EXCLI Journal 2009;8:74-88 – ISSN 1611-2156
Received: April 28, 2009, accepted: May 3, 2009, published: May 5, 2009

Review article:

A PRACTICAL OVERVIEW OF QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP

Chanin Nantasenamat[1], Chartchalerm Isarankura-Na-Ayudhya[1], Thanakorn Naenna[2], Virapong Prachayasittikul[1,*]

[1] Department of Clinical Microbiology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[2] Department of Industrial Engineering, Faculty of Engineering, Mahidol University, Nakhon Pathom 73170, Thailand
[*] Corresponding author: Telephone: 662-441-4376, Fax: 662-441-4380
E-mail: mtvpr@mahidol.ac.th

ABSTRACT

Quantitative structure-activity relationship (QSAR) modeling pertains to the construction of predictive models of biological activities as a function of structural and molecular information of a compound library. The concept of QSAR has typically been used for drug discovery and development and has gained wide applicability for correlating molecular information with not only biological activities but also with other physicochemical properties, which has therefore been termed quantitative structure-property relationship (QSPR). Typical molecular parameters that are used to account for electronic properties, hydrophobicity, steric effects, and topology can be determined empirically through experimentation or theoretically via computational chemistry. A given compilation of data sets is then subjected to data pre-processing and data modeling through the use of statistical and/or machine learning techniques. This review aims to cover the essential concepts and techniques that are relevant for performing QSAR/QSPR studies through the use of selected examples from our previous work.

Keywords: quantitative structure-activity relationship, QSAR, quantitative structure-property relationship, multivariate analysis

INTRODUCTION

Drug discovery has often evolved from serendipitous and fortuitous findings, for example, the discovery of penicillin by Alexander Fleming in 1928 triggered the Antibiotic Revolution which contributed tremendously to humankind's quest of longevity. If not by chance, such discoveries may be achieved through random systematic experimentation or chemical intuition where combinatorial libraries are synthesized and screened for potent activities. Such approach is extremely time consuming, labor intensive, and impractical in terms of costs. A more lucrative solution to this problem is to rationally design drugs using computer-aided tools via molecular modeling, simulation, and virtual screening for the purpose of identifying promising candidates prior to synthesis.

Quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) makes it possible to predict the activities/properties of a given compound as a function of its molecular substituent. Essentially, new and untested compounds possessing similar molecular features as compounds used in the development of QSAR/QSPR models are likewise assumed to also possess similar activities/properties. Several successful

EXCLI Journal 2009;8:74-88 – ISSN 1611-2156
Received: April 28, 2009, accepted: May 3, 2009, published: May 5, 2009

QSAR/QSPR models have been published over the years which encompass a wide span of biological and physicochemical properties. QSAR/QSPR has great potential for modeling and designing novel compounds with robust properties by being able to forecast physicochemical properties as a function of structural features. The popularity of QSAR/QSPR has seen exponential growth as illustrated by a literature search in Scopus for research articles with QSAR, QSPR, structure-activity relationship, and structure-property relationship as keywords (Figure 1).

This review covers the essential concepts and history of QSAR/QSPR as well as the components involved in the development of QSAR/QSPR models. Several examples from our previous research and relevant equations are presented.

BRIEF HISTORY OF QSAR

QSAR has its origins in the field of toxicology whereby Cros in 1863 proposed a relationship which existed between the toxicity of primary aliphatic alcohols with their water solubility (Cros, 1863). Likewise, Crum-Brown and Fraser (Crum-Brown and Fraser, 1868-1869) postulated the linkage between chemical constitution and physiological action in their pioneering investigation in 1868 as follows:

> *"performing upon a substance a chemical operation which shall introduce a known change into its constitution, and then examining and comparing the physiological action of the substance before and after the change"*

Shortly after, Richet (1893), Meyer (1899), and Overton (1901) separately discovered a linear correlation between lipophilicity (e. g. oil-water partition coefficients) and biological effects (e. g. narcotic effects and toxicity). By 1935, Hammett (1935, 1937) introduced a method to account for substituent effects on reaction mechanisms through the use of an equation which took two parameters into consideration namely the (i) substituent constant and the (ii) reaction constant.

