Advanced PLS Technique Focusing on Visualization and Chemical Interpretation - SOMPLS Analysis of Serine Protease Inhibitors -

Kiyoshi Hasegawa\textsuperscript{a}, Kimito Funatsu\textsuperscript{b,*}

\textsuperscript{a}Chugai Pharmaceutical Company, Kamakura Research Laboratories, Kajiwara 200, Kamakura, Kanagawa 247-8530, Japan

\textsuperscript{b}University of Tokyo, Department of Chemical System Engineering, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan

(Received July 12, 2010; Accepted October 13, 2010)

In quantitative structure activity relationships (QSAR), partial least squares (PLS) are of particular interest as a statistical method. Since successful applications of PLS to QSAR data set, PLS has evolved for coping with more demands associated with complex data structures. Especially, PLS variants focusing on visualization and chemical interpretation are highly desirable for molecular design. In this paper, we employed the self-organized map PLS (SOMPLS) approach to predict multiple inhibitory activities against three serine protease receptors (Factor Xa, Tryptase and urokinase-type Plasminogen Activator (uPA)). Retrosynthetic Combinatorial Analysis Procedure (RECAP) fingerprints were used as chemical descriptors that express the existence of specific substructure in the molecule. From the SOMPLS analysis and the subsequent correlation map, essential fragments for each serine protease were easily identified. From the correlation map, we designed best combinations of fragments at each substituent position for each serine protease protein. The essential fragments could be validated from X-ray crystal structures of serine protease receptors in computer graphics. SOMPLS is an unique approach that makes data-mining feasible from visualization of structure-activity data biased to ligand-based view point.

Keywords: QSAR, PLS, SOMPLS, Serine protease inhibitors, RECAP fingerprints

\textsuperscript{*}funatsu@chemsys.t.u-tokyo.ac.jp

Copyright 2010 Chemical Society of Japan
1. Introduction

It is crucial to establish the relationships between chemical structures and their biological activities in order to achieve the objective of performing better and fewer experiments. A quantitative description of the relationships is provided by the quantitative structure activity relationships (QSAR). If such a relationship can be derived from a structure-activity data, the model equation allows medicinal chemists to predict the biological activity of new molecule in advance. Partial least squares (PLS) are of particular interest because it can analyze data comprising numerous X variables that have strongly collinear and noisy characteristics. It can simultaneously model multiple biological activities Y. Furthermore, it can provide us several prediction regions and diagnostic plots as the statistical measures [1]. Since successful applications of PLS to QSAR data set, PLS variants focusing on visualization and chemical with complex data structures [2]. Especially, PLS has evolved for coping with more demands associated with complex data structures [2]. Especially, PLS variants focusing on visualization and chemical interpretation are highly desirable for molecular design.

Recently, Melssen et al. have proposed bi-directional Kohonen (BDK) in modeling multi-class problems [3]. BDK consists of two separate Kohonen maps: one for the input X and another map for the output Y. In a BDK network, the similarity between an output object Y and the units in the output map determines to a high extent the formation of the units in the input map, whereas the similarity between an input object X and the units in the input map drives dominantly the adaptations of the units in the output map. Hence, in a BDK network, two maps are updated in an alternating bi-directional way. BDK is open and transparent, allowing direct visualization and interpretation of the content of the underlying model. After that, the same group has expanded the BDK concept to its regression type. They have introduced a hyphenated regression technique that couples the transparency of BDK with the modeling power of PLS [4]. This technique is called as self-organized map PLS (SOMPLS). Analogous to the kernel based regression method such as support vector machine (SVM) [5], the kernel matrix serves as input descriptors to PLS. The kernel matrix is calculated by the similarity between the input descriptors and the weight vectors of the input map of BDK. Melssen et al. have applied SOMPLS to a calibration problem of six metabolites with unknown molecular concentrations [4]. We have carried out first application of SOMPLS to QSAR data set of three serine protease inhibitors (Thrombin, Trypsin and Factor Xa) with VolSurf descriptors [2].

