ISSN : 0974 - 7435 Volume 6 Issue 1

Trade Science Inc.

BioTechnology

An Indian Journal

Review

BTAIJ, 6(1), 2012 [16-21]

A probabilistic approach for the analysis of free-energy distribution in proteins

G.Padma*, C.Vijayalakshmi

Department of Mathematics, Sathyabama University, Chennai, (INDIA)

E-mail : govindanpadma1970@gmail.com; vijusesha2002@yahoo.co.in

Received: 27th September, 2011 ; Accepted: 5th December, 2011

ABSTRACT

Probabilistic inference in an MRF involves computing Marginal distributions over the random variables in the graph. Graph formalism is an effective and efficient representation for multivariate independence structure for both model construction and for inference. The use of graphs to represent independence structure in multivariate probability models has been pursued in a relatively independent fashion across a wide variety of research disciplines. Traditional graphical models decompose the joint distribution as a product of functions of subsets of variables. However, a number of rigorous approximation algorithms have been devised for performing inference in MRFs. Factor graphs are more useful for describing models that involve a large number of overlapping relationships between variables. When compared to Bayesian Networks (BNs) and Markov Random Fields (MRFs) Factor graph model decomposes interactions between variable. While functional relationships between variables in BNs and MRFs must be determined by identifying parent-child clusters or maximal cliques, Factor graphs explicitly identify functional relationships. Any Bayesian network or Markov random field can be represented as a factor graph. Belief propagation algorithm is used for finding the Marginal probability of the any hidden Variable conditioned on the observed variable. The algorithm is designed by passing real valued functions called messages along the edges between the nodes. This paper analyses the application of the belief propagation on Biological Markov random field with an example.

KEYWORDS

Markov Random Fields;
Bayesian Networks;
Factor graphs;
Belief networks.

INTRODUCTION

Many studies in recent years address the challenge of constructing protein–protein interaction networks. Several experimental assays, such as yeast two-hybrid[13] and tandem affinity purification[11] have facilitated high-throughput studies of protein–protein interactions on a genomic scale. Some computational approaches aim to detect functional relations between proteins, based on various data sources such as phylo-

genetic profiles[10] or mRNA expression[3]. Other computational assays try to detect physical protein–protein interactions by, for example, evaluating different combinations of specific domains in the sequences of the interacting proteins[12].

While the above combined approaches lead to an improvement in prediction, they are still inherently limited by the treatment of each interaction independently of other interactions. By explicitly modeling such dependencies, we can leverage observations from varied sources to produce better joint predictions of the protein interaction network as a whole. As a concrete example, consider the budding yeast proteins Pre7 and Pre9, given in Figure 1.

Figure 1

These proteins were predicted to be interacting by a computational assay[12]. However, according to a large-scale localization assay[5], the two proteins are not co-localized; Pre9 is observed in the cytoplasm and in the nucleus, whilePre7 is not observed in either of those compartments. Based on this information alone, it can be concluded that an interaction between the two proteins is improbable. However, additional information on related proteins may be relevant. For example, interactions of Pre5 and Pup3 with both Pre9 and Pre7 were reported by large scale assays[9, 10]. This example illustrates two reasoning patterns that we would like to allow in our model. First, to encode that certain patterns of interactions (e.g., within complexes) are more probable than others. Second, an observation relating to one interaction should be able to influence the attributes of a protein (e.g., localization), which in turn will influence the probability of other related interactions.

Proteins are chains of simpler molecules called amino acids. An amino acid unit in a protein is called a residue. Each amino acid has a chemical group called the side-chain. This group distinguishes one amino from another. Amino acids are joined end to end during protein synthesis by the formation of peptide bonds. Formation of a succession of peptide bonds generates the backbone upon which the side chains are hanged. Predicting side-chains conformation given the backbone structure is a central problem in protein-folding and molecular design. For many the applications it is not enough to calculate a single, most probable side chain configuration but the interest is mainly in the calculation of the free energy distributions. Due to the importance of this problem in computational biology large number of special algorithms has been developed.

Biomolecular systems are governed by changed in free energies provides better understanding of bio molecular interactions and the ability of optimize them. A probabilistic representation confers several advantages, including that it provides a structural changes due to changes in temperature, pH, ligand binding and mutation can become an inference problems over the model. Recent advances in inference algorithms for graphical models such as generalized belief propagation is a rigorous approximation to the free energy of the system[14]. These free energy estimates are accurate enough to perform non-trivial tasks within structural biology.

