Journal of Theoretical Biology 359 (2014) 120–128

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A set of descriptors for identifying the protein–drug interaction in cellular networking

Loris Nanni [a,*], Alessandra Lumini [b], Sheryl Brahnam [c]

[a] DEI, University of Padua, viale Gradenigo 6, Padua, Italy
[b] DISI, Università di Bologna, Via Venezia 52, 47521 Cesena, Italy
[c] Computer Information Systems, Missouri State University, 901S. National, Springfield, MO 65804, USA

HIGHLIGHTS

- Protein–drug interactions.
- Ensemble of machine learning system.
- Matrix representation of a protein for extracting different descriptors.
- Position specific scoring matrix for describing a protein.

ARTICLE INFO

Article history:
Received 13 February 2014
Received in revised form
2 June 2014
Accepted 6 June 2014
Available online 17 June 2014

Keywords:
Drug–target interaction
Amino-acid sequence
Position specific scoring matrix
Machine learning
Ensemble of classifiers

ABSTRACT

The study of protein–drug interactions is a significant issue for drug development. Unfortunately, it is both expensive and time-consuming to perform physical experiments to determine whether a drug and a protein are interacting with each other. Some previous attempts to design an automated system to perform this task were based on the knowledge of the 3D structure of a protein, which is not always available in practice. With the availability of protein sequences generated in the post-genomic age, however, a sequence-based solution to deal with this problem is necessary. Following other works in this area, we propose a new machine learning system based on several protein descriptors extracted from several protein representations, such as, variants of the position specific scoring matrix (PSSM) of proteins, the amino-acid sequence, and a matrix representation of a protein. The prediction engine is operated by an ensemble of support vector machines (SVMs), with each SVM trained on a specific descriptor and the results of each SVM combined by sum rule. The overall success rate achieved by our final ensemble is notably higher than previous results obtained on the same datasets using the same testing protocols reported in the literature.

MATLAB code and the datasets used in our experiments are freely available for future comparison at http://www.dei.unipd.it/node/2357.

1. Introduction

Predicting drug–target interactions is a crucial step in the drug discovery process, which itself is critical to the discovery of new medicines (Knowles and Gromo, 2003). Methods typically used to discover these interactions include text mining the literature (Zhu et al., 2005), docking simulations (Rarey et al., 1996; Chou et al., 2003), combining chemical structure, genomic sequence, and 3D structure information (Yamanishi et al., 2008). Experimental 3D structure of a target protein is essential for identifying drug–target interactions, but reliable 3D structures are not always available. For such cases, one common solution is to create a homology model based on the structure of a related protein (Chou, 2004; Jorgensen, 2004; Hillisch et al., 2004), but not all proteins have sufficiently high sequence similarity with a known 3D protein structure. In general, current methods for determining 3D protein structures are very slow and costly, and finding templates that are suitable for the homologous technique, as well as for other structural bioinformatics tools (Chou, 2004), are limited. With the availability of protein sequences generated in the post-genomic age, the pace of drug development could be accelerated if sequence-based computational methods were developed for predicting interactions between drugs and proteins (Xiao et al., 2013).

[*] Corresponding author.
E-mail address: loris.nanni@unipd.it (L. Nanni).

http://dx.doi.org/10.1016/j.jtbi.2014.06.008
0022-5193/© 2014 Elsevier Ltd. All rights reserved.

Pioneering work in this area includes the work of Yamanishi et al. (2008) and He et al. (2010). In Yamanishi et al. a computational method was proposed to identify the interaction between drug and target proteins from the integration of genomic and chemical spaces. In He et al. a similar method was proposed based on functional groups and biological features.

More recent works providing servers has focused on G-protein-coupled receptors (GPCRs) (Xiao et al., 2013) and ion channels. Being the largest family of cell surface receptors (Xiao et al., 2013), GPCRs are the targets of many drugs. They are involved in many diseases, such as, cancer, diabetes, and a number of neurodegenerative, inflammatory, and respiratory disorders. Indeed, more than 50% of prescription drugs on the market today act by targeting GPCRs (Chou, 2005). Ion channels are also excellent drug targets since dysfunctions in ion channels may lead to one of the so-called channelopathies: epilepsy, arrhythmia, and type II diabetes, which are treated with drugs that modulate the ion channels (Kaczorowski et al., 2008).

