BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 5 2012, pages 745–746
doi:10.1093/bioinformatics/bts031

Data and text mining

Advance Access publication January 17, 2012

COPICAT: a software system for predicting interactions between proteins and chemical compounds

Yasubumi Sakakibara[1,]*, Tsuyoshi Hachiya[1], Miho Uchida[2], Nobuyoshi Nagamine[1], Yohei Sugawara[1], Masahiro Yokota[1], Masaomi Nakamura[1], Kris Popendorf[1], Takashi Komori[2] and Kengo Sato[1]

[1]Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Yokohama 223-8522 and [2]Advanced Technology Research and Development Institute, INTEC Inc., Tokyo 136-8637, Japan

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Since tens of millions of chemical compounds have been accumulated in public chemical databases, fast comprehensive computational methods to predict interactions between chemical compounds and proteins are needed for virtual screening of lead compounds. Previously, we proposed a novel method for predicting protein–chemical interactions using two-layer Support Vector Machine classifiers that require only readily available biochemical data, i.e. amino acid sequences of proteins and structure formulas of chemical compounds.

In this article, the method has been implemented as the COPICAT web service, with an easy-to-use front-end interface. Users can simply submit a protein–chemical interaction prediction job using a pre-trained classifier, or can even train their own classification model by uploading training data. COPICAT's fast and accurate computational prediction has enhanced lead compound discovery against a database of tens of millions of chemical compounds, implying that the search space for drug discovery is extended by >1000 times compared with currently well-used high-throughput screening methodologies.

Availability: The COPICAT server is available at http://copicat.dna.bio.keio.ac.jp. All functions, including the prediction function are freely available via anonymous login without registration. Registered users, however, can use the system more intensively.

Contact: yasu@bio.keio.ac.jp

Received on September 24, 2011; revised on January 9, 2012; accepted on January 11, 2012

1 INTRODUCTION

Due to the limitations in experimental methods for determining interactions between proteins and chemical compounds, a comprehensive computational approach is useful in the early stages of the drug discovery process. Although 3D structure-based methods such as docking analysis (Morris *et al.*, 1998) have been studied intensively in this field, they are still time-consuming and not really feasible for genome-wide application.

To address the problem, we developed a comprehensively applicable statistical prediction method for interactions between any protein and chemical compound, that requires only protein sequence data and chemical structure data and utilizes the statistical learning method of Support Vector Machines (SVM) (Nagamine and Sakakibara, 2007; Nagamine *et al.*, 2009). We have shown the usefulness of our approach in searching potential ligands binding to human androgen receptors from >19 million chemical compounds and verifying these predictions by *in vitro* binding (Nagamine *et al.*, 2009).

2 METHODS AND IMPLEMENTATION

We approach the problem as binary classification of protein–chemical pairs whose abstractive identities are represented numerically. An amino acid sequence of protein is converted to a multi-dimensional feature vector representing trimer frequencies, while a chemical compound is converted to a feature vector representing frequencies of substructure occurrences in the structural formula.

We obtained a 'positive' sample set, i.e. a set of protein–chemical pairs that have been proven to interact with each other via biological assays, from the DrugBank database (Wishart *et al.*, 2008). In addition, SVM-based classifiers require a 'negative' sample set, containing protein–chemical pairs that do not interact with each other, which can be extracted randomly from the complement set of the positive sample set. Using the resultant positive and negative protein–chemical pair sets, we trained two-layer SVMs. First, we trained each of the multiple first-layer SVMs with small sample sets designed using different criteria. Next, using another larger sample set, we trained a second-layer SVM whose input is a set of probabilities output from the first-layer SVMs.

