Methods 110 (2016) 97–105

Contents lists available at ScienceDirect

Methods

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

Prediction of residue-residue contact matrix for protein-protein interaction with Fisher score features and deep learning

Tianchuan Du [a,*], Li Liao [a,*], Cathy H. Wu [a,b], Bilin Sun [a]

[a] Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA
[b] Center for Bioinformatics & Computational Biology, University of Delaware, Newark, DE, USA

ARTICLE INFO

Article history:
Received 13 April 2016
Accepted 3 June 2016
Available online 6 June 2016

Keywords:
Deep learning
Stacked autoencoders
Deep neural networks
Protein-protein interaction
Contact matrix
Machine learning

ABSTRACT

Protein-protein interactions play essential roles in many biological processes. Acquiring knowledge of the residue-residue contact information of two interacting proteins is not only helpful in annotating functions for proteins, but also critical for structure-based drug design. The prediction of the protein residue-residue contact matrix of the interfacial regions is challenging. In this work, we introduced deep learning techniques (specifically, stacked autoencoders) to build deep neural network models to tackled the residue-residue contact prediction problem. In tandem with interaction profile Hidden Markov Models, which was used first to extract Fisher score features from protein sequences, stacked autoencoders were deployed to extract and learn hidden abstract features. The deep learning model showed significant improvement over the traditional machine learning model, Support Vector Machines (SVM), with the overall accuracy increased by 15% from 65.40% to 80.82%. We showed that the stacked autoencoders could extract novel features, which can be utilized by deep neural networks and other classifiers to enhance learning, out of the Fisher score features. It is further shown that deep neural networks have significant advantages over SVM in making use of the newly extracted features.

1. Introduction

Protein-protein interactions (PPIs) play essential roles in many biological processes. The cost, time and other limitations associated with the current experimental methods to detect protein-protein interaction have motivated the development of computational methods for predicting PPIs [1–3]. Acquiring knowledge of the interfacial regions between interacting proteins is not only helpful in annotating functions for proteins, but also critical for structure-based drug design and disease treatment [4–6]. Despite a lot of effort and progresses that have been made in PPIs predictions, most computational methods can only predict whether two proteins interact or not, but could not tell which residues on these two proteins are actually in contact, although such information obviously can be valuable for further understanding the interaction mechanisms and hence for designing modulation of the interaction via mutagenesis. Furthermore, even some methods can predict residue contact information, many of them require protein 3D structures which are not easily available, for example, docking [7]. Thus, it is desirable that computational method can predict the detailed residue-residue contact information from pure protein sequences.

The interaction of a sequence pair can be viewed as a contact matrix with rows and columns corresponding to the residues in the two interacting sequences respectively, and the element in the matrix indicates whether the corresponding pair of residues interact or not [8]. The contact matrix is a binary matrix, in which 1 means the two corresponding residues are in contact and 0 means the two corresponding residues are not in contact. For example, the element of a value 1 at (row 1, column 2) means that the first residue in sequence A (A_1) is in contact with the second residue in Sequence B (B_2) in Fig. 1. Like many prediction problems, one key step in residue-residue contact matrix prediction is to extract useful, predictive features from protein sequence. A Support Vector Machines (SVMs)-based method has shown some progress in contact matrix prediction [8]. In the research, Fisher score features of proteins, extracted from interaction profile Hidden Markov Models (ipHMMs), were used with SVMs to predict the protein residue-residue contact. Following is a brief explanation of the method. As the first step, protein domains of proteins are identified and profiled using ipHMMs. The ipHMM architecture takes account both structural information and sequence data. Each

* Corresponding authors at: Department of Computer and Information Sciences, 101 Smith Hall, Newark, DE 19716, USA.
E-mail addresses: tdu@udel.edu (T. Du), lliao@cis.udel.edu (L. Liao), wuc@dbi.udel.edu (C.H. Wu), sunbilin@udel.edu (B. Sun).

http://dx.doi.org/10.1016/j.ymeth.2016.06.001
1046-2023/

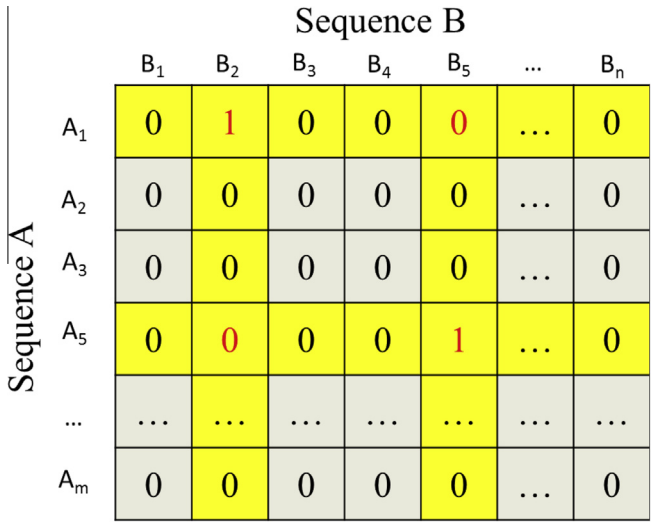

Fig. 1. Contact matrix example of an interacting sequence pair. One means two corresponding residues are in contact and zero means two corresponding residues are not in contact. Each row/column represents a residue in sequence A/B. The yellow rows/columns indicate those residues are interface residues (residues in contact with any residue from the other sequence).