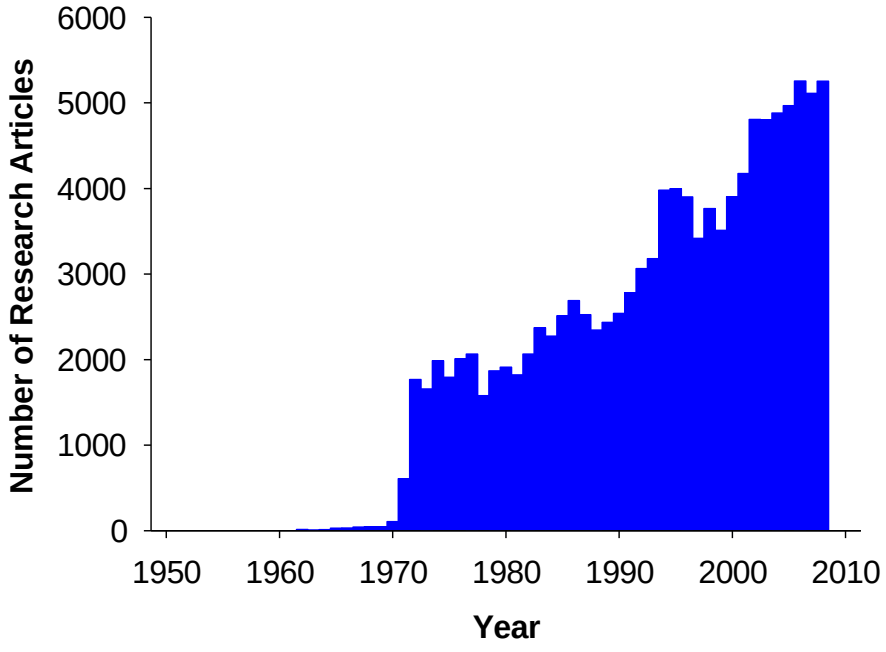

Figure 1: Number of research articles in the field of QSAR/QSPR.

EXCLI Journal 2009;8:74-88 – ISSN 1611-2156
Received: April 28, 2009, accepted: May 3, 2009, published: May 5, 2009

Complementing the Hammett's model, Taft proposed in 1956 an approach for separating polar, steric, and resonance effects of substituents in aliphatic compounds (Taft, 1956). The contributions from Hammett and Taft set forth the mechanistic basis for QSAR/QSPR development by Hansch and Fujita (1964) in their seminal development of the linear Hansch equation which integrated hydrophobic parameters with Hammett's electronic constants. An insightful account on the development of QSAR/QSPR can be found in the excellent book by Hansch and Leo (1995).

DEVELOPMENT OF QSAR MODEL

The construction of QSAR/QSPR model typically comprises of two main steps: (i) description of molecular structure and (ii) multivariate analysis for correlating molecular descriptors with observed activities/properties. An essential preliminary step in model development is data understanding. Intermediate steps that are also crucial for successful development of such QSAR/QSPR models include data pre-processing and statistical evaluation. A schematic representation of the QSAR process is illustrated in Figure 2.

Data understanding

Data understanding is a crucial step that one should not overlook as it helps the researcher to become familiar with the nature of the data prior to actual QSAR/QSPR model construction thereby reducing unnecessary errors or labors that would otherwise occur. An added benefit is that such preliminary observations can often lead to the identification of interesting associations or relationships to study. However, before exploring the data it is essential that thorough literature search on relevant background information pertaining to the biological or chemical system of interest is performed.

This can be achieved through what is known as exploratory data analysis which often starts with simple observation of the data matrix particularly the variables (also known as attributes or fields), its corresponding data types, and the data samples (also called records).