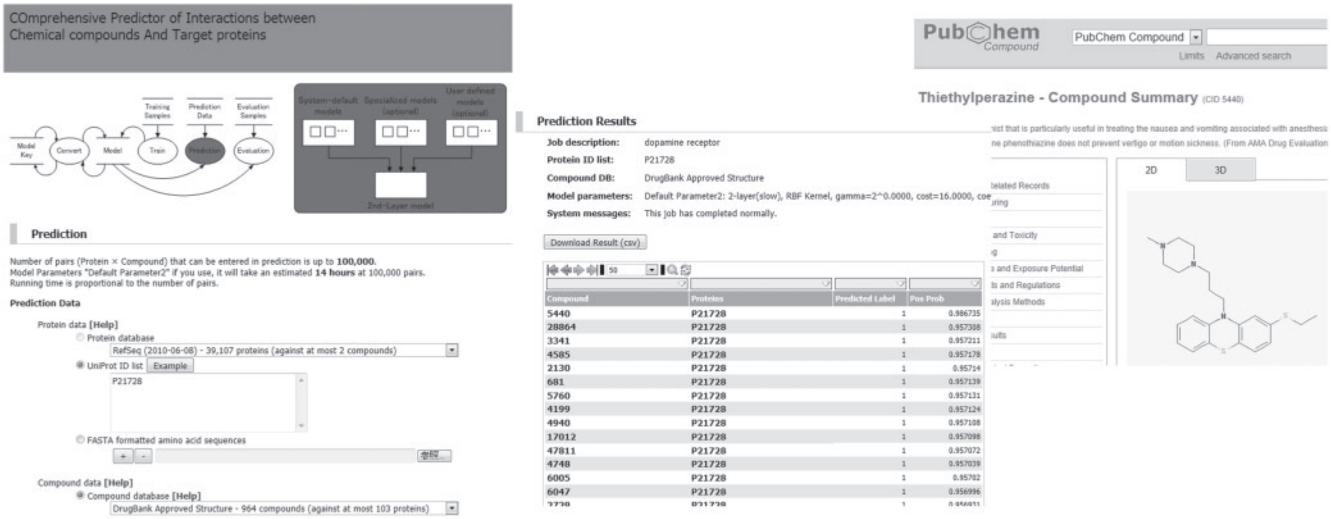
The prediction performances were evaluated by 10-fold cross-validation using the DrugBank database. Sensitivity, specificity, precision and accuracy values obtained are 0.954, 0.999, 0.984 and 0.997 respectively. Details of the algorithms and their prediction accuracy are given in Nagamine and Sakakibara (2007) and Nagamine *et al.* (2009).

The prediction method has been implemented as the web-based software system COPICAT (COmprehensive Predictor of Interactions between Chemical compounds And Target proteins), available as the COPICAT web service. The core SVM modules of COPICAT use LIBSVM (Chang and Lin, 2011) with the rest of the system written mainly in Perl and C++. The web interface is built using Java Servlets, Apache Tomcat and MySQL. COPICAT integrates SVM-based prediction programs and associated large databases of proteins and chemical compounds.

3 COPICAT WEB SERVER AND FEATURES

The COPICAT web service offers easy use of the programs without having to download the code, store huge datasets or execute the programs on computationally intensive CPUs with large memories.

*To whom correspondence should be addressed.



COmprehensive Predictor of Interactions between Chemical compounds And Target proteins

Prediction

Number of pairs (Protein × Compound) that can be entered in prediction is up to **100,000**.
Model Parameters "Default Parameter2" if you use, it will take an estimated **14 hours** at 100,000 pairs.
Running time is proportional to the number of pairs.

Prediction Data

Protein data **[Help]**

○ Protein database
RefSeq (2010-06-08) - 39,107 proteins (against at most 2 compounds)

● UniProt ID list Example
P21728

○ FASTA formatted amino acid sequences

Compound data **[Help]**

● Compound database **[Help]**
DrugBank Approved Structure - 964 compounds (against at most 103 proteins)

Prediction Results

Job description: dopamine receptor
Protein ID list: P21728
Compound DB: DrugBank Approved Structure
Model parameters: Default Parameter2: 2-layer(slow), RBF Kernel, gamma=2^0.0000, cost=16.0000, coe
System messages: This job has completed normally.