In this paper, we employed the SOMPLS approach to predict multiple biological activities of another three serine protease inhibitors (Factor Xa, Trypsase and urokinase-type Plasminogen Activator (uPA)). This data set is larger than the previous serine protease data set and it would be highly expected that the performance of SOMPLS can be fully validated. Retrosynthetic Combinatorial Analysis Procedure (RECAP) fingerprints were used as chemical descriptors that express the existence of specific substructure in the molecule [6]. From the SOMPLS analysis and the subsequent correlation map, essential fragments for each serine protease were easily identified. Also, to each chemical structure, three protease inhibitory activities were predicted simultaneously thanks to the PLS framework. From the correlation map, we designed best combinations of fragments at each substituent position against each serine protease protein. The essential fragments could be validated from X-ray crystal structures of serine protease receptors in computer graphics. The rationale why these fragments show specific biological activity could be inspected.

SOMPLS would be proved to be an powerful approach in chemogenomics from the previous [2] and this study. The purpose of chemogenomics is to find any interesting inhibitors against an orphan target receptor whose function has not been understood and its inhibitor has not been known [7]. SOMPLS is an unique approach that makes data-mining feasible from visualization of structure-activity data biased to ligand-based view point.

2. Material and Methods

2.1. Data set

In this study, the data set of three serine protease inhibitors (Factor Xa, Trypsine and uPA) was used. The data set is originally 15840 combinatorial libraries against five serine protease receptors (Chymotrypsin, Factor Xa, Trypsin, Trypsine and uPA) [8]. They were synthesized based on UGI reactions consisting from 24 isocyanides, 44 aldehydes and 15 amines. General UGI reaction is shown in Figure 1. Among them, inactive compounds having the IC$_{50}$ values more than 100 uM were removed. The data set of Chymotrypsin and Trypsin inhibitors were entirely removed due to small number of active compounds. The 1390 active compounds against uPA, factor Xa and Trypsase receptors were selected to construct the full matrix for the subsequent SOMPLS analysis. The first 1200 compounds in the time course were used as the training set. The next 190 compounds were used as the validation set. Log(1/IC$_{50}$) value was used as Y descriptor block. This data set was kindly provided by Prof. Gisbert Schneider at ETH Zurich.
2.2. RECAP fragments

RECAP starts by collecting a set of chemical structures. Using the set of eleven fragmentation rules proposed in the original literature [6], each structure is then subject to be cleaved exhaustively into fragments. These rules are applied such that effectively all susceptible bonds are cleaved in a single pass. Therefore, no intermediate structures appear in the final list of fragments. Eleven fragmentation rules are shown in Figure 2. All fragments obtained are collected for the subsequent analysis. The analysis includes the existence or non-existence of each fragment which is given by binary manner. If each specific RECAP fragment exists, the bit value is set to be one, otherwise zero. After RECAP, the data set of serine protease inhibitors produced 93 unique fragments. That is, the molecular structure was represented as the 93 bit strings. The 93 RECAP fragments were used as X descriptor block. RECAP was performed by the written scripts in Molecular Operating Environment (MOE) [9].

2.3. SOMPLS

The algorithm of SOMPLS is briefly described in the followings [4]. First, a bi-directional Kohonen (BDK) network is supervised. A BDK network consists of two coupled Kohonen self-organizing feature maps. The first map (referred to as Xmap) deals with the topology of the input data X. The second map (Ymap) having the same size as that of the Xmap embeds the multivariate structure present in the corresponding output space Y. In the first updating pass of a BDK network, only the weights of the units in the Xmap are adapted. This is repeated until all objects are presented once in a random order to the BDK network. The position of the winning unit corresponds to the location in the map for which the minimum in \( S_{ListX}(i,k) \) occurs. \( S_{ListX}(i,k) \) is defined as follows.

\[
S_{ListX}(i,k) = (1 - \alpha(t))S(X_i, X_{map_j}) + \alpha(t)S(Y_i, Y_{map_j})
\]

Here \( S(X, X_{map}) \) denotes the Euclidean distances between an input object \( X \) and all the unit weights in the Xmap. The same applies to \( S(Y, Y_{map}) \) but now for the associated output object \( Y \) and the units in the Ymap. The parameter \( \alpha(t) \) regulates the relative weight between the similarities \( S(X, X_{map}) \) and \( S(Y, Y_{map}) \). In the second updating pass, the Ymap units of the BDK network are adapted according to the following equation.

\[
S_{ListY}(i,k) = \alpha(t)S(X_i, X_{map_j}) + (1 - \alpha(t))S(Y_i, Y_{map_j})
\]

As a whole, the units in the Xmap and Ymap are updated in bi-directional manner.