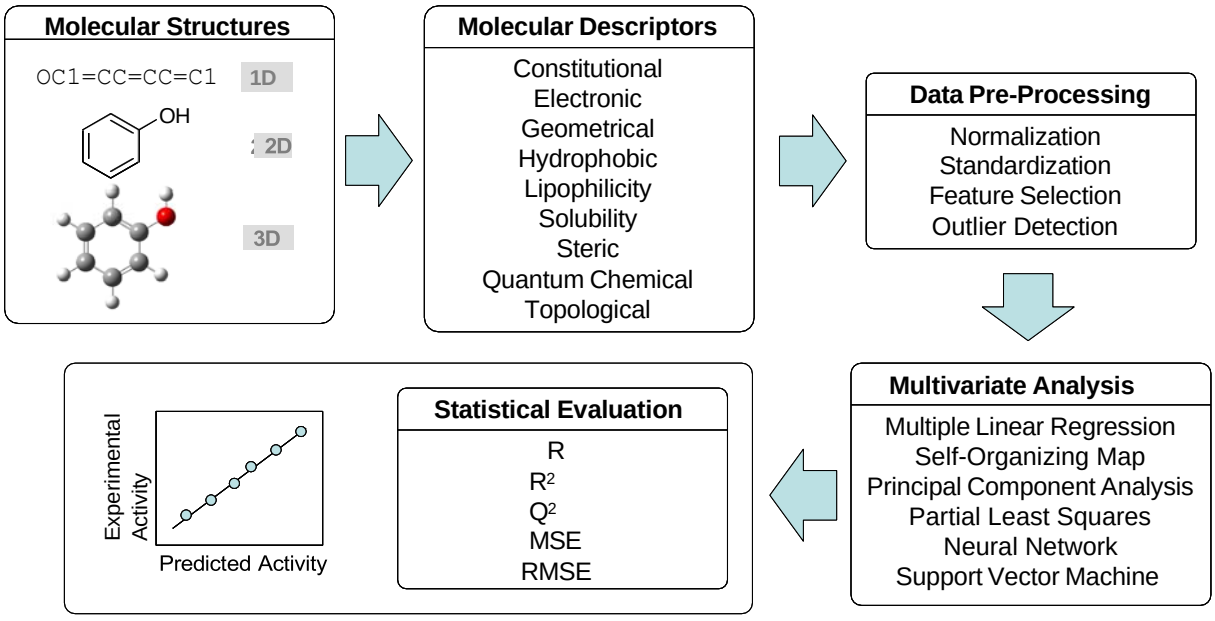

Figure 2: Schematic overview of the QSAR process.

EXCLI Journal 2009;8:74-88 – ISSN 1611-2156
Received: April 28, 2009, accepted: May 3, 2009, published: May 5, 2009

As applied to the QSAR discipline, variables represent molecular descriptors; data samples represent each unique compound; data types refer to the characteristics or the kinds of data the particular value is represented as, which essentially is qualitative or quantitative in nature. Qualitative data types are interpreted as categorical labels while quantitative data types are amendable to arithmetic operations. A more in-depth look into the nature of the data can be performed via a simple scatter plot of the variables.

Molecular descriptors

Molecular descriptors can be defined as the essential information of a molecule in terms of its physicochemical properties such as constitutional, electronic, geometrical, hydrophobic, lipophilicity, solubility, steric, quantum chemical, and topological descriptors. A more in-depth explanation of molecular descriptors can be found in the literature (Helguera et al., 2008; Karelson et al., 1996; Katritzky and Gordeeva, 1993; Labute, 2000; Randić, 1990; Randić and Razinger, 1997; Xue and Bajorath, 2000) and a more extensive treatment in the encyclopedic *Handbook of Molecular Descriptors* (Todeschini and Consonni, 2000). From a practical viewpoint, molecular descriptors are chemical information that is encoded within the molecular structures for which numerous sets of algorithms are available for such transformation.

