Computational Biology and Chemistry 65 (2016) 21–28

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

A post-decoding re-ranking algorithm for predicting interacting residues in proteins with hidden Markov models incorporating long-distance information

Colin Kern, Li Liao*

Department of Computer and Information Science, University of Delaware, Newark, DE 19716, USA

ARTICLE INFO

Article history:
Received 11 March 2016
Received in revised form
22 September 2016
Accepted 22 September 2016
Available online 29 September 2016

Keywords:
Protein–protein interaction
Genetic algorithm

ABSTRACT

Protein–protein interactions play a central role in the biological processes of cells. Accurate prediction of the interacting residues in protein–protein interactions enhances understanding of the interaction mechanisms and enables *in silico* mutagenesis, which can help facilitate drug design and deepen our understanding of the inner workings of cells. Correlations have been found among interacting residues as a result of selection pressure to retain the interaction during evolution. In previous work, incorporation of such correlations in the interaction profile hidden Markov models with a special decoding algorithm (ETB-Viterbi) has led to improvement in prediction accuracy. In this work, we first demonstrated the sub-optimality of the ETB-Viterbi algorithm, and then reformulated the optimality of decoding paths to include correlations between interacting residues. To identify optimal decoding paths, we propose a post-decoding re-ranking algorithm based on a genetic algorithm with simulated annealing and show that the new method gains an increase of near 14% in prediction accuracy over the ETB-Viterbi algorithm.

1. Introduction

Protein–protein interactions (PPI) play essential roles in many biological processes in the cell. A central task in systems biology is to study the cell in a holistic way, essentially by reconstructing various biological networks including protein–protein interaction networks. Despite the advancements in experimental technologies, such as yeast two-hybrid (Y2H) systems and coimmunoprecipitation (CoIP), for detecting PPIs (Uetz et al., 2000), the current experimental methods are still noisy, giving inconsistent results from different methods (Braun et al., 2009). In addition, the cost and other limitations inherent in the experimental methods, such as detecting transient interacting partners, have motivated development of computational methods for predicting PPIs.

To better understand why and how two proteins interact, it is important to identify the residues that are involved in protein–protein and protein–ligand interaction; such knowledge also has practical impact, such as serving as a guide in designing mutants to modulate the interaction affinity for different purposes, which is common in drug design. Due to the selection pressure exerted on the amino acids on the interacting interface, the sequence segments, called domains, in which the interface resides tend to be highly conserved during evolution. Therefore, domain identification can be used as a first step towards detecting the interacting interface. However, domain identification poses a challenging task itself. While highly conserved, the domain sequences nonetheless have sustained mutations as well and therefore no clear sequence patterns based on amino acid compositions can be readily available for domain identification. This issue manifests in the mismatches when sequences that contain the same domain are aligned in a multiple sequence alignment. Hidden Markov models (HMMs) are probabilistic models that are trained to capture both the commonalities and sequence variations for a set of proteins, and have been successfully applied in identifying protein domain families. In the Pfam database, HMMs have been built for many common protein domains and families and can be used to identify domain occurrence for query sequences (Finn et al., 2006).

When it comes to detecting interface domains, because the interaction sites impose strong constraints, it is important to incorporate these constraints into the computational methods for more accurate identification of the domains and interacting residues. However, many proteins have not had their structures determined from X-ray crystallography experiments due the cost and experimental difficulties, and hence lack the interaction site information that can be derived from the structure in interacting complexes. At present, the dataset of structurally resolved interacting

* Corresponding author.
E-mail address: liliao@udel.edu (L. Liao).

http://dx.doi.org/10.1016/j.compbiolchem.2016.09.015
1476-9271/

complexes remains relatively small. To tackle this issue, a method called interaction profile hidden Markov model (ipHMM) has been developed (Friedrich et al., 2006). This new model is based on the ordinary profile hidden Markov model (pHMM) (Eddy, 1998) but modifies the model architecture by introducing new match states to specifically represent residues on the interface. Trained with interface sequences that are determined based on 3D structure of protein complexes, the model can then be used to predict interacting domains for proteins with no experimentally determined structural information. The work of Friedrich et al. (2006) reports improved accuracy in identifying interacting domains and interacting residues when using ipHMM as compared to ordinary profile HMMs.

However like most hidden Markov models, the ipHMMs are not suitable for modeling long-distance correlations due to the use of an essentially first-degree Markov chain for the hidden states. Although in principle long-distance correlations can be accommodated by introducing high degree Markov chains, the computational complexity will increase significantly. On the other hand, significant correlations among the interacting residues, sometimes separated by dozens of amino acids in the primary structure, have been reported (Crooks and Brenner, 2004; Gonzalez, 2009). We have previously developed a novel decoding algorithm for HMMs to incorporate this long-distance correlation into the path prediction (Kern et al., 2013). While this algorithm shows significant improvement in the prediction accuracy, it is not guaranteed to produce the highest scoring path like the unmodified Viterbi algorithm is. In this paper, we developed a method of post-decoding re-ranking to find paths through ipHMM that better capture the long-distance correlations and, as a result, further improve prediction accuracy.