Next, the similarity matrix obtained between the objects and the weight vectors of the converged BDK Xmap is weighted by a kernel function. This weighted similarity matrix will hereafter be referred to as the kernel matrix. The kernel function employed is the Pearson VII Universal Kernel (PUK). The PUK kernel function can be expressed as follows:

\[
K(i, j) = \frac{1}{1 + \left(\frac{2 \beta}{\sigma} \|X_i - X_{map_j}\|^2\right)^{1/\alpha}}
\]

The core of the PUK is formed by the Euclidean distance between the input object \( X_i \) and the BDK weight vector \( X_{map_j} \). In this equation, the parameter \( \sigma \) determines the width (sharpness) of the Pearson VII function. The parameter \( \alpha \) controls the actual shape (tailing) of the function. The constant \( \beta \) in the equation is used as an internal normalization factor. This normalization assures that all distances between the input objects and the map weights are in the range [0–1].

Finally, the relationship between the kernel matrix and biological activity is investigated by P.L.S. For each

Figure 1: Combinatorial libraries based on UGI reaction.

Figure 2: Eleven fragmentation rules in RECAP.
combination of the parameters ($\sigma$ and $\omega$), the number of components of the PLS model is optimized by the internal cross-validation.

The SOMPLS analysis was performed by the MATLAB [10] code developed in the Melssen’s group [4].

3. Results and Discussion

The optimal size of the BDK network was determined to be 10*10 after several attempts. SOMPLS gave us 38-component PLS model with following performance: In the training set, $R^2 = 0.783, 0.582$ and 0.460 for Factor Xa, Tryptase and uPA, respectively. In the test set, $Q^2 = 0.643, 0.602$ and 0.425 for Factor Xa, Tryptase and uPA, respectively. The $\sigma$ and $\omega$ values in PUK function were optimized to be 4.0 and 1.0 by internal cross-validation. The BDK map for Y (Ymap) is shown in Figure 3. In this figure, red and blue colors indicate active and inactive molecules, respectively. From this figure, the distribution of active compounds of Factor Xa is similar to that of Tryptase. On the other hand, uPA is far different from Factor Xa and Tryptase. This ligand-based information cannot be obtained from only the similarities of amino acid sequences.

Figure 3: Ymap of three serine protease inhibitors (Factor Xa, Tryptase, uPA). Red and blue colors indicate active and inactive molecules, respectively.

The correlation map between Ymap and 93 RECAP fragments is shown in Figure 4 top. The x-axis is the ID number of 93 RECAP fragments and the y-axis means Factor Xa, Tryptase and uPA from the top. The red, yellow, and blue colors indicate the positive, zero, and negative coefficient values, respectively. From this figure, we designed best combinations of $R_1$, $R_2$ and $R_3$ for each serine protease. The designed best combination matrix is shown in Figure 4 bottom row. This figure says that as for $R_3$ substituent, Factor Xa has high flexibility for making accommodation two different types of scaffold. This was confirmed by three X-ray crystal structures (PDB codes: 1G2L of Factor Xa, 2BM2 of Tryptase, 1EJN of uPA). Factor Xa has the alanine residue around $R_3$ substituent. On the other hand, Tryptase and uPA have the serine residues. Due to the electrostatic nature of serine, Trptase and uPA seem to restrict the freedom of inhibitor in the binding pocket (Figure 5).

Figure 4 top row: Correlation map between Ymap and RECAP fragments. X-axis is the ID number of 93 RECAP fragments and y-axis means Factor Xa, Tryptase and uPA from the top. Red, yellow, and blue colors indicate the positive, zero, and negative coefficient values, respectively.

Figure 4 bottom row: Best combinations of $R_1$, $R_2$ and $R_3$ for each serine protease.

Figure 5: Three X-ray crystal structures around $R_3$ substituent (Factor Xa: white, Tryptase: cyan, uPA: green).

In contrast, Figure 4 bottom row indicates that around $R_1$ substituent, the fragments with the flat shape are favorable for Factor Xa and Tryptase. On the other hand,
for uPA, the small size of fragments are better. uPA has the small binding pocket around R₁ compared to other two proteins and it restricts the flexible accommodation of fragments (Figure 6).