Such descriptors could be calculated using general quantum chemical software such as Gaussian (Frisch et al., 2004), Spartan (Wavefunction, 2004), GAMESS (Gordon and Schmidt, 2005; Schmidt et al., 1993), NWChem (Kendall et al., 2000), Jaguar (Schrödinger, 2008), MOLCAS (Karlström et al., 2003), Q-Chem (Shao et al., 2006), Dalton (Angeli et al., 2005), and MOPAC (Stewart, 2009) or specialized software such as DRAGON (Talete srl, 2007; Tetko et al., 2005), CODESSA (Katritzky et al., 2005), ADRIANA.*Code* (Molecular Networks GmbH Computerchemie, 2008), and RECON (Sukumar and Breneman, 2002). Once the molecular descriptors have been calculated it will serve as independent variables for further construction of the QSAR model.

Modeled activities/properties

The activities and properties that can be modeled by QSAR/QSPR are dependent variables of the QSAR model. These dependent variables are assumed to be influenced by the independent variables which are the molecular descriptors. A variety of biological and chemical properties have successfully been modeled using the QSAR approach, such parameters are summarized in Table 1.

Table 1: Summary of biological and chemical properties explored in QSAR studies.

Biological properties	Chemical properties
Bioconcentration	Boiling point
Biodegradation	Chromatographic retention time
Carcinogenicity	Dielectric constant
Drug metabolism and clearance	Diffusion coefficient
Inhibitor constant	Dissociation constant
Mutagenicity	Melting point
Permeability	Reactivity
Blood brain barrier	Solubility
Skin	Stability
Pharmacokinetics	Thermodynamic properties
Receptor binding	Viscosity

EXCLI Journal 2009;8:74-88 – ISSN 1611-2156
Received: April 28, 2009, accepted: May 3, 2009, published: May 5, 2009

Data pre-processing

Data pre-processing can be considered to be one of the most important phase of data mining as it helps to ensure the integrity of the data set before proceeding further with data mining analysis. Essentially, the quality of a data mining analysis is a function of the quality of the data to be analyzed. This is often summarized by the "garbage in–garbage out" rule. Therefore, to obtain reliable QSAR models it is important to handle the data with great care.

Data cleaning

The preliminary steps in data pre-processing typically requires *data cleaning* as raw data often contain anomalies, errors, or inconsistencies such as missing data, incomplete data, and invalid character values which may cause trouble for data mining software if left untreated. This matter is made complicated when informations are consolidated from various sources as such data would need to be prepared to conform to designated criteria and redundant information would also need to be eliminated.

Data transformation

There exists a great deal of variability in the range and distribution of each variable in the data set. However, this may pose a problem for data mining algorithms such as neural network which involves distance measurements in the learning step. Such situation is handled by applying statistical techniques such as min-max normalization or z-score standardization. In min-max normalization, the minimum and maximum value of each variable is adjusted to a uniform range between 0 and 1 according to the following equation:

$$x_{normalized} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

where $x_{normalized}$ represents the min-max normalized value, x_i represents the value of interest, x_{min} represents the minimum value, and x_{max} represents the maximum value.

In z-score standardization, essentially the variable of interest is subjected to statistical operation to achieve mean center and unit variance according to the following formula:

$$x_{ij}^{stnd} = \frac{x_{ij} - \bar{x}_j}{\sum_{i=1}^{N} \left(x_{ij} - \bar{x}_j\right)^2 / N}$$

where x_{ij}^{stnd} represents the standardized value, x_{ii} represents the value of interest, $\bar{x}_j$ represents the mean, and N represents the sample size of the data set. The equation is essentially the difference of the value of interest and its mean followed by a division operation with the numerator, which is the variance. Practically, both normalization and standardization requires statistical operation to be applied to each individual value using the global parameter of each variable such as its minimum value, mean, or variance.