The free energy is defined as G = E – TH, where E is the enthalpy of the system, T is the absolute temperature and H is the entropy of the system. The entropy estimates are difficult because they involve sums over an exponential number of states. Hence the entropy term is often ignored altogether under the assumption that it does not contribute significantly to the free energy[1] proved that energy functions comprising sums of pairwise interactions cannot distinguish a protein's native structure from decoy structures. Hence entropy contributions become significant when the structures are similar. It is proved that the native structure is usually the one with highest entropy and hence[8] entropy should be included in the energy calculations.

A MARKOV RANDOM FIELD MODEL FOR PROTEIN STRUCTURE

A protein consists of finite number of atoms across one or more polypeptide chains. A configuration of the protein corresponds to the geometry of each of its consistent atoms. An accurate representation of the protein is the ensemble of configurations at room temperature. Let the configuration of a protein be regarded as a random variable in some configurational space C. Let x denote the random variable which corresponds to the configuration of the entire set of atoms in the protein.

Review

The probability of a configuration $x_c \in C$ with internal energy E_c is

$$P(x = x_c) = \frac{1}{z}\exp\left(\frac{-E_c}{K_B T}\right)$$

where $z = \sum_{x_c \in C} e^{-E_c}$ is the partition function, K_B is the Boltzmann's constant and T is the absolute temperature in Kelvin.

The entire set of atoms is partitioned as 2 disjoint subsets as backbone and side bone chains. Backbone atoms refers to those that are common to all 20 amino acid types, while side chain atoms are those that differ among the different kinds of amino acids.

Let the backbone variables are represented by X_b and the side bone variables are represented by X_s, then the protein representing the conformation of backbone atoms are the

$X_b = \{x_{b_1}, x_{b_2}, ...\}$ and $X_s = \{x_{s_1}, x_{s_2}, ...\}$ where each x_{s_i} represents the conformation of the side chain atom of residue i. Let the energies of X_b, X_s are E_b, E_s. The joint distribution and the partition function be defined as

$P(X = X_c) = P(X_b = x_b)\, P(X_s = x_s / X_b = x_b)$

$$Z = \sum_{x_b} e^{\left(-\frac{E_b}{K_B T}\right)} Z_b \qquad (1)$$

where $Z_b = \sum_{x_s} e^{\left(-\frac{E_s}{K_B T}\right)}$ is the partitions function over the side chain space with a fixed backbone.

Given a specific backbone trace X_b due to the nature of the physical forces in action, pairs of residues distally located according to trace are expected to exert very little direct influence on one another. Such residues are independent of each other when conditioned on X_b. These conditional independencies are compactly independencies are compactly encoded as a Markov Random field (MRF).

An MRF is a probability distribution on over a graph and can be represented as a tuple (x, ε, ϕ) where the set of random variables in the multivariate distribution are the set of vertices X_s and X_b. The edges $e \in \varepsilon$ join residues that are directly dependent on each other and $\phi = (\phi_1, \phi_2, \ldots, \phi_m)$ is a set of functions called as factors over the random variables. In general an MRF encodes the following conditional independencies for each vertex X_i and for any set of vertices X′ not containing X_i as

$P(X_i / X', \text{Neighbors}(X_i)) = P(X_i / \text{Neighbors}(X_i))$

That is a random variable X_i is conditionally independent of every other set of nodes in the graph, given its immediate neighbors in the graph[7].

$$P(X_s = x_s / X_b = x_b) = \frac{1}{Z_b}\prod_{\varphi_i \in \varphi} \varphi_i(x\varphi_i) \qquad (2)$$

where Z_b is the so-called partition function.

