Chemometrics and Intelligent Laboratory Systems 135 (2014) 166–171

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Application of orthogonal L-shaped PLS to chemogenomic data and its chemical interpretation from predictive and orthogonal latent variables

Kiyoshi Hasegawa [a], Kimito Funatsu [b,*]

[a] Chugai Pharmaceutical Company, Kamakura Research Laboratories, Kajiwara 200, Kamakura, Kanagawa 247-8530, Japan
[b] The University of Tokyo, Department of Chemical System Engineering, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan

ARTICLE INFO

Article history:
Received 12 February 2014
Received in revised form 18 April 2014
Accepted 20 April 2014
Available online 26 April 2014

Keywords:
Chemogenomics
Orthogonal PLS
Orthogonal LPLS
Adenosine receptor
Chemical interpretation

ABSTRACT

We carried out a validation study of the orthogonal L-shaped PLS (OLPLS) method using chemogenomic data based on adenosine receptor inhibitor activity measurements. Using OLPLS, the ligand and protein descriptors could be connected to eight adenosine receptor inhibitor activities. The fingerprints representing specific chemical substructures on the ligands were used as the ligand descriptors, while z-scales were used as the protein descriptors. Three clusters were observed in the chemical and protein spaces from the predictive scores and loadings. From these, the predictive and orthogonal ligand structure fragments towards three adenosine receptors could be successfully elucidated. The predictive fragment for the human adenosine 2A receptor was confirmed by comparison to the X-ray crystal structure. As expected, the orthogonal fragments contained no physicochemical features required for specific interaction with the adenosine receptors.

1. Introduction

Until recently, drug discovery has long been a multi-disciplinary effort aimed at optimizing ligand properties such as potency, selectivity and bio-availability towards single molecular targets. It is estimated that out of the 25,000 human genes thought to encode the approximately 3000 druggable targets [1], only 800 of these protein targets have been investigated by the pharmaceutical industry [2]. Moreover, compared with the 10^{60} virtual ligand compounds present in various databases, medicinal chemistry has provided only ten million chemical structures, the latter using the technology of the miniaturization and parallelization of compound synthesis [3]. Only a small fraction of the ligands describing current chemical space have therefore been tested on only a fraction of the entire protein space [4].

Chemogenomics is a relatively new inter-disciplinary field that attempts to fully match target and ligand spaces, so as to ultimately identify all potential ligands of all targets [5]. Chemogenomics has received significant attention in the pharmaceutical industry because of the discovery of new inhibitory ligands of various primary targets as well as the potential unfavorable off-targets of various ligands that can result in negative side effects. By common definition, chemogenomic data comprises a two-dimensional matrix, where proteins are reported as columns and ligands as rows, and where reported values are usually inhibitor activities. The chemogenomic matrix can therefore be described as two matrices consisting of ligand and protein descriptors.

A bi-modal PLS approach, termed L-shaped PLS (LPLS), is used to connect ligand and protein descriptors to their biological activities. The LPLS approach explores relationships between the matrix columns and rows by building bi-modal models [6]. Besides constructing a regression model for the response matrix X_1 and the ligand matrix X_2, LPLS builds a further regression model connecting the weights or loadings of X_1 to the protein matrix, X_3 [7,8]. Orthogonal LPLS (OLPLS) was later developed by combining LPLS with the orthogonal concept [9] for separating predictive and orthogonal variations from data sources; variations that chemogenomic datasets inherently include [10]. This is important from the point of chemical interpretation as it helps to mitigate the risk of over-fitting the data.

Many chemogenomic modeling studies have adopted a kernel approach combined with a non-linear method and then chemical interpretability is less biased [11,12]. Chemical interpretation is valuable for generating hypotheses and knowledge, which are the final goals of molecular design. The OLPLS method may also prove suitable for chemical interpretation, and the corresponding diagnostic plots could guide the design of novel inhibitors against orphan protein targets. However, OLPLS has only been applied to analytical chemical datasets thus far [10], and has not yet been tested in chemogenomic studies focusing on chemical interpretation.

In this paper, we validated the application of OLPLS to chemogenomic data using adenosine receptor inhibitor activity data as the dataset. By using OLPLS, we were able to connect the ligand and protein descriptors to eight adenosine receptor inhibitor activities (towards the rat adenosine 1, 2A, 2B, 3 and human adenosine 1, 2A, 2B, 3 receptors). For ligand descriptors, we used the ECFP_6 fingerprints representing specific

* Corresponding author.
E-mail address: funatsu@chemsys.t.u-tokyo.ac.jp (K. Funatsu).

http://dx.doi.org/10.1016/j.chemolab.2014.04.015
0169-7439/

Table 1
SVR models for eight adenosine receptor targets. Num represents the number of molecules with observed inhibitor activity towards the targets. C, Nu and Sigma are parameters in the SVR models. R^2 and Q^2 represent the squared and ten-fold cross-validated correlation coefficient values, respectively.

