D

Data Mining of Chemogenomics Data Using Activity Landscape and Partial Least Squares

Kiyoshi Hasegawa
Chugai Pharmaceutical Company, Kamakura Research Laboratories, Japan

Kimito Funatsu
University of Tokyo, Japan

INTRODUCTION

Until recently, drug discovery has long been a multi-disciplinary effort to optimize molecular structures. It is estimated that, out of the 25000 human genes supposed to encode for 3000 proteins (Russ, 2005, p. 1607), only 800 proteins have currently been investigated by the pharmaceutical industry (Paolini, 2006, p. 805). In 10^{60} chemical space, medicinal chemists have provided only 10 million chemical structures using the technology of the miniaturization and parallelization of molecular synthesis (Lipinski, 2004, p. 855). Therefore, only a small fraction of molecules describing the current chemical space has been tested on a fraction of the entire protein space. Chemogenomics is the new inter-disciplinary field, which attempts to fully match protein space and chemical space, and ultimately identify all active molecules of all proteins (Caron, 2001, p. 464-470).

By definition, chemogenomics data is a two-dimensional matrix, where proteins are usually reported as columns and molecules as rows, and where reported values are usually biological activities. Since this matrix is sparse, several computational methods are actively developed to supplement time-consuming and costly experiments. They are either designed to fill rows and thus profile a molecule towards a heterogeneous set of proteins or to fill columns and thus identify active molecules for an existing protein (Rognan, 2007, p. 38).

In the field of drug design, data mining methods for predicting molecular selectivity within protein families are highly attractive. This is because molecules having biological activities against multi-proteins will cause many unfavorable side effects and toxicities. Molecular selectivity is an index indicating how many proteins are influenced by a single molecule. Lower molecular selectivity indicates safer molecule for human. From the perspective of above mentioned-safety, visualization of molecular selectivity against multi-proteins is of great value (Agrafiotis, 2007, p. 5926; Schneider, 2009 p. 258; Lounkine, 2010, p. 68). Activity landscape (AL) is a sophisticated-graphical representation for this purpose (Hasegawa, 2010, p. 793; Bajorath, 2012, p. 463). However, AL concept has intrinsic limitation. AL is primarily descriptive in nature, rather than predictive. AL is designed to provide access to complex structure-activity relationship patterns in large data sets, but not to answer which molecules to be synthesized next. Accordingly, a supplementary predictive method is needed.

BACKGROUND

In this study, we combined AL and partial least squares (PLS) for analyzing the aminergic G protein-coupled receptor (GPCR) data set. PLS is a statistical method that bears some relation to principal components regression. It finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Each AL was created from the inhibitory activity values of molecules against each GPCR (Peltason, 2010, p. 1021). After assembling all ALs, the inhibitory activity values in ALs were correlated with the sequence data of GPCRs by PLS (Hasegawa, 2012, p. 766). AL for new GPCR could be estimated

DOI: 10.4018/978-1-4666-5888-2.ch165

from the established PLS models. We successfully predicted the inhibitory activity values for the external molecules not included in the training data set.

MATERIAL AND METHODS

Data Set

We collected a data set of human aminergic GPCR inhibitors from the GVK data base (http://www.gvkbio.com/informatics.html). The aminergic GPCRs, also known as seven-transmembrane domain receptors constitute a large protein family of receptors that sense molecules outside the cell and activate inside signal transduction pathways and, ultimately, cellular responses. GPCRs are the important target proteins for pharmaceutical industry. GVK data base is a commercial data base for depositing chemical structures and target proteins and the associated biological activities compiling from all literatures and patents. The inhibitory activity was expressed as the logarithm of the reciprocal IC_{50} value (pIC_{50}), where IC_{50} represents the micro-molar concentration at which 50% inhibition is achieved. This data set is the same as used in a previous study (Hasegawa, 2013, p. 85).

The inhibitory activity data stored in the GVK data base are incomplete and there are many missing data points for pairs of molecules and GPCRs. To estimate the missing data points, we performed a PLS analysis against each GPCR. In this case, the extended-connectivity fingerprints of depth 6 (ECFP_6) were used as chemical descriptors (Rogers, 2010, p. 742). ECFP_6 is a binary-based fingerprint for molecular representation. Each bit in ECFP_6 represents a specific substructure within a molecule. PLS models with Q^2 values greater than 0.5 were used to predict the missing data points for GPCRs. Q^2 represents the squared correlation coefficient value derived from cross-validation (CV). CV is a pseudo-assessment method for the prediction ability of the model using the internal data splitting method. For consistency, observed inhibitory activity values were replaced by their predicted values. Table 1 shows the aminergic GPCRs used in this study. The total data set comprised a matrix of 6185 molecules against 16 GPCRs. The ECFP_6 calculation and PLS analysis were performed using Pipeline Pilot of Accelrys (http://accelrys.co.jp/).

Atom Coloring

The regression coefficient value obtained from the PLS model provides useful information on how substructures of molecules are related to the inhibitory activity against the specific GPCR. The original numerical digit is of limited utility; thus the regression numerical coefficient value was transformed into the atom coloring format used in the previous Bayesian analysis of CYP3A4 substrate/non-substrate classification (Hasegawa, 2010, p. 19).

Analogous to the Bayesian classification, the atom score was derived from the regression coefficient value of each substructure in the PLS model. The regression coefficient value of each ECFP_6 substructure was divided by the number of heavy atoms present in the substructure, and the resulting score value was assigned to each atom (Metz, 2007). The atom scores in a molecule were highlighted by the five-graded colors. The atom scores were calculated using the original script written in our laboratory in the R environment (http://www.r-project.org/).

AL

AL is defined by the distance between any pair of molecules with their biological activities (Bajorath, 2012, p. 463). The distance between two molecules was calculated as the Euclidean distance between their ECFP_6 descriptors. The Euclidean distance was defined according to the following equation:

$$\delta_{ij} = \sqrt{N_i + N_j - 2N_{ij}} \quad (1)$$

where N_i and N_j denote the number of ECFP_6 binary bins present in molecules i and j, respectively. N_{ij} denotes the number of binary bins shared by both molecules. Multi-dimensional scaling (MDS) was used to project multi-dimensional data into 2D chemical space. MDS aims to preserve the relative distance between any pair of molecules by minimizing the deviation from the ideal relationships (Borg, 2005).

Biological activity values were added to the data points in 2D chemical space in order to generate AL. In general, however, the data points are sparse and unevenly distributed and must be interpolated to obtain coherent chemical space. For this purpose, a