. These computations are also used to predict C(2)H(2) zinc finger transcription factor binding sites in cis-regulatory regions of their target genes (Cho et al., 2008).

Certain Markov matrix (MMM) values are used to characterize numerically 81 sequences of type III RNases and 133 proteins of a control group. Also one MMM-QSAR and one classic hidden Markov model (HMM) is developed based on the same data. The MMM-QSAR shows a discrimination power of RNAses from other proteins of 97.35% without using alignment, which is a result as good as for the known HMM techniques. Furthermore, the MMM-QSAR model predicts the new RNase III with the same accuracy as other classical alignment methods (Agüero-Chapín et al., 2008).

## Conclusion

Graphical probabilistic models are of much importance in Systems biology, especially in analyzing and modeling biological networks. Bayesian Networks have large applications in almost every field of life science ranging from gene expression analysis, genetic/metabolic network analysis and pathway modeling. Gaussian Networks are applied to analyze various interaction networks like protein-protein, gene-gene and gene-protein. Pathway modeling is also done based on this method. Maximum Likelihood is used in phylogenetic estimates, study genetic cross-over, pathway modeling and gene expression analysis. Density Estimation is useful for certain immunological or clinical trials, metabolic network analysis and pathway modeling. Helmholtz Machine (HM) is used in studying metabolic activities of brain and nervous system. Latent Variable Models (LVM) is used for studying various regulatory networks, pathway modeling and gene expression profiles. Generative Topographic Mapping (GTM) is used in microarray analysis, gene expression level analysis and pathway modeling. Lastly, Hidden Markov Models (HMM) are used in protein structure analysis, sequence analysis, metabolic pathway analysis, gene expression analysis and promoter region identification.

## References

1. Agarwal R, Bayardo RJ, Srikant R (2000) Athena: Mining-Based Interactive Management of Text Database, Proceedings of the 7th International Conference on Extending Database Technology: Advances in Database Technology 365-379.

2. Agüero-Chapín G, Gonzalez-Díaz H, de la Riva G, Rodríguez E, Sanchez-Rodríguez A, et al. (2008) MMM-QSAR recognition of ribonucleases without alignment: comparison with an HMM model and isolation from Schizosaccharomyces pombe, prediction, and experimental assay of a new sequence. J Chem Inf Model 48: 434-448.

3. Beyene J, Tritchler D, Bull SB, Cartier KC, Jonasdottir G, et al. (2007) Multivariate analysis of complex gene expression and clinical phenotypes with genetic marker data. Genet Epidemiol 1: S103-S109.

4. Burger L, van Nimwegen E (2008) Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Mol Syst Biol 4:165.

5. Cho SY, Chung M, Park M, Park S, Lee YS (2008) ZIFIBI: Prediction of DNA binding sites for zinc finger proteins. Biochem Biophys Res Commun 369: 845-848.

6. Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) Introduction to Algorithms, McGraw-Hill.

7. DeCaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M, et al. (2007) Conrad: gene prediction using conditional random fields. Genome Res 17: 1389-1398.

8. Demetrescu C, Emiliozzi S, Italiano GF (2004) Experimental analysis of dynamic all pairs shortest path algorithms. In Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'04).

9. Demetrescu C, Finocchi I, Italiano GF (2003) Algorithm engineering, Bulletin of the EATCS 79: 48-63.

10. Deng X, Geng H, Ali HH (2006) Joint learning of gene functions-a Bayesian network model approach. J Bioinform Comput Biol 4: 217-239.

11. Estivill-Castro V, Houle ME (2001) Data Structures for

Minimization of Total Within-Group Distance for Spatio-temporal Clustering, Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery 91-102.

12. Estivill-Castro V, Houle ME (2001) Fast minimization of total within-group distance. In J. Fong and M. Ng, editors, Proceedings of the International Workshop on Mining Spatial and Temporal Data in conjunction with the fifth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-2001) 72-81.

13. Goutsias J (2006) A hidden Markov model for transcriptional regulation in single cells. IEEE/ACM Trans Comput Biol Bioinform 3: 57-71.

14. Gupta M, Qu P, Ibrahim JG (2007) A temporal hidden Markov regression model for the analysis of gene regulatory networks. Biostatistics 8: 805-820.

15. Han J, Kamber M (2000) Data mining: concepts and techniques, Morgan Kaufmann Publishers Inc.

16. Hall I, Özyurt LO, Bezdek J (1999) Clustering with a genetically optimized approach. IEEE Transactions on Evolutionary Computation 3: 103-112.

17. Hu S (2004) Optimal time points sampling in pathway modeling. Conf Proc IEEE Eng Med Biol Soc 1: 671-674.

18. Jin S, Zhang Y, Yi F, Li PL (2008) Critical role of lipid raft redox signaling platforms in endostatin-induced coronary endothelial dysfunction. Arterioscler Thromb Vasc Biol 28: 485-490.

19. Joung JG, Shin D, Seong RH, Zhang BT (2006) Identification of regulatory modules by co-clustering latent variable models: stem cell differentiation. Bioinformatics 22: 2005-1011.

20. Justenhoven C, Hamann U, Schubert F, Zapatka M, Pierl CB, et al. (2008) Breast cancer: a candidate gene approach across the estrogen metabolic pathway. Breast Cancer Res Treat 108: 137-149.

21. Levitsky VG, Ignatieva EV, Ananko EA, Turnaev II, Merkulova TI, et al. (2007) Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. BMC Bioinformatics 8: 481.

22. Li P, Zhang C, Perkins EJ, Gong P, Deng Y (2007) Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks. BMC Bioinformatics 8 Suppl 7: S13.

23. Marashi SA, Kargar M, Katanforoush A, Abolhassani H, Sadeghi M (2007) Evolution of 'ligand-diffusion chreodes' on protein-surface models: a genetic-algorithm study. Chem Biodivers 4: 2766-2771.

24. McBryde ES, Pettitt AN, Cooper BS, McElwain DL (2007) Characterizing an outbreak of vancomycin-resistant enterococci using hidden Markov models. J R Soc Interface 4: 745-754.

25. McKinney BA, Crowe JE, Voss HU, Crooke PS, Barney N, et al. (2006) Hybrid grammar-based approach to nonlinear dynamical system identification from biological time series. Phys Rev E Stat Nonlin Soft Matter Phys 73: 021912.

26. Mirabeau O, Perlas E, Severini C, Audero E, Gascuel O, et al. (2007) Identification of novel peptide hormones in the human proteome by hidden Markov model screening. Genome Res 17: 320-327.

27. Miranda-Saavedra D, Barton GJ (2007) Classification and functional annotation of eukaryotic protein kinases. Proteins 68: 893-914.

28. Pe'er D (2005) Bayesian network analysis of signaling networks: a primer. Sci STKE 2005: pl4.

29. Sgourakis NG, Bagos PG, Papasaikas PK, Hamodrakas SJ (2005) A method for the prediction of GPCRs coupling specificity to G-proteins using refined profile Hidden Markov Models. BMC Bioinformatics 6: 104.

30. Stayrook KR, Rogers PM, Savkur RS, Wang Y, Su C, et al. (2008) Regulation of human 3 alpha-hydroxysteroid dehydrogenase (AKR1C4) expression by the liver X receptor alpha. Mol Pharmacol 73: 607-612.

31. Tonella P (2001) Concept analysis for module restructuring. IEEE Transactions on Software Engineering 27: 351-363.