Download Result (csv)

Compound	Proteins	Predicted Label	Pos Prob
5440	P21728	1	0.986735
28864	P21728	1	0.957308
3341	P21728	1	0.957211
4585	P21728	1	0.957178
2130	P21728	1	0.95714
681	P21728	1	0.957139
5760	P21728	1	0.957131
4199	P21728	1	0.957124
4940	P21728	1	0.957108
17012	P21728	1	0.957098
47811	P21728	1	0.957072
4748	P21728	1	0.957039
6005	P21728	1	0.95702
6047	P21728	1	0.956996

Thiethylperazine - Compound Summary (CID 5440)

nist that is particularly useful in treating the nausea and vomiting associated with anesthesi
ne phenothiazine does not prevent vertigo or motion sickness. (From AMA Drug Evaluation

elated Records
uring
and Toxicity
g
and Exposure Potential
ls and Regulations
alysis Methods
ults

Fig. 1. Screenshots of a prediction job page (left), and the prediction results page (middle) on the COPICAT server and a linked PubChem DB page (right).

Prediction job: user chooses 'Prediction' from the side menu and submits the user's input data consisting of pairs of proteins and chemical compounds. The prediction result is returned as a list of protein–chemical pairs with classification labels '1 (interact)' or '−1 (non interact)' and their probabilities. Each protein and chemical compound is linked to a UniProt or PubChem database entry.

Prediction against databases: the COPICAT web service performs a prediction job against a database by receiving the user's chemical compound or protein data as a query and returning the prediction results against all molecules in one of the system-side databases. The following databases are available for use in the predictions: 'Ensembl' (79,063 entries), 'RefSeq' (39,107 entries), 'DrugBank Approved Drug Targets' (456 entries), 'DrugBank Approved Structure' (964 entries) and 'PubChem (23,130,811 entries)'. Prediction against PubChem is realized as a specialized function 'Chemical-BLAST'.

Training job: when a user uploads his own training data, an 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.

(1) Sample data: the default positive samples were obtained from DrugBank. In addition, users can upload their own positive samples.

(2) First-layer model addition: users can choose to add pre-trained models to the first layer in three ways: (i) the user's own previously trained model. (ii) the system-side specialized models: 'Nuclear Receptor', 'Ion Channel', 'GPCR', and 'Enzyme'. (iii) other user trained models via the model circulation function.

Circulation of user-trained models: COPICAT web service provides a function to circulate a user-trained model to other users wishing to combine the model with their own two-layer model. In this function, each user-trained model is assigned a model ID. Users, who are allowed to access the model, can reload the model through the ID certification process.

Job example: a prediction job with 'dopamine receptor' (UniProt ID: P21728) submitted against 'DrugBank Approved Structures' returns the prediction result containing 22 candidate compounds such as CID:5440 (Thiethylperazine, a dopamine antagonist), CID:28864 (Lisuride) and CID:681 (Dopamine) each of which is linked to the PubChem database, as shown in Figure 1.

Prediction accuracy: we tested the accuracy of prediction jobs against databases. In prediction jobs against 'DrugBank Approved Drug Targets' with a query of every Drugbank approved structure, 99.9% of the known true target proteins were listed as the first rank (highest probability) except multiple targets.

Funding: Bioinformatics research and development from the Japan Science and Technology Agency, in part.

Conflict of Interest: none declared.

REFERENCES

Bolton,E. *et al*. (2008) PubChem: integrated platform of small molecules and biological activities. Chapter 12 In. *Annual Reports in Computational Chemistry*, Vol. 4, American Chemical Society, Washington, DC.

Chang,C.C. and Lin,C.J. (2011) LIBSVM : a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27.

Morris,G.M. *et al*. (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, **19**, 1639–1662.

Nagamine,N. and Sakakibara,Y. (2007) Statistical prediction of protein chemical interactions based on chemical structure and mass spectrometry data. *Bioinformatics*, **23**, 2004–2012.

Nagamine,N. *et al*. (2009) Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening. *PLoS Comput. Biol.*, **5**, e1000397.

Wishart,D.S. *et al*. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–D906.