protein can be characterized by either a distinct domain or a combination of domains. The evolutionarily conserved domains are well defined by the Pfam database [9,10]. The interacting properties protein domains can be profiled by ipHMMs. The construction of ipHMM was based on the ordinary profile hidden Markov model [11] by adding to the model architecture new states explicitly representing residues on the interacting interface [11,12]. Here, a match state of the classical pHMM is split into a non-interacting (M_{ni}) and an interacting match state (M_i) as shown in Fig. 2. More details for building the ipHMMs can be found in the methods section. The ipHMMs can be applied to predict interacting residues for individual protein sequence directly to predict interacting sites [12,13], however, it cannot tell how the interacting residues are paired up like the contact matrix. Thus as the next step, each residue for a member domain sequence is represented as a 20-dimensional vector of Fisher score features derived from the ipHMM such that the feature vectors of two residues can be fed into machine learning models for classification. The Fisher score vectors characterize how similar a residue in the protein sequence is compared with the family profile at that position, which have different pattern between interacting and non-interacting residues. The use of Fisher vectors to represent protein sequences was first proposed by Jaakkola (1999) in the context of detection of remote protein homologues and was later adopted for other applications in bioinformatics [14–16]. Feature score features can characterize each residue in the sequence in a way that captures how it contributes to the alignment of the sequence with the whole family as an ipHMM, and it has shown the ability to discriminate protein interactions [8,17]. At last, the features were used to training SVM models for classification. While achieving good overall performance, the previous method does not perform as well in differentiating true contact points from false positives when they are interface residues (the yellow rows/columns as shown in Fig. 1). Especially, it may not be the best way to use Fisher scores features to a machine learning classifier directly since residue-residue interactions are complicated processes, and there might be hidden features that could better represent the Fisher score features. Thus, we introduced deep learning techniques, specifically stacked autoencoders, to learn abstract features out of Fisher score features to predict residue-residue contact matrix.

Deep learning is a set of machine learning algorithms which attempt to learn multiple-layered models of inputs. The deep neural networks are composed of multiple levels of non-linear operations [18–21]. A central idea [22] of deep learning is referred to as greedy layerwise unsupervised pre-training, which is to learn a hierarchy of features one level at a time. The greedy layerwise unsupervised pre-training [19,23,24] is based on training each layer with an unsupervised learning algorithm, taking the features produced at the previous level as input for the next level. It is then straightforward to supply the extracted features either as input to a standard supervised machine learning classifier (such as SVMs or Random Forests) or as initialization for a deep supervised neural network (DNN). Stacked autoencoders are a typical class of those deep learning algorithms [25,26]. An autoencoder neural network is an unsupervised learning algorithm that applies backpropagation, setting the target values to be equal to the inputs [26,27]. An autoencoder takes an input x in $[0,1]^d$ and first maps it (with an encoder) to a hidden representation h through a deterministic mapping: $h = f(Wx + b)$, where f is a non-linear function, such as the sigmoid function, $f(z) = 1/(1 + \exp(-z))$ or rectified linear unit (ReLU), $f(z) = max(0, z)$. The sigmoid function is used in this paper. The latent representation, h, is then mapped back (with a decoder) into a reconstruction z of the same shape as x. The mapping happens through a similar transformation, e.g.: $z = f(W'h + b')$. The parameters of this model are optimized such that the average

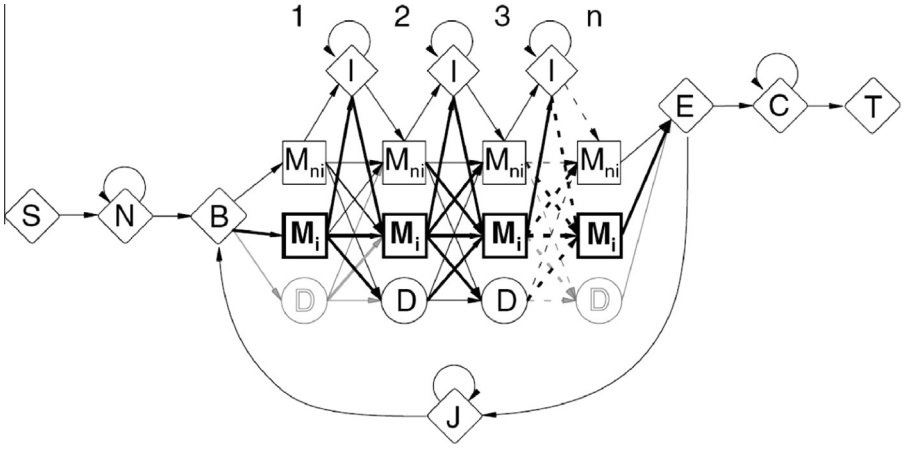

Fig. 2. Architecture of the interaction profile hidden Markov model. The ipHMM architecture follows the restrictions and connectivity of the HMM architecture. The match states of the classical HMM are split into non-interacting (Mni) and interacting (Mi) match states. Image credit for Friedrich et al. Bioinformatics, 2006.