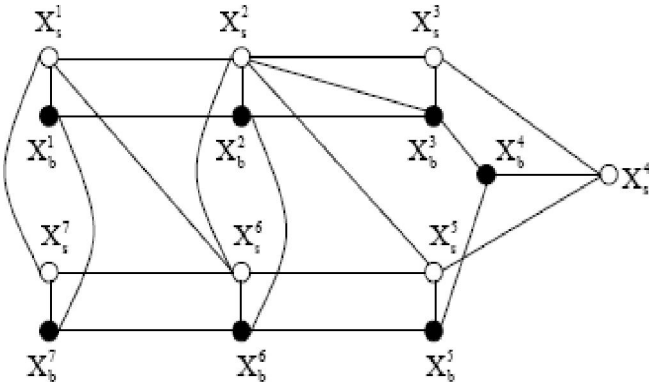

Figure 2 : Part of the random field induced by the outlined residues x_s^i 's are hidden variables representing the rotameric state, the visible variables are the backbone atoms in conformations x_b.

$$Z_b = \sum_{X_i}\prod_{\varphi_i \in \varphi} X_{\varphi_i}$$

Also $\varphi_i(x_\varphi) = \exp\left(-\frac{E(x_{\varphi_i})}{K_B T}\right)$ where x_{φ_i} a is the set of atoms that serve as arguments to φ_i and $E(x_{\varphi i})$ is the potential energy of those atoms as defined by a molecular force field. The joint distribution is obtained by simply multiplying (2) with the probability of a particular backbone conformation x_b according to (1). Thus the probability of a given state is simply the product of the functions, suitably normalized.

PROBABILISTIC INFERENCE AND FREE ENERGY CALCULATIONS

Probabilistic inference in an MRF involves computing the marginal distributions over the random variables in the graph. Inference and free-energy approxi-

Review

mations require the estimation of a partition function. The belief propagation algorithm starts with random initial beliefs[4], and then use messages passing between nodes to converge on a final set of beliefs.

Informally each node updates its own beliefs base on the beliefs of its neighbors in the graph and the value of the potential functionϕ. When the algorithm converges the final beliefs can be used to obtain the partition function, and hence the free energy.

If the MRF happens to form a tree (i.e., a graph with no cycles). Belief propagation is exact and takes O (|ε|) times, where |ε| is the number of edges in the graph. Here the algorithm used in generalized belief propagation on MRF which encodes the conditional distribution with the estimation of the partition function for each backbone configuration X_b[6].

Thus the undirected graphical model (MRFs) characterized by the variable X and the potential function ϕ is a bipartite graph (X, F) called a factor graph. If it is restricted with pairwise potentials then the equivalent factor graph for the MRF of Figure 2 is shown as Figure 3.

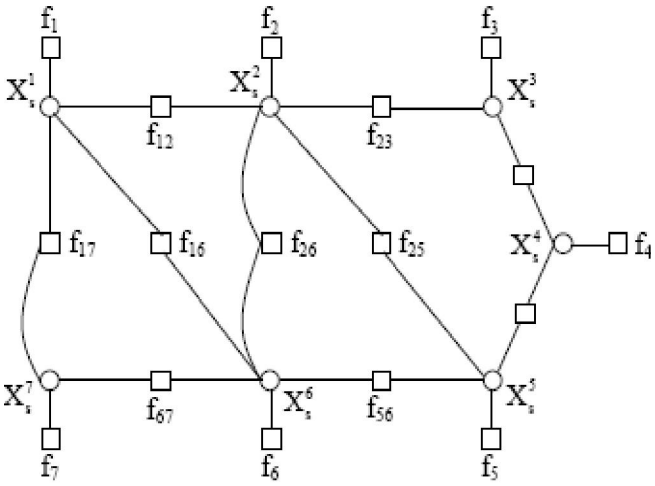

Figure 3 : Factor graph representation for Figure 2. The observed variables corresponding to the backbone atoms can be replaced by a factor f_i at each side chain variable.

The generalized belief propagation (GBP) is a message passing algorithms that approximates the true marginals. This differs from the belief propagation (BP) algorithm in the size of its regions that estimates the free energy. GBP computes fixed points of the more general region based free energy. Let the pseudo message for a region R with parents P(R) and children O(R) be given by $\mathbf{n}^0_{R\to P}(X_r) = \tilde{f}_R(X_R) \prod_{p'\in P(R)} mp' \to R(x_r) \prod no$

$$\mathbf{m}^0_{R\to o}(X_o) = \sum_{x_R \setminus x_o} \tilde{f}_R(X_R) \prod_{p\in P(R)\setminus p} \mathbf{m}_p \to \mathbf{R}(x_R) \prod_{O'\in O(R)\setminus O} \mathbf{n}_{o'} \to_R (x_{o'})$$

