**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., 2000) 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, 2001) 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.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
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