No	Targets	Num	C	Nu	Sigma	R2	Q2
1	Rat_A1	2216	2	0.4	0.0313	0.823	0.608
2	Rat_A2A	2051	1	0.5	0.0313	0.824	0.656
3	Rat_A2B	803	1	0.6	0.0313	0.805	0.579
4	Rat_A3	327	2	0.4	0.0313	0.883	0.492
5	Human_A1	1635	2	0.4	0.0313	0.823	0.518
6	Human_A2A	1526	1	0.6	0.0313	0.863	0.645
7	Human_A2B	780	1	0.6	0.0313	0.863	0.635
8	Human_A3	1661	2	0.4	0.0313	0.866	0.602

chemical substructures. For protein descriptors, 27 z-scales representing nine unique amino acids were used. We identified three clusters in the chemical and protein spaces from the resulting predictive scores and loadings. From this data, the predictive and orthogonal fragments on the ligand structures required for association with three adenosine receptors (human adenosine 2A, 2B, 3 receptors) were successfully elucidated. The predictive fragment for human adenosine 2A receptor (a furan ring) was confirmed by comparison to the X-ray crystal structure. As expected, the orthogonal fragments (methylene carbon or ether oxygen) contained no physicochemical features required for specific interaction with the adenosine receptors.

2. Materials and methods

2.1. Adenosine receptor inhibitor activity data

We used previously published inhibitor activity datasets for rat and human adenosine receptors [13]. The inhibitor activity was represented by the logarithm of the reciprocal K_i value (pK_i). Affinity towards the A1 receptor on HEK293 cell membranes was determined using [^{3}H]DPCPX as the radioligand. Affinity towards the A2A receptor on CHO cell membranes was determined using [^{3}H]ZM241385 as the radioligand. Affinity towards the A2B receptor on CHO cell membranes was determined using [^{3}H]PBS603 as the radioligand. Affinity towards the A3 receptor on HEK293 cell membranes was determined using [^{3}H]PSB11 as the radioligand. Final established K_i values were calculated using a nonlinear regression curve-fitting program. In total, 10,999 annotated data points for eight adenosine receptors proved feasible. The 10,999 data points were provided by Dr. Andreas Bender's group at the University of Cambridge [13]. The number of experimentally determined data points against each adenosine receptor is shown in the column labeled 'Num' in Table 1.

The complete inhibitor activity dataset was not available and there were data missing for ligand molecules and adenosine receptor targets. To fill in the missing elements in the inhibitor matrix, we performed a respective support vector regression (SVR) analysis against each adenosine receptor target [14]. The ECFP_6 fingerprints of chemical structures [15] were used as input for SVR. The output from SVR is the predicted inhibitory activity of the chemical structure. The established SVR models were used to predict the missing data for each adenosine receptor target. The optimal parameters (C, Nu and Sigma) for each SVR model were determined by ten-fold cross-validation. For consistency, the observed inhibitor activity values were replaced by the predicted values. Table 1 shows the statistical measures of the eight SVR models. In total, a full matrix of 4898 ligand compounds with potential activity towards eight adenosine targets was constructed. The SVR analysis was performed using R scripts on a Linux machine.

2.2. Ligand and protein descriptors

The ligand descriptors used in the OLPLS analysis were the ECFP_6 fingerprints [15]. These descriptors are the same as those used for constructing the full adenosine receptor inhibitor activity dataset. The 10,000 fingerprints were generated by referring to a pre-defined substructure dictionary. That is, where a molecule had a specific substructure, the corresponding bit was defined as one in 10,000 binary codes. Otherwise, the corresponding bit was defined as zero. After setting the binary codes, the bits below the variance of 0.05 were removed until a final total of 33 bits remained.

Z-scales were used as the protein descriptors. Z-scales were initially developed as descriptors of amino acids, and contain three variables labeled as z_1, z_2 and z_3 [16]. These parameters were determined by principal component analysis (PCA) of 29 physico-chemical parameters for all 20 natural amino acids. The first, second and third principal components correspond in turn to z_1, z_2, and z_3. These are tentatively interpreted as hydrophobic, steric, and electronic properties, respectively.

With the z-scales derived from the amino acids, the amino acid sequences of the adenosine receptor proteins were translated into a vector of numbers. Because the sequence of the active site of the receptors can be aligned, comparisons of resulting vectors should directly represent the protein variation between adenosine receptors. The z-scales of each aligned sequence residue form the uniform protein descriptor matrix. Among the 15 amino acid residues forming the active site, six residues (AA_01, AA_02, AA_04, AA_06, AA_09, and AA_15) are strictly conserved and were not included in the protein matrix. For each of the nine variable amino acids, three z-scales were assigned and a total of 27 z-scales were therefore used as protein descriptors. Table 2 shows the protein descriptors representing the amino acid residues of eight adenosine receptors.

2.3. OLPLS

The OLPLS approach was introduced by Lofstedt et al. in 2012 as a method for exploring consistent patterns of co-variation between three data matrices arranged in an L-shaped system, where X_2 and X_3 give additional descriptors of the columns and rows of X_1, respectively [10]. OLPLS allows for studies of predictive and orthogonal variations in both the column and row data. OLPLS takes advantage of this bimodal arrangement and allows the analysis and interpretation to

Table 2
Protein descriptors representing the amino acid residues of eight adenosine receptors. The one-letter symbols represent the amino acid residue notation.

Proteins	AA_03	AA_05	AA_07	AA_08	AA_10	AA_11	AA_12	AA_13	AA_14
Rat_A1	E	N	L	H	Q	K	S	I	Y
Rat_A2A	E	N	L	H	H	A	P	M	Y
Rat_A2B	E	N	V	H	D	K	K	M	N
Rat_A3	R	S	L	S	K	I	E	M	C
Human_A1	E	N	L	H	H	K	S	T	Y
Human_A2A	E	N	L	H	H	A	L	M	Y
Human_A2B	E	N	V	H	N	K	K	M	N
Human_A3	V	S	L	S	E	V	Q	L	Y