Neurocomputing 206 (2016) 81–88

Contents lists available at ScienceDirect

Neurocomputing

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

Relevance search for predicting lncRNA–protein interactions based on heterogeneous network

Jianghong Yang [a], Ao Li [a,b,*], Mengqu Ge [a], Minghui Wang [a,b]

[a] School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China
[b] Centers for Biomedical Engineering, University of Science and Technology of China, Hefei 230027, China

ARTICLE INFO

Article history:
Received 12 August 2015
Received in revised form
30 October 2015
Accepted 20 November 2015
Available online 31 May 2016

Keywords:
lncRNA–protein interaction
Protein–protein interaction information
Heterogeneous network
Relevance search

ABSTRACT

lncRNA plays important roles in many biological and pathological processes. lncRNA–protein interaction is the most common way of lncRNA performing their functions. Thus, predicting lncRNA–protein interaction is very significant to understand the nature of lncRNA. Earlier methods to predict RNA–protein interaction always adopt classification algorithms using features extracted from RNA and protein themselves. But their performance is not good enough in lncRNA–protein interaction prediction because of lncRNA's low conservation. In this paper, we try to use information implicit in the topologies of biological network associated with lncRNA to solve this problem. Firstly we construct a heterogeneous lncRNA–protein network which incorporate protein interaction information. We use an algorithm called HeteSim which can evaluate relevance between heterogeneous objects to perform relevance search in the lncRNA and protein heterogeneous network. The relevance search results are used to help the identification of lncRNA–protein interactions.

1. Introduction

The operational definition of long noncoding RNAs (lncRNAs) are transcripts of RNA polymerase II, which are longer than 200 nucleotides and lack of open reading frame that is able to be translated into proteins [1]. Huge amount of research on lncRNAs' function shows us that lncRNAs display important roles in many biological and pathological processes including regulating gene expression in all level, post-translational regulation for proteins activity, communication between cells, organization of protein complexes, as well as recombination process [2]. In general, lncRNAs exert functions by interfacing with corresponding RNA-binding proteins. Thus, identifying lncRNA interacted proteins is significant to understand lncRNAs' intricate functions and molecular mechanism in disease progression and cellular circuitry.

In fact, several computational approaches have been developed to predict interaction between RNA and protein (Table 1). The pioneer work can be traced back to 2011. A computational method to predict ncRNA–protein interaction was proposed by Pancaldi et al. [3]. They used random forest (RF) and support vector machine (SVM) classifiers with more than 10 features extracted from RNA and protein physical properties, protein secondary structure, RNA localization and untranslated regions. Then, a method called RPISeq was introduced by Muppirala et al. [4] in the same year. They also used RF and SVM classifiers but with the features only got from protein and RNA sequences. Bellucci et al. developed an approach named catRAPID [5] by exploiting physicochemical properties of RNA and protein including secondary structure, van der Waals forces and hydrogen bond to analyze and predict the RNA–protein interactions and domains. In 2013, Wang et al. [6] presented a method depends on Naive Bayes and extended Naive Bayes classifiers utilizing similar triad features to those of Muppirala et al.'s work. Lncpro tool was developed by Lu et al. [7] using Van der Waals propensities, hydrogen bonding and secondary structure. In 2015, Suresh et al. introduced their method called RPI-Pred [8] which used the traditional SVM classifier but with both sequences and predicted structures peculiarities of protein and RNA.

However, most of these methods use classification algorithms based on the properties of RNA and protein themselves. However, when it comes to lncRNA, most of them are not effectual enough for ignoring specialty of lncRNA. Even through a large number of small RNAs, such as snoRNAs or microRNAs, exhibit high conservation across diverse species [9], large amounts of research have illustrated the fact that lncRNAs generally exhibit low conservation [10]. This phenomenon bring difficulties to use classifier based on intrinsic properties of lncRNAs and proteins. In our study,

* Corresponding author at: School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China.
E-mail address: aoli@ustc.edu.cn (A. Li).

http://dx.doi.org/10.1016/j.neucom.2015.11.109
0925-2312/

Table 1
Computational approaches to predict interaction between RNA and protein.

Method	Interaction	Model	Feature (information)
Pancaldi et al.	mRNA–protein	RF and SVM classifier	Protein secondary structure, RNA localization untranslated regions
RPISeq	RNA–protein	RF and SVM classifier	Only features from protein and RNA sequences
catRAPID	ncRNA–protein	Constructed interaction matrix	Secondary structure, van der Waals forces, hydrogen bond
Wang et al.	RNA–protein	Naïve Bayes and extended Naive Bayes classifiers	Features from protein and RNA sequences
lncPro	lncRNA–protein	Development of catRAPID	Van der Waals, hydrogen bonding and secondary structure
RPI-Pred	ncRNA–protein	SVM classifier	Sequences feature, predicted Structures peculiarities of protein and RNA

