International Journal of Pharmaceutics 502 (2016) 61–69

Contents lists available at ScienceDirect

International Journal of Pharmaceutics

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

Prediction and characterization of P-glycoprotein substrates potentially bound to different sites by emerging chemical pattern and hierarchical cluster analysis

Xianchao Pan[a,b], Hu Mei[a,b,*], Sujun Qu[b], Shuheng Huang[b], Jiaying Sun[c], Li Yang[a,b], Hua Chen[d,**]

[a] Key Laboratory of Biorheological Science and Technology (Ministry of Education), Chongqing University, Chongqing 400044, China
[b] College of Bioengineering, Chongqing University, Chongqing 400044, China
[c] College of Chemistry and Chemical Engineering, Sichuan University of Arts and Science, Dazhou 635000, Sichuan, China
[d] College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400044, China

ARTICLE INFO

Article history:
Received 13 December 2015
Received in revised form 27 January 2016
Accepted 14 February 2016
Available online 17 February 2016

Keywords:
P-glycoprotein
Emerging chemical pattern
Substrat e-binding site
Prediction
Hierarchical cluster analysis
Molecular docking

ABSTRACT

P-glycoprotein (P-gp), an ATP-binding cassette (ABC) multidrug transporter, can actively transport a broad spectrum of chemically diverse substrates out of cells and is heavily involved in multidrug resistance (MDR) in tumors. So far, the multiple specific binding sites remain a major obstacle in developing an efficient prediction method for P-gp substrates. Herein, emerging chemical pattern (ECP) combined by hierarchical cluster analysis was utilized to predict P-gp substrates as well as their potential binding sites. An optimal ECP model using only 3 descriptors was established with prediction accuracies of 0.80, 0.81 and 0.74 for 803 training samples, 120 test samples, and 179 independent validation samples, respectively. Hierarchical cluster analysis (HCA) of the ECPs of P-gp substrates derived 2 distinct ECP groups (ECPGs). Interestingly, HCA of the P-gp substrates based on ECP similarities also showed 2 distinct classes, which happened to be dominated by the 2 ECPGs, respectively. In the light of available experimental proofs and molecular docking results, the 2 distinct ECPGs were proved to be closely related to the binding profiles of R- and H-site substrates, respectively. The present study demonstrates, for the first time, a successful ECP model, which can not only accurately predict P-gp substrates, but also identify their potential substrate-binding sites.

1. Introduction

Multidrug resistance (MDR) is a major pitfall in effective treatment of cancer, wherein chemotherapy drugs are undesirably exported from target cells by membrane-embedded pumps (Gottesman et al., 2002). P-glycoprotein (P-gp), one of the most prevalent of these efflux pumps, belongs to the ATP-binding cassette (ABC) superfamily of membrane transporters. This transporter is a single polypeptide containing 1280 residues encoded by ABCB1 gene, and is characterized by two homologous halves with pseudo-2-fold molecular symmetry. Each half consists of one transmembrane domain (TMD) responsible for substrate translocation and one cytoplasmic nucleotide-binding domain (NBD) for ATP binding and hydrolysis. P-gp can pump a wide range of structurally diverse anticancer drugs out of cells in an ATP-dependent manner (Eckford and Sharom, 2009). Thus, over-expression of P-gp in cancer cells seriously reduces intracellular concentrations of most chemotherapeutics and impairs

Abbreviations: P-gp, P-glycoprotein; ABC, ATP-binding cassette; MDR, multidrug resistance; EP, emerging pattern; ECP, emerging chemical pattern; ECPG, emerging chemical pattern group; TM, transmembrane; TMD, transmembrane domain; NBD, nucleotide-binding domain; SBS, substrate-binding site; QSAR, quantitative structure-activity relationship; CAEP, Classification by Aggregating Emerging Patterns; HCA, hierarchical cluster analysis; Acc, accuracy; Spe, specificity; TP, true positive; TN, true negative; FP, false positive; FN, false negative; AMW, average molecular weight scaled on the number of atoms; nHacc, the number of H-bond acceptors; tPSA, total polar accessible surface area.

* Corresponding author at: College of Bioengineering, Chongqing University, No. 174 Shangzhengjie, Shapingba, Chongqing 400044, China. Tel.: +86 23 65102507.
** Corresponding author at: College of Chemistry and Chemical Engineering, Chongqing University, No. 174 Shangzhengjie, Shapingba, Chongqing 400044, China. Tel.: +86 23 65106615.
E-mail addresses: meihu@cqu.edu.cn (H. Mei), chenhuacqu@aliyun.com (H. Chen).

http://dx.doi.org/10.1016/j.ijpharm.2016.02.022
0378-5173/

bioavailability. Hence, an efficient method for accurately predicting P-gp substrates is crucial for designing chemotherapeutics with good bioavailability.

To date, the binding profiles of P-gp substrates have not been fully understood, mainly due to substrate promiscuity and multiple substrate-binding sites (SBSs) in P-gp transmembrane domain. Shapiro and Ling proposed the existence of at least two SBSs, i.e., H-site and R-site registered for Hoechst 33342 and rhodamine-123, respectively (Shapiro and Ling, 1997b). According to their research, Hoechst 33342, quercetin, and colchicine would preferentially bind to H-site, while rhodamine-123, doxorubicin, daunorubicin, and other anthracyclines R-site. Other researches also declared that there are at least two main SBSs for P-gp substrates (Chufan et al., 2013; Dey et al., 1997; Loo et al., 2003a,b; Loo and Clarke, 1999; Martin et al., 2000; Pleban et al., 2005; Shapiro et al., 1999).