In Xiao et al. (2013), descriptors representing a drug compound and a GPCR are fused and fed into a Fuzzy K-NN prediction engine. The drug compound is formulated by its 2D fingerprint, a 256 bit-string encoding of molecular structure and properties (Eckert and Bajorath, 2007). This is one of many types of structural representations suggested in the literature. Other types include using the physicochemical properties (Laurent et al., 2006), chemical graphs (Gregori-Puigjane et al., 2011), topological indices and 3D pharmacophore patterns and molecular fields (Ren, 2002). GPCR is represented with a gray model pseudo amino acid composition (PseAAC). PseAAC (Chou, 2001), which replaces the simple amino acid composition AAC, has been used as a protein representation in a large number of applications, such as discriminating outer membrane proteins (Hayat and Khan, 2011) and predicting protein structural class (Zou et al., 2011). PseAAC composition represents a protein sequence with a discrete model without completely losing its sequence order information. The model is composed of a set of more than 20 discrete factors, where the first 20 factors represent the components of its conventional amino acid (AA) composition while the remaining factors incorporate some of its sequence order information using various modes (e.g., a series of rank-different correlation factors along a protein chain). The gray model (Ding, 1989) is then used in the encoding process. The drug and GPCR features are fused and fed into the Fuzzy K-NN engine to determine whether or not they form a GPCR–drug pair. Likewise, in Xiao et al. (2013), the ion channel and drug components are represented by fusing a drug component 2D fingerprint representation with a protein PseAAC representation. The fusion is then fed into a Fuzzy K-NN engine to determine whether or not they form an ion channel–drug pair. Two recent papers which also addressed drug–protein interactions in cellular networking are Min et al. (2013), Xiao et al. (2013).

As stated above one of the cornerstones for the authors' prediction method is the pseudo amino acid composition (PseAAC). To avoid losing many important information hidden in protein sequences, the pseudo amino acid composition (Chou, 2001, 2005) or Chou's PseAAC (Lin et al., 2013) was proposed to replace the simple amino acid composition (AAC) for representing the sample of a protein. For a brief introduction about Chou's PseAAC, visit the Wikipedia web-page at http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition. Ever since the concept of PseAAC was proposed by Chou (2001), it has rapidly penetrated into almost all the fields of protein attribute prediction, such as identifying bacterial virulent proteins (Nanni et al., 2012), predicting supersecondary structure (Zou et al., 2011), predicting protein quaternary structure (Zhang et al., 2008), predicting enzyme family and sub-family classes (Zhou et al., 2007), predicting protein subcellular location (Zhang et al., 2008), predicting protein submitochondria locations (Zeng et al., 2009; Nanni et al., 2008), identifying risk type of human papillomaviruses (Esmaeili et al., 2010), predicting G-Protein-Coupled Receptor Classes (Gu et al., 2010), predicting cyclin proteins (Mohabatkar, 2010), predicting GABA(A) receptor proteins (Mohabatkar et al., 2011), and classifying amino acids (Georgiou et al., 2009), among many others. Because it has been widely and increasingly used, in addition to the web-server 'PseAAC' (Shen et al., 2008) built in 2008, recently three powerful open access soft-wares, called 'PseAAC-Builder' (Du et al., 2012), 'propy' (Cao et al., 2013), and 'PseAAC-General' (Du et al., 2014), were established: the former two are for generating various modes of Chou's special PseAAC; while the 3rd one for those of Chou's general PseAAC.

As demonstrated by a series of recent publications (e.g., Feng et al., 2013; Min et al., 2013; Xu et al., 2013; Fan et al., 2014; Guo et al., 2014; Liu et al., 2014; Qiu et al., 2014) and summarized in a comprehensive review (Chou and Shen, 2009), to develop a really useful statistical predictor for a biomedical system, we need to address the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the statistical samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public or to make the code available. Below, in the next sections, let us describe how to address these problems.

The aim of this work is to increase the performance of previous predictors for identifying protein–drug interactions using an ensemble of SVMs as the prediction engine. Each SVM is trained using a different protein descriptors (detailed in Section 2) based on the following representations: the position specific scoring matrix (PSSM) of the proteins, the amino-acid sequence, and a matrix representation of a protein. Moreover, we report on the performance of new proteins descriptors introduced in this paper. The features that describe a given interaction are obtained by concatenating a protein descriptor with the 2D molecular fingerprint of the drug. We test our system on three large datasets already well-studied in the literature. As reported in Section 3, our system significantly outperforms previous approaches in the tested datasets.

2. Pattern representation and feature extraction

In this study we deal with the protein–drug interaction problem using a machine learning approach. Our basic idea is to find a compact and effective representation of proteins+drugs that is based on a fixed length encoding scheme and that can be coupled with a general purpose classifier. This approach has been applied to several other biological problems, such as subcellular localization and protein–protein interactions, with positive results (Chou and Shen, 2007; Nanni et al., 2010).

Since the aim of our system is to predict the interaction between a protein–drug pairing, we use a descriptor which combines a descriptor for proteins with a descriptor for drugs. Moreover, because the focus of this study is on protein representations, we investigate different protein representations combined with one fixed representation for drugs.

Our classification system is an ensemble of classifiers trained using the different descriptors as illustrated in Fig. 1. Two types of protein representations are considered: one based on the amino-acid sequence and one a matrix representation. From each representation several descriptors are extracted. The first type of representation includes the simple amino acid sequence (AAS), while the second type includes four different representations: (i) position