Journal of Theoretical Biology 437 (2018) 149–162

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

S-FLN: A sequence-based hierarchical approach for functional linkage network construction

A. Jalilvand [a], B. Akbari [a,*], F. Zare Mirakabad [b]

[a] Department of Electronic and computer engineering,Tarbiat Modares University, Tehran, Iran
[b] Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran

ARTICLE INFO

Article history:
Received 31 December 2016
Revised 27 July 2017
Accepted 18 October 2017
Available online 26 October 2017

Keywords:
Network modeling
Network construction
Ensemble learning
Functional linkage network (FLN)
Link prediction

ABSTRACT

The functional linkage network (FLN) construction is a primary and important step in drug discovery and disease gene prioritization methods. In order to construct FLN, several methods have been introduced based on integration of various biological data. Although, there are impressive ideas behind these methods, they suffer from low quality of the biological data. In this paper, a hierarchical sequence-based approach is proposed to construct FLN. The proposed approach, denoted as S-FLN (Sequence-based Functional Linkage Network), uses the sequence of proteins as the primary data in three main steps. Firstly, the physicochemical properties of amino-acids are employed to describe the functionality of proteins. As the sequence of proteins is a more comprehensive and accurate primary data, more reliable relations are achieved. Secondly, seven different descriptor methods are used to extract feature vectors from the proteins sequences. Advantage of different descriptor methods lead to obtain diverse ensemble learners in the next step. Finally, a two-layer ensemble learning structure is proposed to calculated the score of protein pairs. The proposed approach has been evaluated using two biological datasets, *S.Cerevisiae* and *H.Pylori*, and resulted in 93.9% and 91.15% precision rates, respectively. The results of various experiments indicate the efficiency and validity of the proposed approach.

1. Introduction

Recent researches show that perturbations of cellular systems are the main cause of human diseases, especially in molecular networks (Barabási et al., 2011; Goh et al., 2007). Meanwhile, the associated genes to the same or similar diseases commonly reside in the same neighborhoods of molecular networks (Goh et al., 2007). These observations have been found as the basis of many computational methods to associate the unknown genes to certain diseases, Considering that the machine learning methods play an important role in system biology approaches (Hedberg, 2006; Kell, 2006). The majority of these methods are based on functional linkage networks (FLNs) (Wang et al., 2011). FLNs are well-defined data structures which are used to identify disease-related genes. They are extendable to investigate gene cooperation in complex diseases and drug discovery (Apolloni et al., 2011). Moreover, FLNs can be used to assign function classes to unknown genes or proteins, which has been known as a fundamental task in biological researches (Apolloni et al., 2011; Manimaran et al., 2009).

An FLN is defined as a graph in which the nodes represent genes or corresponding proteins and the edges denote functional associations between them. In other words, two proteins are connected in an FLN if some experimental or computational methods indicate that they share the same functionality. In this regard, the process of identifying functional relationships among the proteins is called FLN construction. Fig. 1 shows an overview of the FLN construction process.

Many computational methods have been proposed in the literature to predict the links between the proteins and construct the biological networks. Generally, they can be divided into two categories. In the first category, the main idea is to integrate different data sources to construct the biological network. The data sources include Protein-Protein Interaction (PPI) network, gene fusion, gene neighborhoods, literature mining knowledge, Gene Ontology(GO), and other data (Franke et al., 2006; Lei et al., 2012; Linghu et al., 2009; Wang et al., 2014; Wu et al., 2010; You et al., 2010). In Franke et al. (2006) an FLN is constructed by integrating various types of biological data. The authors employed PPI, microarray co-expression, and GO and applied a Bayesian approach to predict gene pairs that participate in the same GO biological process. Similarly, in Köhler et al. (2008) multiple biological data

* Corresponding author.
E-mail address: b.akbari@modares.ac.ir (B. Akbari).

