Journal of Biomedical Informatics 56 (2015) 300–306

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A novel method based on physicochemical properties of amino acids and one class classification algorithm for disease gene identification

Abdulaziz Yousef, Nasrollah Moghadam Charkari *

Faculty of Electrical & Computer Engineering, Tarbiat Modares University, Tehran, Iran

ARTICLE INFO

Article history:
Received 13 February 2015
Revised 4 June 2015
Accepted 26 June 2015
Available online 2 July 2015

Keywords:
Disease gene identification
Physicochemical properties of amino acid
One class classification
Support Vector Data Description

ABSTRACT

Identifying the genes that cause disease is one of the most challenging issues to establish the diagnosis and treatment quickly. Several interesting methods have been introduced for disease gene identification for a decade. In general, the main differences between these methods are the type of data used as a prior-knowledge, as well as machine learning (ML) methods used for identification. The disease gene identification task has been commonly viewed by ML methods as a binary classification problem (whether any gene is disease or not). However, the nature of the data (since there is no negative data available for training or leaners) creates a major problem which affect the results. In this paper, sequence-based, one class classification method is introduced to assign genes to disease class (yes, no). First, to generate feature vector, the sequences of proteins (genes) are initially transformed to numerical vector using physicochemical properties of amino acid. Second, as there is no definite approach to define non-disease genes (negative data); we have attempted to model solely disease genes (positive data) to make a prediction by employing Support Vector Data Description algorithm. The experimental results confirm the efficiency of the method with precision, recall and *F*-measure of 79.3%, 82.6% and 80.9%, respectively.

1. Introduction

Identifying disease genes is an important issue to enhance our knowledge about disease mechanism, and to improve clinical methods.

Traditional linkage and association studies have been carried out to identify a large number of candidate genes probably related with diseases [1]. Since using experimental approaches to identify disease associated genes from the vast numbers of candidates is an expensive task, the requirement of computational approaches has been taken into account. In this regard, many interesting machine learning methods have been introduced to find the similarity features between unknown genes (candidate genes) with the known disease genes. These methods differ in two ways. First, the type of genomic data used for generating the feature vector, such as protein–protein interactions (PPIs) [2–6], gene expression profiles [7], gene ontology (GO) [8]. Other methods integrate multiple data sources to prioritize candidate disease genes [9–11].

Second, the type of the algorithm which has been used for training the prediction model. However, the majority of the studies consider this issue as two class classification problem. Some studies have defined the known disease gene as positive set and the unknown disease gene as negative set [12,13]. Since the unknown genes set is often comprised with some disease genes, other methods have attempted to reduce the confusion in classification process by selecting a small fraction of unknown genes as negative set [14,15]. Nevertheless, these methods are not robust and reliable enough as the negative set which has been achieved from unknown genes, still suffers from noisy data.

All the above mentioned methods might not be implemented properly, because these methods rely on the information of proteins achieved from prior-knowledge (PPI network, gene ontology, and protein domains) which contains some errors. Moreover, they usually suffer from incompleteness. Therefore, a universal prior-knowledge would be required to tackle this problem. The only data which are available for all proteins and has influential role in solving many problems such as predicting a subcellular locations [16,17], protein–protein interactions [18–20], and protein structural and functional classes [21–23] is the sequences of proteins [24]. On the other hand, there is no information about the negative data (non-disease gene). Also, there is no guarantee of using the unknown genes or fraction of them as a negative data. Hence, using two class classification algorithm may be not appropriate.

* Corresponding author. Tel.: +98 21 82883301; fax: +98 21 82884325.
E-mail addresses: yousef@modares.ac.ir, Azizyousef1@yahoo.com (A. Yousef), charkari@modares.ac.ir (N.M. Charkari).

http://dx.doi.org/10.1016/j.jbi.2015.06.018
1532-0464/

In this paper, we present a novel sequence-based one class classification method for disease gene identification. Since the earliest protein sequences and the structures were determined, it would be clear that the positioning and properties of amino acids are key point to infer many biological processes. For example, the first protein structure, haemoglobin provides a molecular explanation for the genetic disease sickle cell anaemia [25]. Therefore, a sequence-translated method based on physicochemical properties of amino acid, is employed to construct a feature vector for each protein. To improve the performance of the proposed method, some efficient features are selected using Principal Component Analysis (PCA). Since, the mutation of genes is always possible to happen, and also there is a likelihood of this mutation leading to disease [26], there is no available conception asserts that some proteins are not involved in disease (non-disease genes). Hence, we have attempted to only train the positive data (disease genes) using Support Vector Data Description (SVDD) algorithm. We have used two type of data as negative data to evaluate the model. The fist type has selected randomly from unknown dataset. While the second one has been selected from the likely negative data used in positive-unlabeled technique [15].