where $\tilde{f}_R(X_R) = (\prod_{a\in A_r} f_a(x_a))^{\omega R}$ and then compensating for over counting by defining the actual messages as

$$\mathbf{n}_{R\to P}(X_r) = (\mathbf{n}^0_{R\to P}(X_r))^{\beta R} (\mathbf{m}^0_{R\to o}(X_o))^{\beta R-1}$$

$$\mathbf{m}_{R\to P}(X_r) = (\mathbf{n}^0_{R\to P}(X_r))^{\beta R-1} (\mathbf{m}^0_{R\to o}(X_o))^{\beta R}$$

where ω_R is the weight given to region R, p_R the number of parents of region R, and $\beta_R = p_R/(2p_R + \omega_R - 1)$

The beliefs at R, are then given by

$$\mathbf{b}_R(X_r) = \tilde{f}_R(X_R) \prod_{O\in O(R)} \mathbf{n}_o \to_R (x_o) \prod_{p\in P(R)} \mathbf{m}_p \to \mathbf{R}(x_p)$$

Note that if $\beta_R = 1$ this algorithm becomes equivalent to running BP directly on the region graph.

The algorithm is typically started with randomly initialized messages and run until the beliefs converge. If it does converge, GBP is guaranteed to find a fixed point of the region based free energy.

Inferences from the probabilistic graphical models have been used for a number of problems in the area of secondary structure prediction[2]. The applications of graphical models to tertiary structure are limited to application of Hidden Markov models (HMM)[7]. HMMs make severe independence assumptions to allow for efficient learning and inference.

The focus is mainly on the problem of computing entropy using marginal probabilities for the unobserved variables, X_s. Variables has been replaced by edges to a factor f_i representing the interactions between these variables can be dropped from the factor graph by replacing their interactions with each side chain variable by a factor. Hence the probability of a particular conformation can be expressed using the factor notation

$$\mathbf{P}(X_s) = \frac{1}{Z} \prod_{fa\in F} \mathbf{fa}(X_s^a)$$

Where X_s^a is a set of variables connected to the factor 'fa' in the factor graph.

Review

APPROXIMATING FREE ENERGY

A corollary of the second law of thermodynamics is that a physical system seeks to minimize its freeenergy. Thus, the most accurate entropy estimates are obtained when the system has the least free energy. Under the assumption of constant temperature, the free energy of a system is given by

G= E -TH

Where E is the enthalpy of the system, T the temperature and H, the entropy. If we associate a belief b(x) with state x, this can be rewritten as

$$G = \sum_{x} b(x)E(x) + \sum_{x} b(x) \ln (b(x))$$

Where the first term and second terms on the right are the enthalpic and entropic contributions respectivelyand the summation is over all possible x. Intuitively, the enthalpic term corresponds to the energy of thesystem. However, the second law of thermodynamics dictates that not all energy can be used to do work. The free energy is the energy left to be used to do work after deducting the energy that is "lost" which isthe entropic deduction mentioned above. There has been a considerable amount of work by physicists at developing approximations to estimatethese terms. The popular methods are based on approximating the free energy using a region based free energy. Intuitively, the idea is to breakup the factor graph into a set of regions, R, each containing multiple factors f_R and variables X_R, computethe free energy over the region using estimates of the marginal probability over X_R, and then approximatethe total free energy by the sum of the free energies over these regions. Since the regions could overlap, contributions of nodes—factors or variables—which appear in multiple regions have to be subtracted out, so that each node is counted exactly once. This can be done by associating weights w_{Ri} to the contribution of every node in region region $R_i \varepsilon R$, in such a way that the sum of weights of the regions that the node appears in sums to one.