Over the past few decades, *in silico* quantitative structure-activity relationship (QSAR) models have been intensively proposed to predict P-gp substrates (Bikadi et al., 2011; Broccatelli, 2012; Crivori et al., 2006; de Cerqueira Lima et al., 2006; Desai et al., 2013; Gombar et al., 2004; Hammann et al., 2009; Huang et al., 2007; Levatic et al., 2013; Li et al., 2014a; Poongavanam et al., 2012; Schwaha and Ecker, 2011; Wang et al., 2005, 2011; Xue et al., 2004). There is general agreement that molecular weight or volume (Bikadi et al., 2011; Levatic et al., 2013), number of hydrogen acceptors (Desai et al., 2013; Li et al., 2014a), polar surface area (Desai et al., 2013), molecular shape (Broccatelli, 2012; Schwaha and Ecker, 2011), polarizability (Bikadi et al., 2011), and hydrophobicity (Broccatelli, 2012; Crivori et al., 2006; Wang et al., 2011) are important for substrate binding.

Although the available QSAR models have shown good predictive performances, there are many obvious drawbacks or limitations. Firstly, the sizes of datasets are generally quite small, which results in limited coverage of chemical space and poor extrapolabilities of resulting models. Secondly, the transport activities of P-gp substrates are often measured by different experimental methods, and many methods, e.g., ATPase and calcein-AM, even have intrinsic biases, which lead to the lack of confidence and often conflicting results. For example, doxorubicin classified as a P-gp substrate (Gottesman et al., 2002; Mechetner et al., 1998), was determined as a nonsubstrate in Polli's work (Polli et al., 2001). Thirdly, regression-based QSAR methods are inappropriate in many cases, where P-gp substrates tend to bind to different sites. Lastly, the available models often lack interpretabilities, due to the complexities of QSAR approaches.

Recently, emerging pattern (EP) has been introduced in chemoinformatics as a powerful tool for compound classification, especially when a few positive samples are available. Emerging pattern (EP) approach is a machine learning methodology developed in computer science to identify class-specific feature patterns for label prediction (Dong and Li, 1999; Dong et al., 1999; Li et al., 2000, 2001). This method was subsequently adopted in bioinformatics to predict gene expression patterns (Li and Wong, 2002), and then introduced in chemoinformatics termed as emerging chemical pattern (ECP) for compound classification (Auer and Bajorath, 2006, 2008b; Namasivayam et al., 2014, 2013a, b; Pan et al., 2014; Sherhod et al., 2012, 2014), and conformation analysis (Auer and Bajorath, 2008a).

In this study, ECP modeling combined by hierarchical cluster analysis (HCA) was successfully applied to predict and characterize P-gp substrates potentially bound to different sites. The results showed that ECP method can capture the subtle structural differences between P-gp substrates and nonsubstrates, and the resulting ECP model can not only accurately predict P-gp substrates, but also identify their different binding profiles and binding sites. The prediction results of the ECP model were further proved to be consistent with the experimental and molecular docking results. Taken together, this paper provided a promising all-in-one ECP model for predicting P-gp substrates as well as substrate-binding sites.

2. Materials and methods

2.1. Dataset

P-gp substrates and nonsubstrates were extracted from a dataset published by Levatic et al. (2013). In brief, Levatic et al. (2013) correlated expression levels of P-gp mRNA with cytotoxicity activities of ~13,000 compounds against 60 human cancer cell lines. The 'substrate' and 'nonsubstrate' classes were created according to two independent criteria: 'difference' and 'correlation' criterion. After a strict process of sample screening, a dataset of 934 samples (448 substrates and 486 nonsubstrates) was constructed (Levatic et al., 2013). To the best of our knowledge, this is the largest publicly available dataset for *in silico* researches. The structures of all compounds in SMILES format are freely available for download at http://pgp.biozyne.com.

In this paper, 11 metal-containing samples (7 substrates and 4 nonsubstrates) were removed from the original dataset (Table S1 in Supporting Information). Then, the remaining 923 samples (441 P-gp substrates and 482 nonsubstrates) were divided into training and test sets in accordance with literature (Levatic et al., 2013). The training set including 803 samples (386 substrates and 417 nonsubstrates, substrates/nonsubsrates = 0.93) was used for ECP modeling. The test set including 120 samples (55 substrates and 65 nonsubstrates, substrates/nonsubsrates = 0.85) was used for model validation.

In addition, an independent validation dataset collected from Broccatelli's work (Broccatelli, 2012) was used to further validate the performance of the derived models. After removing duplicated structures to the above dataset, 179 qualified samples were obtained, comprising 72 substrates and 107 nonsubstrates (Table S2 in Supporting Information).

2.2. Structural description and feature selection

After removing counterions and adding hydrogens, all molecules were optimized by MMFF94 force field (Sybyl 8.1, http://www.tripos.com). The optimal conformation of each sample was then used for structural description by PreADMET (version 2.0, http://preadmet.bmdrc.org). A total of 140 descriptors were calculated for each sample, including atom and bond counts, physicochemical, electronic, pharmacophoric and molecular surface properties. After removing the descriptors with lower variances, 89 descriptors were retained for feature selection by backward logistic regression, of which the entry and removal probability were set to 0.05 and 0.1, respectively.

2.3. Emerging chemical patterns

2.3.1. ECP concept

To derive an ECP classifier, a continuous descriptor must be firstly discretized into intervals, which generates a set of descriptor-value pairs. Here, the value is a numerical interval into which the descriptor falls. A subset of all available descriptor-value pairs can be considered as a chemical pattern. The frequency of a pattern x in a training set D is defined as the support of x in D (Eq. (1)), abbreviated $\mathrm{supp}_D(x)$.

$$\mathrm{supp}_D(x) = \frac{\mathrm{count}_D(x)}{|D|} \quad (1)$$