Available online at www.sciencedirect.com
ScienceDirect
journal homepage: www.elsevier.com/locate/bbe

Original Research Article

Virus–human protein–protein interaction prediction using Bayesian matrix factorization and projection techniques

Q1 Esmaeil Nourani [a,*], Farshad Khunjush [b,c], F.E. Sevilgen [d]

[a] Department of Information Technology, Faculty of Information Technology and Computer Engineering, Azarbaijan Shahid Madani University, Tabriz, Iran
[b] Department of Computer Science and Engineering, School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran
[c] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
[d] Department of Computer Engineering, Gebze Technical University, Kocaeli, Turkey

ARTICLE INFO

Article history:
Received 22 October 2017
Received in revised form
31 March 2018
Accepted 23 April 2018
Available online xxx

Keywords:
Bioinformatics
Protein–protein interaction
Pathogen host interaction
Interaction prediction
Kernelized projection

ABSTRACT

Pathogens infect host organisms by exploiting host cellular mechanisms and evading host defence mechanisms through molecular pathogen–host interactions (PHIs). Discovering new interactions between pathogen and human proteins is very crucial in understanding the infection mechanisms. By analysing interaction networks, the interactions responsible for infectious diseases can be detected and new drugs disabling these interactions can be delivered. In this paper, we propose a method based on Bayesian matrix factorization for predicting PHIs along with a projection-based technique and combine the results by employing an ensemble method. Furthermore, two features, *target similarity* and *attacker similarity*, are utilized for the first time in the literature for PHI prediction. The advantages of the proposed methods are two folds. Firstly, they relieve the need for negative samples which is significant since there is no available dataset providing negative samples for most of the pathogenic systems. Secondly, the experiments demonstrate that the proposed approach outperforms state-of-the-art methods; roughly 20% of top 50 predictions are among recently validated interactions. So, the search space for wet-lab experiments to obtain validated interactions can be considerably narrowed down from a huge number of possible interactions.

* *Corresponding author at*: Department of Information Technology, Faculty of Information Technology and Computer Engineering, Azarbaijan Shahid Madani University, Kilometere 35, Tabriz/Azarshahr Road, Tabriz, Iran.
E-mail address: ac.nourani@azaruniv.ac.ir (E. Nourani).

https://doi.org/10.1016/j.bbe.2018.04.006

Please cite this article in press as: Nourani E, et al. Virus–human protein–protein interaction prediction using Bayesian matrix factorization and projection techniques. Biocybern Biomed Eng (2018), https://doi.org/10.1016/j.bbe.2018.04.006

1. Introduction

One of the major threats to human health is still infectious diseases, despite of continuous research effort for a long time. This is mostly due to the fact that novel variants of pathogens pose the problem of drug resistance. Novel antimicrobial therapeutics to fight drug resistance is only possible thorough exploration of pathogen–host interaction (PHI) systems [1]. However, current PHI knowledge is limited due to the time-consuming and expensive experimental methods for validating PHIs. Therefore, PHI prediction is worthwhile to enlighten the infection mechanisms in the case of limited experimentally verified PHI data. Moreover, predicted PHIs are excellent candidates for experimental verification.

Intra-species protein–protein interaction (PPI) prediction (within a single organism) is considered by researchers for more than a decade [2], while the prediction of PPIs between different organisms (inter-species PPI prediction) recently has attracted more attention. Pathogen–host interaction (PHI) refers to an interaction encountered between proteins of a pathogen and a host organism.