Acknowledgement

We would like to thank Prof. Gisbert Schneider at ETH Zurich for providing us UGI combinatorial data. We also would like to appreciate Dr. Kataoka at Ryoka System for writing the scripts to enumerate RECAP fragments in MOE.

References

[1] K. Hasegawa, K. Funatsu, in Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques (Eds: H. Lodhi, Y. Yamanishi), IGI publishing in press.
[2] K. Hasegawa, K. Funatsu, Curr. Comput.-Aided Drug Des., 2010, 6, 254-262.
[3] W. Melssen, R. Wehrens, L. Buydens, Chemom.Intell.Lab.Syst., 2006, 83, 99-113.
[4] W. Melssen, B. Ustun, L. Buydens, Chemom.Intell.Lab.Syst., 2007, 86, 102-120.
[5] K. Hasegawa, K. Fuantsu, Curr. Comput.-Aided Drug Des., 2010, 6, 24-36.
[6] X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, J.Chem.Inf.Comput.Sci., 1998, 38, 511-522.
[7] D. Rognan, Mol. Inf., 2010, 29, 176-187.
[8] A. Schuller, G. Schneider, J. Chem. Inf. Model., 2008, 48, 1473-1491.
[9] Molecular Operating Environment (MOE): http://www.chemcomp.com/software-moe2009.htm
Chemical Computing Group, 1010 Sherbrooke St. W, Suite 910 Montreal, Quebec, Canada.
[10] MATLAB: http://www.mathworks.com/products/matlab/
The MathWorks, 3 Apple Hill Drive Natick, MA 01760-2098, USA.

4. Conclusion

In this paper, we employed the SOMPLS approach to predict multiple inhibitory activities against three serine protease receptors. RECAP fingerprints were used as chemical descriptors. From the SOMPLS analysis and the subsequent correlation map, essential fragments for each serine protease were easily identified. From the correlation map, we designed best combinations of fragments at each substituent position for each serine protease protein. The essential fragments could be validated from X-ray crystal structures of serine protease receptors in computer graphics.

SOMPLS would be proved to be an powerful approach in chemogenomics from the previous [2] and this study. SOMPLS is an unique approach that makes data-mining feasible from visualization of structure-activity data biased to ligand-based view point.
図示化と化学的解釈に特化した新しいPLS手法 - セリンプロテアーゼ阻害剤のSOMPLS解析 -

長谷川 清a、船津 公人b

a 中外製薬株式会社・鎌倉研究所（〒247-8530 神奈川県鎌倉市穂原200番地）
b 東京大学大学院工学系研究科（〒113-8656 東京都文京区本郷7-3-1）

(Received ; Accepted )

定量的構造活性相関 (QSAR) では、partial least squares (PLS) が統計的手法として注目されてきた。QSARデータセットに対する応用例以来、PLSは、より複雑なデータ構造に対応できるように進化してきた。そこで、特に、図示化と化学的解釈に特化したPLS法が、分子設計ではより求められるようになってきた。本研究でわれわれは、自己組織化マップPLS (SOMPLS)を3つのセリンプロテアーゼ阻害剤（Factor Xa, Tryptase, urokinase-type Plasminogen Activator (uPA)）に応用し、複数の阻害活性を予測した。分子中の特定の部分構造の存在を表すRetrosynthetic Combinatorial Analysis Procedure (RECAP) フラグメントをchemical descriptorとして利用した。SOMPLS解析とその後のcorrelation mapから、それぞれのセリンプロテアーゼに必要な必須フラグメントを容易に見つけることができた。Correlation mapから、われわれはそれぞれのセリンプロテアーゼ阻害剤に対してそれぞれの置換基の最適フラグメントの組み合わせを設計することができた。必須フラグメントについては、コンピュータグラフィックス上セリンプロテアーゼ阻害剤のX線結晶構造から合理的に説明できた。SOMPLSは、構造活性データの図示化というリガンドの観点に特化したユニークなデータマイニング手法であると言える。

キーワード: 定量的構造活性相関、PLS、SOMPLS、セリンプロテアーゼ阻害剤、RECAPフラグメント

funatsu@chemsys.t.u-tokyo.ac.jp