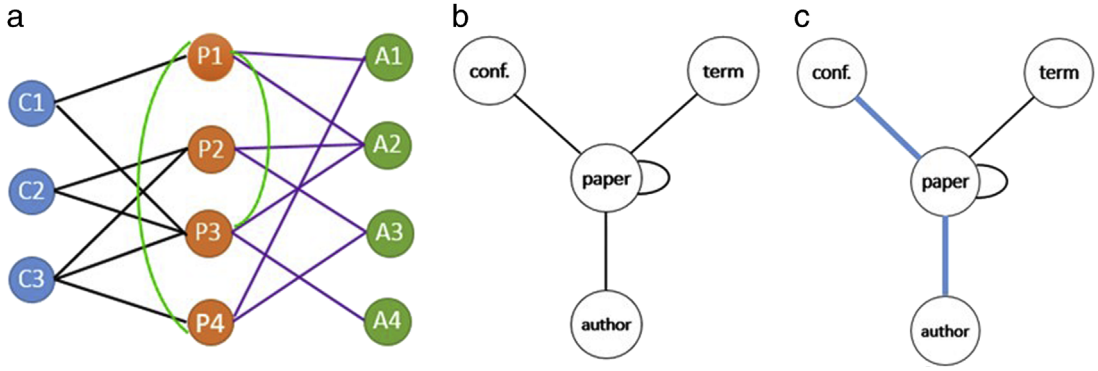

Fig. 1. Heterogeneous information network: (a) bibliographic information network's schema; (b) bibliographic information network; and (c) bold line represent the relevance path. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

we try to solve this problem from a completely new point, the biological network.

Nowadays, network science has extensively being used in many fields including biology related fields. A biological network means network that be applied to biological systems. Network science provides a quantifiable description for the networks that characterize various biological systems [11] and relationships between disease and biological factors [12,13], and has made rapid advances. In our study, we represent and analyze interactions among lncRNAs and proteins as a heterogeneous network (heterogeneous lncRNA–protein network). Then, to predict interactions between lncRNA and protein we adopt a relevance search algorithm called HeteSim [14]. Our choice for HeteSim is motivated by the following two facts: Fist, it is a symmetric, path-constrained relevance measure method which can evaluate relevance between objects of different types; Second, this method performs efficiently on evaluating relatedness of heterogeneous objects, such as Li et al.'s work [15]. They have used HetsSim in drug-target interaction mining and got excellent results.

2. Materials and methods

To avoid difficulties caused by lncRNAs' low conservation, the problem of predicting interaction between lncRNAs and proteins is interpreted as a relevance search task in heterogeneous lncRNA–protein network. We adopt the path-constrained algorithm relevance measurement called HeteSim to perform lncRNA–protein interaction prediction.

The definition of heterogeneous network is a directed graph $G=(V,E)$ accompanied by a link type mapping function $\varphi: E\to R$ and an object type mapping function $\tau: V\to A$. Each object $v\in V$belongs to a specific object type $\tau(v)\in A$, and each link $e\in E$ belongs to one particular relation type $\varphi(e)\in R$. The network is referred to as heterogeneous network if either objects types or relation types are multiple [16]. And network schema, $TG=(A,R)$, is defined as a meta template of a heterogeneous network $G=(V,E)$, which is a directed graph defined on object types A, with edges as relations from R, with object type mapping $\tau: V\to A$and link mapping $\varphi: E\to R$. Many interconnected, large scale data sets such as social, biology, medicine even business can construct corresponding heterogeneous information network, for instance, the bibliographic information network (Fig. 1a, b) [16]. What's more, the relevance path is defined over the network schema which represents relationship among object types, as the Fig. 1**c** shown.

2.1. Algorithm

To measure the relatedness between objects with different types in heterogeneous network, Shi et al. [14] have proposed a novel measurement called HeteSim. HeteSim is a path-based relevance measurement using pair-wise random walk (RW) model. But HeteSim cannot match requirements of our study perfectly, so we make a little adjustment.

In a given relevance path $R=R1\circ\cdots\circ Rl$ $(A_1\Longrightarrow^{R_1}A_2\Longrightarrow^{R_2}\cdots\Longrightarrow^{R_1}A_{l+1}$, A_i means the i-th node type), HeteSim measures the similarity between two objects x and y ($x\in R_1\cdot A_1$ and $y\in R_l\cdot A_{l+1}$) according to the relevance score:

$$Hetesim(x,y|R1\circ R2\circ\cdots\circ Rl-1\circ Rl)=\frac{1}{|O(x|R1)||I(y|Rl)|}\sum_{i=1}^{|O(x|R_1)|}\sum_{j=1}^{|I(y|R_l)|}Hetesim(O_i(x|R_1),I_j(y|R_l)|R_1\circ R_2\circ\cdots\circ R_{l-1}) \quad (1)$$

$O(x|R1)$ means the out-neighbors of x based on relation$R1$ along the path (i means the i-th out-neighbor), and $I(y|Rl)$ presents the in-neighbors of y based on relationRl against the path (j means the j-th in-neighbor) [14]. The relevance between x and y is