In situations where the data does not have a Gaussian (normal) distribution, simple mathematical functions can be applied to achieve normality or symmetry in the data distribution. A commonly used approach is to apply logarithmic transformation on the variable of interest in order to achieve distribution approaching normality. This is typically performed on dependent variables such as the modeled biological/chemical properties of interest whereby IC_{50} may be transformed to $\log IC_{50}$ or $-\log IC_{50}$. Practically, such mathematical operation is applied to each individual value of a given variable of interest.

Feature or variable selection

Typical data sets often contain redundant or noisy variables which make it more difficult for learning algorithms to discern meaningful patterns from the input data set of interest. For example, a data set may contain 1,500 variables but only 15 of those may contain unique and useful information while the rest may contain redundant in-

EXCLI Journal 2009;8:74-88 – ISSN 1611-2156
Received: April 28, 2009, accepted: May 3, 2009, published: May 5, 2009

formation to the aforementioned variable set. Therefore, such multicollinearity of the variables in the data set would need to be treated before proceeding with data mining analysis in order to reduce unnecessary computational resources that are required in model construction.

Similarly, feature fusion is another interesting approach for reducing the dimensionality of the variable matrix while keeping the core information intact. This is performed by merging the information of two or more variables through mathematical operations. A more in-depth treatment of such issue is addressed in the literature (Bosse et al., 2007; Goodman et al., 1997; all and McMullen, 2004; Torra, 2003).

Multivariate analysis

Multivariate analysis is essentially an approach to quantitatively discern relationships between the independent variables (e. g. molecular descriptors) and the dependent variables (e. g. biological/chemical properties of interest). The classical approach is a linear regression technique typically involving the establishment of a linear mathematical equation:

$$y = a_0 + a_1 x_1 + \mathrm{L} + a_n x_n$$

where y is the dependent variable (e. g. biological/chemical property of interest), a_0 is the y-intercept or baseline value for the compound data set, $a_1 \ldots a_n$ are the regression coefficients calculated from a set of training data in a supervised manner where the independent and dependent variables are known. The equation essentially relates the variation of biological/chemical properties as a function of the variations of the molecular substituents present in the molecular data set. Such linear approach works well for biological/chemical systems in which the phenomenon of interest is of linear nature. However, not all properties are clearly straightforward and may be non-linear in nature, therefore calls upon the use of non-linear approaches in order to properly model such properties. Non-linear techniques such as artificial neural network are a quite popular technique which possesses uncanny capability to model properties of interest. This review article will briefly cover the fundamentals of artificial neural network as an example of a non-linear learning algorithm. Other popular learning methods frequently used in the field of QSAR such as partial least squares regression (Geladi and Kowalski, 1986; Höskuldsson, 1988; Wold et al., 2001) or support vector machine (Chen et al., 2004; Cristianini and Shawe-Taylor, 2000; Wang, 2005) can be found in excellent resources elsewhere.

Artificial neural network

Artificial neural network (ANN) is a pattern recognition technique that closely resembles the inner workings of the brain which is essentially composed of interconnected neurons. The neurons receive its signals via synapses at the axon-dendron junction in which the axon of one neuron relays neurotransmitters to the dendron of another neuron. Such phenomenon is emulated by ANN's architectural design where neuronal units are interconnected to one another. A commonly used architecture as shown in Figure 3 is a three-layer feed-forward network which is comprised of (i) input layer, (ii) hidden layer, and (iii) output layer. The input layer essentially passes information of the independent variables into the ANN system; therefore the number of neuronal units present in the input layer is equal to the number of independent variables in the data set. The connections among neurons are assigned numerical values known as weights. The information from the input layer is relayed to the hidden layer for pattern recognition processing and predictions will then be passed from the hidden layer to the output layer. In a back-propagation algorithm, the error is calculated, which is derived from the difference between the predicted value and the actual value, and if it is acceptable then the learning process will stop otherwise signals will

EXCLI Journal 2009;8:74-88 – ISSN 1611-2156
Received: April 28, 2009, accepted: May 3, 2009, published: May 5, 2009

be sent backwards to the hidden layer for further processing and weight readjustments. This is performed iteratively until a solution is reached and learning is terminated.

