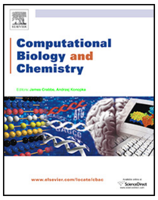
Contents lists available at ScienceDirect

Computational Biology and Chemistry

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



Research Article

Building and analysis of protein-protein interactions related to diabetes mellitus using support vector machine, biomedical text mining and network analysis

Renu Vyas[a,*], Sanket Bapat[b], Esha Jain[b], Muthukumarasamy Karthikeyan[b], Sanjeev Tambe[c], Bhaskar D. Kulkarni[c]

[a] MIT School of Bioengineering Science and Research, ADT University, Loni Kalbhor, Pune, 412201, India
[b] Digital Information Resource Centre (DIRC) & Centre of Excellence in Scientific Computing (CoESC), CSIR-National Chemical Laboratory, Pune, 411008, India
[c] Chemical Engineering and Process Development Division, CSIR-National Chemical Laboratory, Pune, 411008, India

ARTICLE INFO

Article history:
Received 16 July 2015
Received in revised form 7 September 2016
Accepted 19 September 2016
Available online 30 September 2016

Keywords:
Diabetes
SVM
Protein-protein interactions
Machine learning
Protein interaction network

ABSTRACT

In order to understand the molecular mechanism underlying any disease, knowledge about the interacting proteins in the disease pathway is essential. The number of revealed protein-protein interactions (PPI) is still very limited compared to the available protein sequences of different organisms. Experiment based high-throughput technologies though provide some data about these interactions, those are often fairly noisy. Computational techniques for predicting protein–protein interactions therefore assume significance. 1296 binary fingerprints that encode a combination of structural and geometric properties were developed using the crystallographic data of 15,000 protein complexes in the pdb server. In a case study, these fingerprints were created for proteins implicated in the Type 2 diabetes mellitus disease. The fingerprints were input into a SVM based model for discriminating disease proteins from non disease proteins yielding a classification accuracy of 78.2% (AUC value of 0.78) on an external data set composed of proteins retrieved via text mining of diabetes related literature. A PPI network was constructed and analysed to explore new disease targets. The integrated approach exemplified here has a potential for identifying disease related proteins, functional annotation and other proteomics studies.

1. Introduction

Diabetes is a chronic disease, which occurs when the pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces. This leads to an increased concentration of glucose in the blood; this condition is known as *hyperglycaemia*. Diabetes are of 3 types: Type- 1 diabetes, Type- 2 diabetes and gestational diabetes (Defronzo, 1997). According to World Health Organisation, about 346 million people worldwide have been affected with diabetes. Type 2 diabetes is afflicting the third world countries, which are fast becoming the epicentre of this silent killer, mainly due to the rampant urbanization, poor nutrition and sedentary life styles (Buchwald et al., 2009; G. Diabetes Prevention Program Research, 2002). Keeping in view the rapid growth of diabetes and also its critical impact worldwide, it is of major importance to devise global models that can identify novel proteins related to this disease.

Identification of protein-protein interactions (PPIs) is considered as a key strategy for understanding the various mechanisms of any disease (Kann, 2007). Diabetes is a multi-factorial disease wherein a host of crucial interactions are still largely unknown (Virkamaki et al., 1999). Protein interactions are known to correlate with the protein's functional properties and protein interaction networks are frequently utilized to discover the potential biological role of proteins with an unknown function. Both experimental and computational techniques have been used to identify protein-protein interactions (PPIs) in many organisms (Bader and Hogue, 2003). Experimental studies to identify protein-protein interactions (PPIs) have been carried out using techniques such as yeast two-hybrid screens and co-affinity purification followed by mass spectrometry (Phizicky and Fields, 1995). These methods, may be prone to error and may not be specific to proteins in all organisms. Moreover, there is a possibility of a number of false positives in the high throughput data from protein assays (Botstein et al., 2000). Thus use of computational methods for

[*] Corresponding author.
E-mail address: renu.vyas@mituniversity.edu.in (R. Vyas).

http://dx.doi.org/10.1016/j.compbiolchem.2016.09.011
1476-9271/© 2016 Elsevier Ltd. All rights reserved.

predicting proteins in PPI has been intensified. The identification of proteins responsible for human diseases is one of the most challenging tasks in the drug design. Some computational methods for example sequence based, high-throughput database and a combination of both have been used to predict protein-protein interactions (Karlin and Belshaw, 2012). Machine learning methods including Bayesian classifiers, probabilistic decision trees, logistic regression, support vector machines have been employed for predicting the PPIs by using a number of properties of proteins to classify the data (Pizzuti and Rombo, 2016; Oliva and Fernandez-Fuentes, 2016). However, generally the PPI networks are constructed on the basis of sequence data alone.