https://doi.org/10.1016/j.jtbi.2017.10.021
0022-5193/

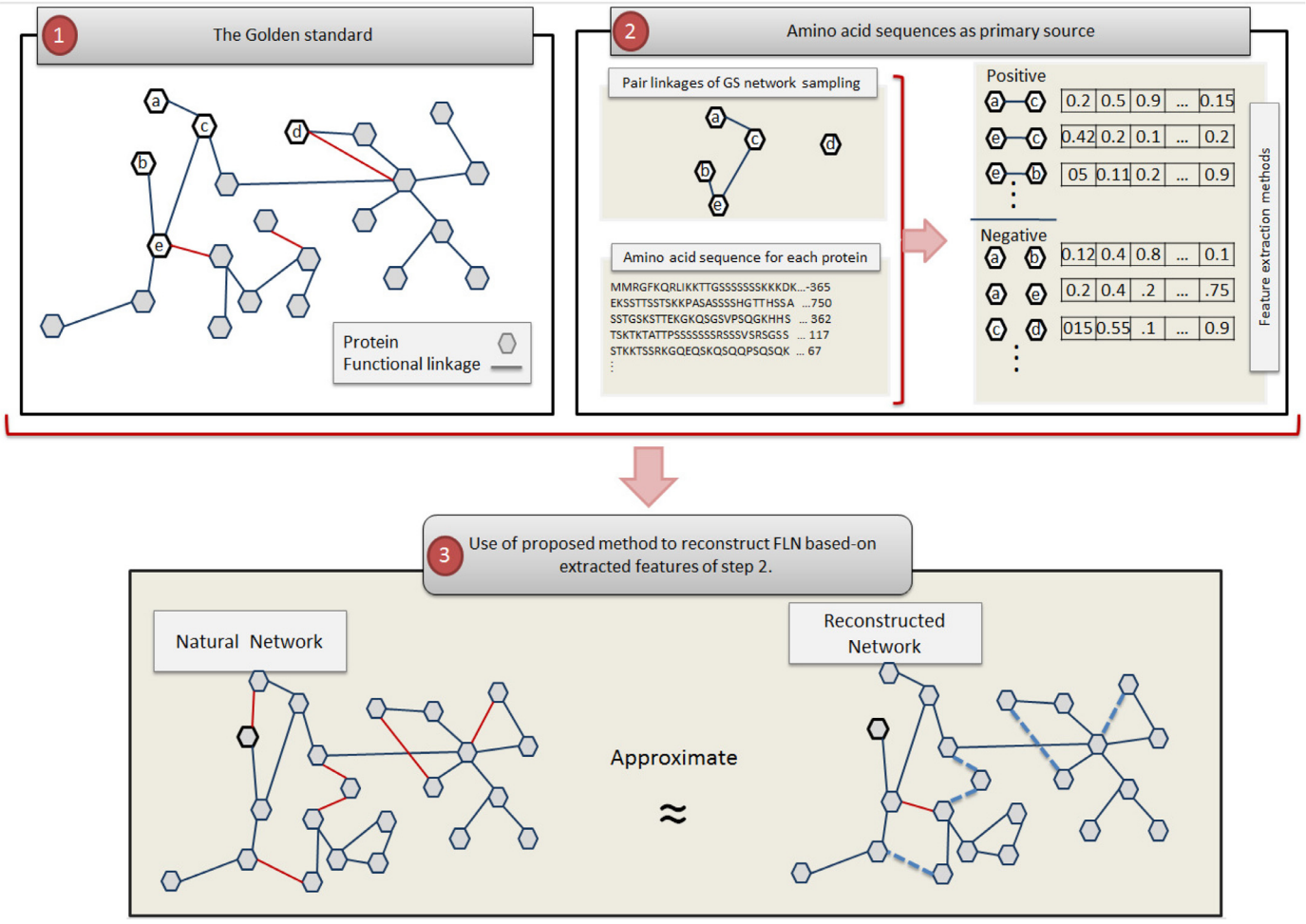

Fig. 1. Functional linkage network reconstruction: the Section 1 shows the gold standard (GS) network has been created by Gene Ontology. The Section 2 shows feature vector extraction step based-on amino acid sequences for positive-negative linkages of GS network. The Section 3 shows use of proposed method to approximate natural network based-on extracted features of protein pairs.

sources by random walk algorithm was used for disease gene prioritization.

In Linghu et al. (2008), a six-dimension feature vector based on the six biological data sources has been proposed. Then, multiple machine learning methods such as Support Vector Machine (SVM), Linear Discriminant Analysis(LDA), Naive Bayes, and Neural Network are applied to construct a reliable FLN. A human FLN by integrating 16 biological data features from 6 model organisms has been constructed by Linghu et al. (2009). Afterwards they use a Naive Bayes classifier to predict functional linkage between genes. Wang et al. (2014) built an FLN of mitochondrial proteins by integrating biological features such as genomic context, gene expression profiles, metabolic pathways and PPI network. A recent and interesting survey on functional linkage network construction methods is provided in Linghu et al. (2013).

Although the results of these methods might be acceptable in non-biological networks, due to low quality of biological data sources they are faced with major challenges in the biological networks. The reason is that the employed feature vectors include some missing values because most of the proteins are not appeared in different data sources. Missing values in integration based methods, causes a negative impact on the prediction accuracy of the protein pairs.

In the second category, a number of methods have been developed to derive information directly from amino acid sequences (Guo et al., 2008; Mei and Zhu, 2014; Shen et al., 2007; Xia et al., 2010b; You et al., 2013; 2014; Yousef and Charkari, 2013). These methods generally are divided into two classes: alignment-based and alignment-free methods. Although, the alignment-based methods obtain high accuracy for some of sequences, their results are inaccurate on the inversion, translocation at substring level, and diverse sequences with the same functionally or unequal lengths (Borozan et al., 2015; Li et al., 2016; Otu and Sayood, 2003).

In this regard, alignment-free methods have been proposed to overcome these issues (Aguiar-Pulido et al., 2012; Agüero-Chapin et al., 2009; Dea-Ayuela et al., 2008; Munteanu et al., 2008a, 2009; Perez-Bello et al., 2009; Vilar et al., 2009; Vinga, 2014). The alignment-free methods include two steps: (i) the protein sequences are transformed into fixed-length feature vectors; (ii) The feature vectors are employed as training set in machine learning algorithms (Fernandez-Lozano et al., 2014; González-Díaz and Riera-Fernández, 2012; Munteanu et al., 2008b; Yao et al., 2014).

A number of methods have been developed that use the sequence information of proteins to predict links in biological networks (Shen et al., 2007; Xia et al., 2010b; Yang et al., 2010; Yousef and Charkari, 2013; Zhang et al., 2011). Some of these methods use physicochemical properties of amino acids to enrich extracted feature vectors (Huang et al., 2016; Xia et al., 2010b; Yousef and Charkari, 2013). In Zhang et al. (2011) a computational approach based on compressed sensing theory is proposed to predict yeast PPI. They have used Auto Covariance (AC) method (Guo et al., 2008) and 7 physicochemical properties to extract the features. In