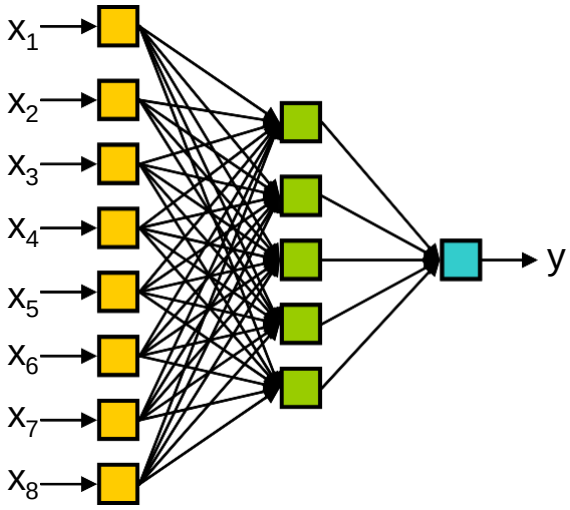

Figure 3: Schematic representation of artificial neural network.

Parameter optimization

In deriving a robust QSAR model, it is essential to optimize the parameters of the learning technique of interest. Such approaches could be performed via a systematic and empirical grid search or via stochastic approaches using techniques such as Monte Carlo or genetic algorithm. A typical systematic grid search is performed from a predetermined minimum to maximum value which essentially is dependent on the parameter to be optimized. The step size between such parameter interval can be initially large in order to minimize computational resources. From this preliminary calculation, the optimal regions are then identified and a more refined parameter search can then be performed using a more stringent approach by narrowing the step size.

Statistical evaluation

In construction of a QSAR model, it is essential to validate the model as well as apply statistical parameters to evaluate its predictive performance.

Model validation

The predictive performance of a data set can be assessed by dividing it into a training set and a testing set. The training set is used for constructing a predictive model whose predictive performance is evaluated on the testing set. Internal performance is typically assessed from the predictive performance of the training set while external performance can be assessed from the predictive performance of the independent testing set that is unknown to the training model. A commonly used approach for internal validation is known as the *N*-fold cross-validation where a data set is partitioned into *N* number of folds. For example, in a 10-fold cross-validation 1 fold is left out as the testing set while the remaining 9 folds are used as the training set for model construction and then validated with the fold that was left out. In situations where the number of samples in the data set is limited, leave-one-out cross-validation is the preferred approach. Analogously, the number of folds is equal to the number of samples present in the data set, therefore one sample is left out as the testing set while the rest is used as the training set for model construction. Finally, validation is performed on the data sample that was left out initially. This is iteratively performed until all data samples are given the chance to be left out as the testing set.

Statistical parameters

Pearson's correlation coefficient (r) is a commonly used parameter to describe the degree of association between two variables of interest. Calculated r value of two variables of interest can take a value ranging from –1 to +1 where the former indicates an indirect (negative) correlation while the latter suggests a direct (positive) correlation. For describing the relative predictive performance of a QSAR model, r is used to measure the correlation between experimental (x) and predicted (y) values of interest in order to observe the variability that exists between the variables. This is calculated according to the following equation:

EXCLI Journal 2009;8:74-88 – ISSN 1611-2156
Received: April 28, 2009, accepted: May 3, 2009, published: May 5, 2009

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}}$$

where r_{xy} is the correlation coefficient between variables x and y, n is the sample size, x is the individual value of variable x, y is the individual value of variable y, xy is the product of variables x and y, x^2 is the squared value of variable x, and y^2 is the squared value of variable y.

Root mean squared error (RMS) is another commonly used parameter for assessing the relative error of the QSAR model. RMS is computed according to the following formula:

$$RMS = \sqrt{\frac{\sum_{i=1}^{n}(x-y)^2}{n}}$$

where *RMS* is the root mean squared error, x is the experimental value of the activity/property of interest, y is the predicted value of the activity/property of interest, and n is the sample size of the data set.

