Chapter 8
Advanced PLS Techniques in Chemometrics and Their Applications to Molecular Design

Kiyoshi Hasegawa
Chugai Pharmaceutical Company, Japan

Kimito Funatsu
University of Tokyo, Japan

ABSTRACT
In quantitative structure-activity/property relationships (QSAR and QSPR), multivariate statistical methods are commonly used for analysis. Partial least squares (PLS) is of particular interest because it can analyze data with strongly collinear, noisy and numerous $X$ variables, and also simultaneously model several response variables $Y$. Furthermore, PLS can provide us several prediction regions and diagnostic plots as statistical measures. PLS has evolved or changed for copying with sever demands from complex data $X$ and $Y$ structure. In this review article, the authors picked up four advanced PLS techniques and outlined their algorithms with representative examples. Especially, the authors made efforts to describe how to disclose the embedded inner relations in data and how to use their information for molecular design.

INTRODUCTION
Establishing relationships between the chemical structures and their activities or properties are crucial to achieve a goal, doing better and fewer experiments. The quantitative description of the relations is the so-called quantitative structure-activity/property relationships (QSAR and QSPR) [Gedeck et al. 2008, Yap et al. 2007].

QSAR studies express the biological activity of compounds as a function of their various structural descriptors and describe how variation of the biological activity depends on change of the chemical structure. If such a relationship can be derived from the structure-activity data, the model equation allows chemists to say with some
confidence that which property has an important role in the biological activity. Furthermore, it also allows chemists some level of prediction. By quantifying physico-chemical properties, it should be possible to calculate in advance what the biological activity of a novel compound might be. Even though a compound is discovered which does not fit the model equation, it implies that some other properties are important and provides us new compound for further investigation.

The first and most important work dealing with QSAR was published by Corwin Hansch and his co-workers in the 1960s [Hansch et al. 1964, Hansch 1969]. In this pioneering framework, the Hansch equation was developed for quantitative approach to describe relationship between chemical structure and biological activity as dependent variables $y$ using linear free energy relationships related parameters as independent variables $X$. The QSAR models are developed using multiple linear regression (MLR) which is a popular classical modeling method.

The MLR model is a powerful technique for optimizing the activity of a chemical compound. With this method a basic assumption is that all the factors involved in variation in biological activity arising from the modification of the molecular structure with a congeneric series can be correlated with concomitant change of physico-chemical parameters. The great advantage of the MLR method is that a causal model is obtained and the physical meaning is obvious. However, the sever conditions must be satisfied to apply MLR. The descriptor variables are orthogonal and the number of compounds is greater than that of descriptors. Otherwise, over fitting results may be obtained and the predictive power of the model is very poor [Hasegawa et al. 2000]. Recently, in QSAR society, new 2D and 3D molecular descriptors have been proposed [Estrada 2008].

The most difficult problem is associated to relatively high uncertainty in molecular descriptors. Thus, search for more informative 2D and/or 3D molecular descriptors has been one of the main concerns in chemoinformatics [Gasteiger 2003]. If the structural information for the molecules investigated is insufficient, the model is biased. In other words, the model is under-fitting in statistical point of views.

Chemical pattern recognition (CPR) is regarded as another method for modeling QSAR and QSPR [Miyashita et al. 1993, Miyashita et al. 1994]. In scientific research, one certainly hopes to establish a global hard model. In quantum chemistry and molecular mechanics, finding a global hard model of relationships between chemical structure and property is desired. But for some complex molecules, it is difficult to perform quantum mechanical calculations. When a global hard model cannot be obtained, chemists frequently utilize other local soft models, such as analogy and similarity. For instance, an empirical rule “like dissolves like” for solubility and the well-known periodic table for chemical elements are classical analogy methods and are commonly used in chemistry. Chemical phenomena are more complex than physical ones and are affected by many unknown factors. So, the real chemical world is typically multivariate. Facing this multivariate chemical world, we must make many assumptions and/or hypotheses in order to obtain a model and then the model loses its strictness or the generality. The Hansch approach using MLR is regarded as a hard model. In CPR, local soft modeling becomes a powerful approach because soft models can be used to predict the related property and activity. The partial least squares (PLS) method can lead to local soft model solutions to chemical problems [Wold et al. 2001].

BACKGROUND

The PLS method was first developed by Herman Wold in the 1960s and was subsequently adopted by his son Svante Wold in the 1980s for regression problems in chemometrics. PLS is of particular interest because, unlike MLR, it can analyze data
Related Content

Learning and Prediction of Complex Molecular Structure-Property Relationships: Issues and Strategies for Modeling Intestinal Absorption for Drug Discovery
[www.igi-global.com/chapter/learning-prediction-complex-molecular-structure/45474?camid=4v1a](www.igi-global.com/chapter/learning-prediction-complex-molecular-structure/45474?camid=4v1a)

Modeling Ecotoxicity as Applied to some Selected Aromatic Compounds: A Conceptual DFT Based Quantitative-Structure-Toxicity-Relationship (QSTR) Analysis
[www.igi-global.com/chapter/modeling-ecotoxicity-applied-some-selected/56448?camid=4v1a](www.igi-global.com/chapter/modeling-ecotoxicity-applied-some-selected/56448?camid=4v1a)

Retrained Classification of Tyrosinase Inhibitors and “In Silico” Potency Estimation by Using Atom-Type Linear Indices: A Powerful Tool for Speed up the Discovery of Leads
[www.igi-global.com/chapter/retrained-classification-tyrosinase-inhibitors-silico/77380?camid=4v1a](www.igi-global.com/chapter/retrained-classification-tyrosinase-inhibitors-silico/77380?camid=4v1a)

Chemoinformatics on Metabolic Pathways: Attaching Biochemical Information on Putative Enzymatic Reactions
[www.igi-global.com/chapter/chemoinformatics-metabolic-pathways/45478?camid=4v1a](www.igi-global.com/chapter/chemoinformatics-metabolic-pathways/45478?camid=4v1a)