The proposed method has been compared with ProDiGe method [14], Smalter's method [12], and yang method [15]. The experimental results achieved 79.3%, 82.6% and 80.9% for precision, recall and *F*-measure, respectively. These results indicate that our method overpassed the current state-of-the art methods in precision.

2. Material and method

In this section, the proposed method for identifying disease genes is described. The method consists of three steps: (1) Translate corresponding gene products (proteins) into numerical feature vector using physicochemical descriptor; (2) Principal Component Analysis (PCA) algorithm is applied to extract appropriate features; (3) training the positive data using SVDD algorithm to make the identification. The proposed method schema is illustrated in Fig. 1.

2.1. Protein sequence translation

Many representation methods have been introduced to extract the information encoded in amino acid sequence, including Geary auto correlation (GA) [27], auto covariance (AC) [28], Normalized Moreau–Broto autocorrelation (NA) [29], and Moran auto-correlation (MA) [30]. These methods are based on different physicochemical properties. In this paper, six joint physicochemical properties of amino acid which were used in many applications (e.g. predicting protein structural, functional classes, protein–protein interactions, sub-cellular locations...) have been selected. Since by adding one more physicochemical property, thirty features will be added to the feature vector, we have selected out the more effective physicochemical properties with the minimum number of these properties to avoid complexity (Time, and Computation). These properties are polarity (POL) [31], residue accessible surface area in tripeptide (RAS) [32], hydrophilicity (HY-PHIL) [33], polarizability (POL2) [34], hydrophobicity (HY-PHOB) [35], solvation free energy (SFE) [36], respectively. The original values of these physicochemical properties for each amino acid were normalized using Min–Max normalization method as shown in Eq. (1):

$$P_{ij} = \frac{P_{i,j} - P_{j,min}}{P_{j,Max} - P_{j,min}} \quad (1)$$

where $P_{i,j}$ is the j-th descriptor value for i-th amino acid, $P_{j,min}$ is the minimum value of j-th descriptor over the 20 amino acids and $P_{j,Max}$ is the maximum value of j-th descriptor over the 20 amino acids. Table 1 shows the normalized physicochemical properties. Since GA method has achieved a good result in other application [20], in this work, we have applied GA method as a representation method.

2.2. Principal Component Analysis (PCA)

To increase the overall performance of the proposed method, we tried to extract the most relevant and useful features from the high-dimensional represented feature vectors (180 features) generated by GA method. PCA is presented as a dimensionality reduction methods. PCA is an appropriate statistical technique to identify patterns in data, and to expiree the data in a way that to high light their similarities and differences. Therefore, utilized to determine significant features which preserves most of the information and removes the redundant components [37]. PCA is a linear combination which transforms one set of variables in R^m space into another set in R^n space containing the maximum amount of variance in the data where n is smaller than m. This is obtained as the following steps:

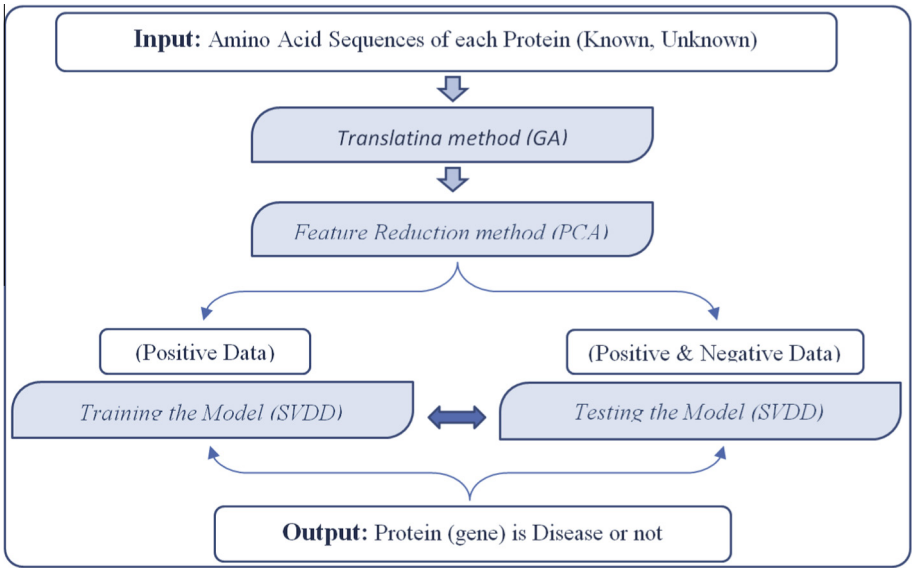

Fig. 1. The schematic of the proposed method. As is shown, the proposed method includes three layers which are representation layer, Feature extraction layer, predictor layer.