F-test

The statistical significance of QSAR models are typically assessed by performing ANOVA and observing the calculated F values, which is essentially the ratio between the explained and the unexplained variance. Comparison of the performance of multiple QSAR models can be performed when all models compared have the same number of degrees of freedom meaning that the same sets of compounds and descriptors are used. Each model yields a calculated F value and the best performing model is identified as those bearing the highest value.

Degrees of freedom take into consideration the number of compounds and the number of independent variables that are present in the data set. This can be calculated using the equation $n - k - l$ where n represents the number of compounds and k represents the number of descriptors. The higher the value becomes the more reliable the QSAR model is.

Outliers

Outlying compounds are those molecules which have unexpected biological activity and do not fit in a QSAR model owing to the fact that such compounds may be acting in a different mechanism or interact with its respective target molecules in different modes (Verma and Hansch, 2005). Similarly, conformational flexibility of target protein binding site (Kim, 2007a) and unusual binding mode (Kim, 2007b) are attributed to be possible source of outliers. Mathematically speaking, an outlier is essentially a data point which has high standardized residual in absolute value when compared to the other samples of the data set. Methods for identification and treatment of outlying compounds are therefore crucial in development of reliable QSAR models (Furusjö et al., 2006). A commonly used approach for detecting outliers is performed by calculating the standardized residuals of all compounds in the data set of a QSAR model.

Predictive QSAR Model

In evaluating the performance of the constructed QSAR model, a commonly used approach in the field of QSAR follows the recommendation of Tropsha (Tropsha et al., 2003) that a predictive QSAR model should possess the following statistical characteristics:

$$q^2 > 0.5$$

$$R^2 > 0.6$$

$$\frac{(R^2 - R_0^2)}{R^2} < 0.1 \text{ or } \frac{(R^2 - R_0'^2)}{R^2} < 0.1$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15$$

where q^2 represents cross-validated explained variance, R^2 represents coefficient of determination (where R_0^2 and $R_0'^2$ repre-

EXCLI Journal 2009;8:74-88 – ISSN 1611-2156
Received: April 28, 2009, accepted: May 3, 2009, published: May 5, 2009

sents predicted versus observed activities and observed versus predicted activities, respectively), slopes k and k' of regression lines passing through the origin.

It should be noted that q^2 is calculated according to the following equation:

$$q^2 = 1 - \frac{\sum_{i=1}^{training} (y_i - y_i)^2}{\sum_{i=1}^{training} (y_i - \bar{y})^2}$$

where y_i is the measured value, y_i is the predicted value, and $\bar{y}$ is the averaged value of the entire data set, and summation applies to all compounds in the training set. Similarly, an external q^2 is calculated using compounds that are previously not used in QSAR model development. This is calculated according to the following equation:

$$q^2_{ext} = 1 - \frac{\sum_{i=1}^{training} (y_i - y_i)^2}{\sum_{i=1}^{training} (y_i - \bar{y})^2}$$

CASE STUDY

In this review, we present examples from our previous QSAR/QSPR investigations on various data sets of biological and chemical systems: (i) recognition of DNA splice junction sites (Nantasenamat et al., 2005a), (ii) prediction of antioxidant activities of phenolics antioxidants (Nantasenamat et al., 2008), (iii) prediction of binding performance of molecularly imprinted polymers (Nantasenamat et al., 2007a; Nantasenamat et al., 2005b; Nantasenamat et al., 2006), (iv) prediction of spectral properties of green fluorescent protein variants (Nantasenamat et al., 2007b), (v) prediction of anti-anthrax activity of furin inhibitors (Worachartcheewan et al., 2009), (vi) prediction of lactonolysis activity of *N*-acylhomoserine lactones. Among these, we select some representative data sets as examples of QSAR/QSPR in action.