IMPLEMENTATION AND RESULTS

The Two-Way GBP algorithm is implemented to compute region graph estimates of free energy and entropy. The factor graph is created by computing inter atomic distances and creatinga factor between residues if the C_α distance between them was lesser than a threshold value. This threshold is largely dictated by the sensitivity of the energy function. For the energy terms we used, we found a threshold of 8.0 Å to be adequate. Each rotamer in the library also had an associated apriori probability which were incorporated into the factor as a prior. The temperature of the system was taken to be 300 K, whichcorresponds to normal room temperature. Then there were two levels of regions. The top level contained "big"regions—regions with more than one variable—while the lower level contained regions representing singlevariables. Since the interaction between residues closest in sequence to be very strong, all factors and nodes were placed between residues within two sequence positions of each other in one region. Each of the rest of the factors, representing edges between residues connected in space, formed "big"regions with two nodes in them. Thus, in the example shown in Figure 2, (X_1s ;X_2s ;X_3s ; f_1; f_2; f_3; f_{12};f_{23}),(X_2s ;X_3s ;X_4s ; f_2; f_3; f_4; f_{23}; f_{34}), and (X_1s ;X_7s ; f_{17}) would be examples of big regions which appear in thetop level, while (X_1s) would be an example of a small region in the lower level. Finally, edges from the "big" regions to all small regions that contain a strict subset of the "big" region's nodes are added. In the aboveexample, the region encompassing X_1s ;X_2s ;X_3s would thus be connected to the small regions corresponding to each of X_1s, X_2s, and X_3s. Since the region graph formalism is very flexible, other equally valid alternatives for creating the graph exist. The best choice of regions will largely depend on the application at hand and the computational constraints. The choice of regions reflects a balance between accuracy and running time by focussing on residues which are expected to be closely coupled together and placing them in bigger regions.

CONCLUSION

Fast and accurate free energy calculations are essential to a number of significant tasks within computational structural biology including structure based protein. Side-chain prediction is an important subtask in the protein-folding problem. It is shown that finding a minimal energy side-chain configuration is equivalent to performing inference in an undirected graphical model.

The graphical model is relatively sparse yet has many cycles and this equivalence is used to assess the performance of approximate inference algorithms in a real-world setting. The probabilistic graphical model based approach to all atom free energy calculations strikes a balance between the physical methods and the speed of the statistical methods. Since it uses all atom force fields when computing internal energies and computes a better approximation of the true partition function of the system. Also this method is competitive with statistical methods in terms of speed.

REFERENCES

[1] M.R.Betan Court, D.Thirumalai; Protein Sci., **8**, 361-369 **(1999)**.

[2] W.Chu et al., Z.Ghahramani, D.Wild; Proc.21stAnn.ICML, 161-168 **(2004)**.

[3] M.B.Eisen, P.T.Spellman, P.O.Brown, D.Botstein; Proc.Natl.Acad.Sci.USA, **95(25)**, 14863-14868 **(1998)**.

[4] G.Padma, C.Vijayalakshmi; Implementation of Belief Propagation Iterative Method on Markov Chains By Designing Bayesian Networks, International Journal of Artificial Intelligent Systems and Machine Learning, Issue **(2011)**.

[5] W.K.Huh, J.V.Falvo, L.C.Gerke, A.S.Carroll, R.W.Howson, J.S.Weissman, E.K.O'Shea; Nature **425**, 686-691 **(2003)**.

[6] H.Kamisetty, E.P.Xing, C.J.Langmead; In Proc.RECOMB, 366-380 **(2007)**.

[7] K.Karplus, R.Karchin, J.Draper et al.; Proteins, **53**, 491-496 **(2003)**.

[8] R.Lilien, B.Stevens, A.Anderson et al.; Journal of Computational Biology, **12**, 740-761 **(2005)**.

[9] H.W.Mewes, J.Hani, F.Pfeiffer, D.Frishman; Nucl.Acids Res., **26**, 33-37 **(1998)**.

[10] M.Pellegrini, E.Marcotte, M.Thompson, D.Eisenberg, T.Yeates; Proc.Natl.Acad.Sci.USA, **96(8)**, 4285-4288 **(1999)**.

[11] G.Rigaut, A.Shevchenko, B.Rutz, M.Wilm, M.Mann, B.Seraphin; Nature Biotechnol., **17(10)**, 1030-1032 **(1999)**.

[12] E.Sprinzak, S.Sattath, H.Margalit; J.Mol.Biol., **327(5)**, 919-923 **(2003)**.

[13] P.Uetz, L.Giot, G.Cagney et al.; Nature, **403(6770)**, 623-627 **(2000)**.

[14] J.S.Yedidia, W.T.Freeman, Y.Weiss; IEEE Trans.Inform.Theory, **51**, 2282-2312 **(2005)**.