COPICAT: A Software System for Predicting Interactions Between Chemical Compounds and Proteins by Using Two-Layer Support Vector Machine

Masahiro Yokota 1  Nobuyoshi Nagamine 1  Yohei Sugawara 1  Kris Popendorf 1
{yokota, nagamine, suga, krisp}@dna.bio.keio.ac.jp

Miho Uchida 2  Tatsuo Kitahashi 2  Takashi Komori 2  Yasubumi Sakakibara 1
{uchida_miho, kitahashi_tatsuo, takashi_komori}@intec-si.co.jp  yasu@bio.keio.ac.jp

1 Department of Biosciences and Informatics, Keio Univ., 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan
2 Bio Business Division, INTEC Systems Institute, Inc., 1-3-3 Shinsuna, Koto-ku, Tokyo 136-0075, Japan

Keywords: drug-target interaction, virtual screening, support vector machine

1 Introduction

Due to the limitations in experimental methods for determining interactions between proteins and chemical compounds, comprehensive in silico approach to the issue is useful in the early stages of drug discovery processes. 3D-structure-based methods such as docking analysis have been studied well in this field, which however are still time-consuming and not really feasible for genome wide application.

To attack the problem, we have developed a novel computational method for predicting drug-target interaction using two-layer Support Vector Machine (SVM) classifiers, which require only readily available biological data, i.e., amino acid sequences of protein and structure formulae of chemical compound [1,2].

2 Method

We consider the problem as binary classification of protein-compound pairs whose abstractive identities are represented numerically: An amino acid sequence of protein is converted to a multi-dimensional feature vector representing trimer histogram, another feature vector is extracted to represent substructure statistics from a structural formula of chemical compound, and then the two vectors are combined into one feature vector as the numerical representation of a protein-compound pair. SVM-based classifier is a promising approach to such a binary classification problem.

We obtained a ‘positive’ sample set, i.e., a set of protein-compound pairs that have been proven to interact each other via biological assays, from DrugBank database [3]. Along with the ‘positive’ sample set, SVM-based classifiers require ‘negative’ sample set, i.e., a set of protein-compound pairs not to interact each other. Such a ‘negative’ sample set can be extracted randomly from the whole complement set of the ‘positive’ sample set. We also have proposed more sophisticated procedure using PSOL algorithm [4].

Using the resultant ‘positive’ and ‘negative’ protein-compound pair sets, we trained two-layer SVMs. First, we trained multiple first-layer SVMs with small sample sets designed with different criteria each other. Next, using another larger sample sets, we trained a second-layer SVM whose input is a set of probabilities output from the first-layer SVMs. We evaluated variety of two-layer SVM structures and sample set design criteria [1,2].

3 Implementation

We have implemented the method as a software system COPICAT, whose web demonstration is available as COPICAT Web System [5]. The core SVM modules of COPICAT use LIBSVM [6] and the rest of the
system is mainly written in Perl. The web interface of COPICAT Web System is built with Java Servlets, Apache Tomcat, and MySQL.

COPICAT Web System provides users with several paths accessing COPICAT functions. Users can simply browse a collection of precompiled prediction results (See Fig.1), submit a prediction job using an example classifier with example input data both being ready at the web site, and even submit a training job by uploading user input data with specifying various training parameter values. When users upload their own training data, a SVM trained with the user data is added to the pre-defined first-layer SVMs and then the second-layer SVM is re-trained using the augmented first-layer outputs with both default sample data set ready at the web site and the user data set. The resultant two-layer SVM classifier belongs to the user, which can be used for another prediction job submitted by the same user.

![COPICAT prediction result database](image)

**Figure 1:** Screenshots of COPICAT prediction result database

**References**