Computational methods for PHI prediction mostly focused on human as the host (and rarely for plants) [3] and primarily utilize sequence information [4,5] and protein structure [6–10] along with known interactions. Some of the computational approaches for analysing PHI network and predicting new PHIs are based on methods originally proposed for studying the social networks. There are many similarities between problems in the context of social networks and the biological networks in bioinformatics. For instance, in the movie rating problem, users rate the movies, and movies are recommended to users based on movie ratings, movie properties, previous ratings of the users and similarities between users and movies [11]. Matrix factorization is a well-known method for recommendation systems where rate matrix is factorized to obtain latent feature matrices of users and movies. Movies are recommended to users based on the extracted latent features. This approach can be applied to similar problems in bioinformatics. For example, in a recent study, kernelized Bayesian matrix factorization (KBMF) [12] is used for predicting drug-target interactions. KBMF extends kernelized matrix factorization with a Bayesian treatment and with an ability to utilize multiple side information sources expressed as different kernels [13].

In this paper, we adapt KBMF for PHI prediction, where interaction matrix is reconstructed to approximate unknown values. Binary interaction matrix between pathogen and host proteins is factorized to extract latent features of proteins. New interaction can be predicted by employing latent features of proteins. The adapted KBMF for PHI prediction produces promising results in comparison with state-of-the-art methods. Furthermore, we present an ensemble method to combine methods and similarity kernels. First, we show that for a single feature, the ensemble of methods improves the results of the individual method. This ensemble combines two projection-based approaches called KBMF and MKPE (multiple kernel preserving embedding) [14]. Then, we try to combine various similarity features using feature ensemble, and we show that combining outputs of KBMF for different kernels using the ensemble method, considerably improves the prediction results. We compare our approach with different state-of-the-art methods for PHI prediction [15–17]. The results of our extensive experiments verify the outperformance and the generalizability of our method for different PHI systems.

2. Materials and methods

2.1. System overview

In this section, we illustrate an overview of the proposed approach for PHI prediction. The idea behind the method is explained and the overall procedure is visualized.

Considering matrix factorization for PHI predictions, latent features of proteins can be extracted by factorizing the binary interaction matrix between pathogen proteins and host proteins. These features can be any characteristics of proteins, which are extracted based on their interaction behaviors. In an analogy by considering latent features as coordinates in a multi-dimensional space, the coordinate distance of pathogen and human proteins can be used to determine whether the proteins are interacting or not. However, to have this ability, coordinates of the pathogen and human proteins should belong to the same space. The projection operation plays the role of mapping coordinates of human and pathogen proteins in the unified target space, where interactions between them can be discovered based on their proximity. For pathogen and human proteins, the projected coordinates in the target space are represented by K-dimensional feature vectors called G_v and G_h, respectively:

$$G_v = \left\{ g_{v,i} \in \mathbb{R}^k \right\}_{i=1}^{N_v}, \quad G_h = \left\{ g_{h,i} \in \mathbb{R}^k \right\}_{i=1}^{N_h} \tag{1}$$

In Eq. (1), N_v and N_h are the numbers of viral and human proteins and $g_{h,i}$ ($g_{v,i}$) is the projected feature vector of ith human (pathogen) protein. Fig. 1 illustrates each protein in a 2D space for visual clarity; however, K-dimensional feature space is considered.

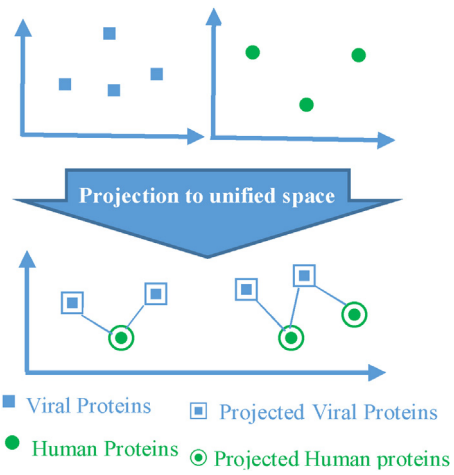

Fig. 1 – Projecting viral and human proteins to a unified space to discover the interactions.

Please cite this article in press as: Nourani E, et al. Virus–human protein–protein interaction prediction using Bayesian matrix factorization and projection techniques. Biocybern Biomed Eng (2018), https://doi.org/10.1016/j.bbe.2018.04.006