2. Method

In this section, we will first briefly review the ipHMM and ETB-Viterbi algorithm as described in the previous work (Kern et al., 2013) so that we can show the sub-optimality with the ETB-Viterbi algorithm. Then we reformulate the optimality problem to include the long-distance correlation. Lastly we present our post-decoding re-ranking method based on a genetic algorithm with simulated annealing.

2.1. Interaction profile hidden Markov models

Like Markov chains, hidden Markov models (HMMs) are models for generating stochastic sequences. However, unlike Markov chains, only the observed sequence is known in an HMM, while the underlying states that correspond to the sequence are hidden. HMMs are useful for modeling problems where easily observed sequences are statistically connected to underlying states which are harder to measure (Rabiner and Juang, 1986). To represent such a model, a directed graph is used where each node corresponds to one of the hidden states S_i ($i = 1$ to N). At each state, a symbol is emitted from a predefined alphabet. The probability of emitting each symbol depends on the current state and is defined as $e_i(x)$ where i is the current state and x is the symbol emitted. After emitting each symbol, the current state changes based on the transition parameters a_{ij}, which is the probability of moving from state S_i to state S_j for every directed edge in the graph. The model parameters for a HMM can be estimated from training data using standard procedures such as maximum likelihood estimation or the Baum-Welch algorithm (Durbin et al., 1998).

HMMs were introduced into bioinformatics in the mid-1990s for analyzing protein and DNA sequences, and have been widely used ever since (Chen and Rost, 2002; Krogh et al., 1994; Eddy, 1995). In 2006, Friedrich et al. (2006) introduced a modified version of the profile HMM, called interaction profile hidden Markov model (ipHMM), which expands the model to include differentiating which residues are part of the protein's interaction with other molecules. The three state types of the pHMM, representing, for each amino acid, whether it is an insertion, deletion, or match when aligned to the protein family, remain in the ipHMM. However, as shown in Fig. 1, the match state type is separated into two distinct states: a non-interacting (M_{ni}) and an interacting match state (M_i). This allows the model to capture statistical differences between conserved amino acids in the protein family that are part of the interaction, and those that are conserved but not part of the interaction.

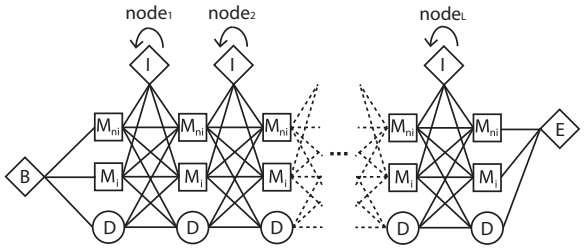

Fig. 1. Architecture of the interaction profile hidden Markov model. The match states of the classical pHMM are split into non-interacting (M_{ni}) and interacting (M_i) match states.

To be useful, an ipHMM has to be first trained, i.e., the parameters $e_i(x)$ and a_{ij} need to be estimated from protein sequences known to be members of a domain family. This is achieved by using a technique called maximum likelihood estimation and is based on a multiple sequence alignment of the member proteins in the domain family, incorporating the annotation of their interaction sites based on the X-ray crystallographic structure of the protein complexes. All residue positions are labeled to indicate whether they are interacting or non-interacting, which can then be used while training the model to capture relevant statistical information to separate interacting positions from those that are non-interacting. However, there are long-distance correlations between these interacting positions that we believe are not sufficiently captured by this model, as we will show in the following subsections.

2.2. Long-distance correlation

A previous study on the presence of long-distance correlations between amino acids in a protein found that the mutual information between any two amino acids decayed quickly as the sequence distance between them increased (Crooks and Brenner, 2004). Even when conditioned on the secondary structure type the pair was in, such as alpha helix or beta sheet, only a distance of about 4 was required for any correlation to become insignificant. However if the correlation is conditioned on pairs of amino acids at key secondary structure positions, such as the boundaries of alpha helices and beta sheets, a significantly stronger correlation emerges (Gonzalez, 2009).

Because the interacting domains of proteins undergo strong selection pressure during evolution, it is reasonable to hypothesize that similarly to the pairs of amino acids on the boundaries of secondary structure domains, stronger correlations may exist between pairs of residues that are interacting compared to those that do not interact (Pazos et al., 1997; Gonzalez et al., 2011). In other words, the amino acids that occur at interacting residues may not be entirely independent of the amino acids occurring at other interacting residues in the same protein. We have previously defined $S(x, y)$ to measure the correlations that exist between any