It is known that physically interacting proteins tend to be involved in the same cellular process, and mutations in their genes may lead to similar disease phenotypes (Estavez et al., 2009). Proteins must interact physically, at least briefly, to form temporary associations to express their biological functions in the cell (Ideker and Sharan, 2008). In the current work, we hypothesize that the probability of physical interaction between two proteins depends upon the 3D structural features derived from the 3D structure of a macromolecule. Based on this knowledge gained by studying 3D structures of 15,000 protein complexes deposited in the Brookhaven Protein Databank, we have generated 1296 binary fingerprint based descriptors encoding the geometric and structural attributes of a protein. We computed binary fingerprints for the proteins related to the type 2 diabetes disease. It is envisaged that the methodology employed here can be used efficiently to distinguish between disease related and non-disease related query proteins.

1.1. Machine learning

One of the main challenges in using the SVM for the prediction of PPIs from the protein sequences is finding a suitable transformation of the protein sequence information present in a fixed number of inputs to be used in SVM training. Many studies in the past have exploited the physiochemical properties of proteins to predict protein-protein interactions (Mei and Zhu, 2016). However, often unequal length inputs are considered in these studies because of the varying lengths of the protein sequences. Thus a method is proposed that converts a protein sequence into fixed-dimensional representative attributes, wherein each feature represents the relationship of amino acid to the protein sequence of interest. The approach is schematically illustrated in Fig. 1.

2. Materials and methods

An in-house developed Java based program was used to generate the binary fingerprints of bit length 1296 for a given protein structure. JProLine, a program developed in our research group was employed for constructing multiple sequence alignment and heatmap generation (Kumar et al., 2016). Another internally developed tool, MegaMiner portal was employed for the rapid intelligent text mining of biomedical records (Karthikeyan et al., 2015).

2.1. Data collection and data set construction

For building the training set, two classes of proteins namely disease related and non-disease related entries were selected. The disease related proteins were retrieved from the PDB and UniProt databases. A total of 1424 entries from the UniProt and 461 entries from the PDB were obtained. These proteins were combined to form the positive data set. For the negative data set equal number of proteins that were not related to the disease were collected from PDB. The final data set consisted of 2653 proteins; half of them were associated with the diabetes disease while the rest were not associated. The model was then applied on an external set of proteins (n = 129) that were extracted via text mining.

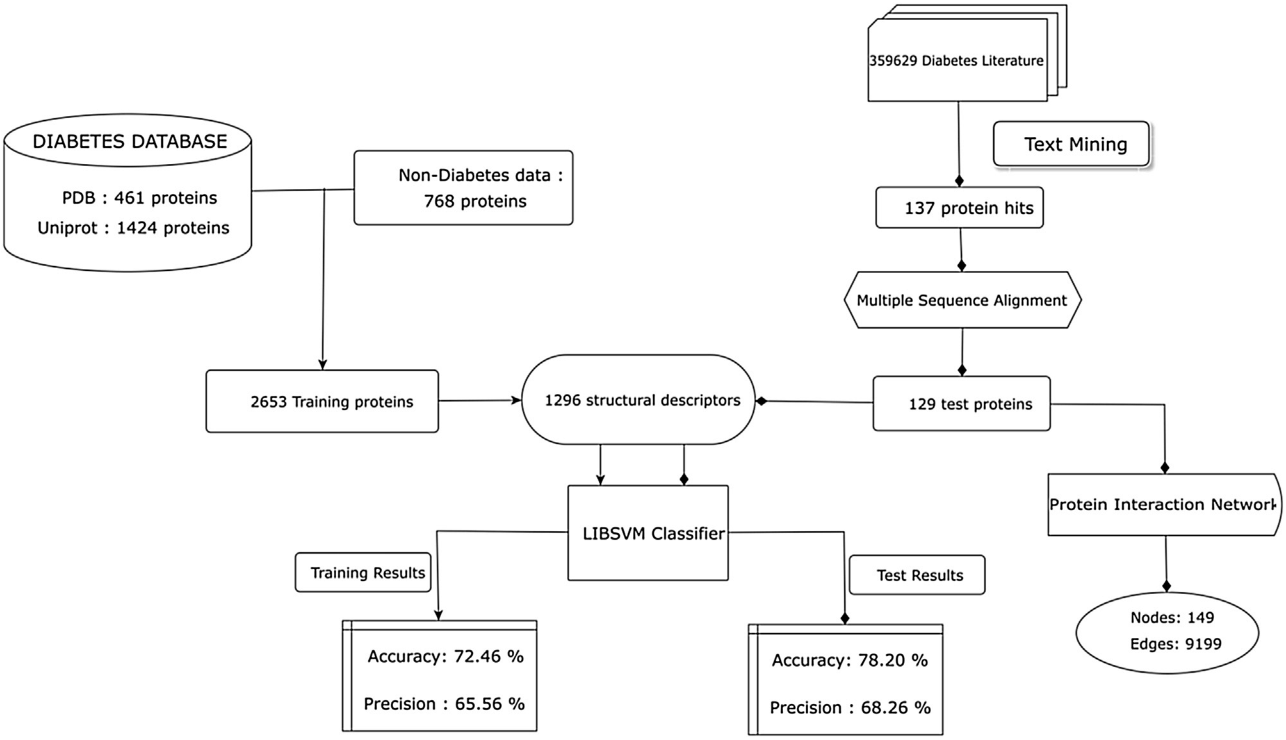


Fig. 1. A schematic representation of the overall workflow in the present work.