Chapter 3.15
Prediction of Compound-Protein Interactions with Machine Learning Methods

Yoshihiro Yamanishi
Mines ParisTech, Institut Curie, Inserm U900, France

Hisashi Kashima
IBM Tokyo Research Laboratory, Japan

ABSTRACT

In silico prediction of compound-protein interactions from heterogeneous biological data is critical in the process of drug development. In this chapter the authors review several supervised machine learning methods to predict unknown compound-protein interactions from chemical structure and genomic sequence information simultaneously. The authors review several kernel-based algorithms from two different viewpoints: binary classification and dimension reduction. In the results, they demonstrate the usefulness of the methods on the prediction of drug-target interactions and ligand-protein interactions from chemical structure data and genomic sequence data.

DOI: 10.4018/978-1-60960-818-7.ch3.15

INTRODUCTION

Most drugs are small compounds which interact with their target proteins and inhibit or activate the biological behavior of the proteins. Therefore, the identification of interactions between compounds (ligands, small molecules, drugs) and proteins (targets) is an important part of genomic drug discovery. Examples of pharmaceutically useful target proteins are enzymes, ion channels, G protein-coupled receptors (GPCRs) and nuclear receptors. Owing to the completion of the human genome sequencing projects, we are beginning to understand the genomic spaces populated by these protein classes. At the same time, the high-throughput screening of large-scale chemical compound libraries with various biological assays

is enabling us to explore the chemical space of possible compounds (Kanehisa et al., 2006, Stockwell, 2000, Dobson, 2004). However, our knowledge about the relationship between the chemical and genomic spaces is very limited.

In 2003 the U.S. National Institutes of Health announced the Roadmap, which contained new chemical genomics initiatives. The aim of chemical genomics research is to relate this chemical space with the genomic space in order to identify potentially useful compounds such as imaging probes and drug leads. Toward the goal, the PubChem database was established at NCBI (Wheeler et al., 2006) in order to store various chemical information about millions of compounds, but the number of compounds with information on their target protein is very limited. This implies that many potential interactions between the chemical and genomic spaces remain undiscovered. There is therefore a strong incentive to develop new methods capable of detecting these potential compound-protein interactions efficiently.

Although some bio-technologies such as binding assays are becoming available, experimental determination of compound-protein interactions remains very challenging and expensive even nowadays. It is therefore of great practical interest to develop effective *in-silico* prediction methods which can both provide new predictions to experimentalists and provide supporting evidence to experimental studies. The computational prediction is expected to increase research productivity toward genomic drug discovery.

In this chapter we review various computational approaches to predict compound-protein interactions from chemical structures and protein sequences. From the viewpoint of machine learning, we formulate the problem of predicting compound-protein interactions, and introduce several supervised machine learning methods which have been recently developed from two different viewpoints: binary classification and dimension reduction. In the results, we show the usefulness of the methods on the predictions of compound-protein interactions from chemical structure data and genomic sequence data. We also discuss the characteristics of the methods, and show some perspectives toward future work.

BACKGROUND

A variety of computational approaches have been developed to analyze and predict compound-protein interactions. One of the most commonly used is docking simulations (Rarey, Kramer, Lengauer, & Klebe, 1996, Cheng et al., 2007). However, the docking cannot be applied to proteins whose 3D structures are unknown, so this limitation is serious for membrane proteins. For example, there are only two GPCRs with 3D structure information (bovine rhodopsin and human β_2-adrenergic receptor) as of writing. Therefore it is difficult to use the docking simulations on a large scale. Another unique approach is text mining which are usually based on key-word searching in a huge number of literatures (Zhu, Okuno, Tsujimoto, & Mamitsuka, 2005), but it suffers from an inability to detect new biological findings and the problem of redundancy in the compound names and protein names in the literature. Recently, a classification of target proteins based on their ligand structures has been performed (Keiser et al., 2007) and an analysis of the drug-target network has revealed characteristic features of its network topology (Yildirim, Goh, Cusick, Barabasi, & Vidal, 2007). However, neither protein sequence information nor chemical structure information were taken into consideration simultaneously.

The current state-of-the-art involves more integrative methods that simultaneously take into account compound chemical structures, protein sequences, and the currently known compound-protein interactions. A straightforward supervised approach for predicting compound-protein interactions is to use binary classification methods where they take compound-protein pairs as an input for machine learning classifiers such as