Recognition of DNA splice junction sites

The deoxyribonucleic acid (DNA) of humans is made up of over three billion nucleotides which contains an estimated number of 30,000 genes that can express over 150,000 different proteins. The amazing fact that a limited number of genes can produce an overwhelming number of different gene products is made possible by a phenomenon known as alternative splicing where the stretch of DNA strands are cleaved at specific regions. Such regions in the DNA are known as exons (coding region) and introns (non-coding region) which are not readily discernible by simple observation of the DNA sequences. Our previous investigation has made it possible to recognize boundaries cleavage regions of the DNA called splice junction sites which are boundaries where splicing occurs.

Splice sites are essentially comprised of 2 types: (i) AG dinucleotide that borders the transition from intron to exon (intron/exon border) and (ii) GT dinucleotide that borders the transitions from exon to intron (exon/intron border). Owing to the fact that a gene is capable of expressing several distinct mRNAs encoding for different proteins, it is therefore important to be able to predict the location of DNA splice sites as it has great potential for the identification of probable gene products in unknown DNA sequences.

In our efforts to develop a computational approach for recognition of DNA splice junction sites, the DNA sequences were transformed to sequences of binary numbers by converting each nucleotide to a four digit binary code where nucleotides adenine, cytosine, guanine, and thymine are represented as 0001, 0010, 0100, and 1000, respectively. Each entry of the data set describes information surrounding the splice junction site, particularly 15 nucleotides upstream and downstream resulting in a total of 32 nucleotides. This information serves as independent variables while the dependent variable is the class of splice junction site which was labeled as one of three possible types (going from 5' to 3' or

left to right of the splice site): (i) Intron-AG-Exon, (ii) Exon-GT-Intron, and (iii) unknown-AG or GT-unknown. The data set is made of a total of 1,424 human DNA sequences that is divided into two portions: (i) a training set of 1,000 sequences and (ii) a testing set of 424 sequences. Various predictive models were developed using three different types of learning algorithm comprising of (i) self-organizing map, (ii) back-propagation neural network, and (iii) support vector machine.

Predicting the antioxidant activity of phenolic antioxidants

Reactive oxygen species (ROS) are produced during normal aerobic metabolism. Antioxidants are biomolecules which scavenge and reduce the deleterious effects of these free radicals. Under normal physiological conditions, equilibrium exists between the production and elimination of free radicals. Such equilibrium may be perturbed by environmental factors to trigger a condition known as oxidative stress, which may result in oxidative damage to various biomolecules such as DNA, RNA, proteins, and membrane lipids. Antioxidant enzymes and compounds that are present inherently in living organisms as well as those acquired from nutrition play crucial role in combating the deleterious effects of ROS. Therefore, the ability to predict the antioxidant activity, in terms of the bond dissociation enthalpy, offers great potential for designing more robust antioxidant compounds.

This was addressed in our previous investigation on the structure-activity relationship of a library of phenolic antioxidants. Multivariate analysis of the QSAR model was performed by support vector machine using molecular descriptors derived from quantum chemical calculations as independent variables to predict the antioxidant activity, which is the dependent variable. The aim of the study was to develop a rapid approach to assess the antioxidant activity of the phenolic antioxidants using readily available quantum chemical descriptors. Such descriptors were calculated at various theoretical levels in order to select the level which gave good performance while at the same time consume minimal computational resources. The theoretical levels consisting of the semi-empirical Austin Model 1 (AM1), Hartree-Fock with 3-21g(d) basis set, Becke's three parameter Lee-Yang-Parr (B3LYP) with 3-21g(d) basis set, and B3LYP with 6-31g(d) basis set were tested with multiple linear regression. Results indicated that AM1 and B3LYP/3-21g(d) were the best performing levels as observed from correlation coefficient of 0.897 and 0.917, respectively, and root mean squared error of 1.974 and 1.777, respectively. Such results outperformed those of HF/3-21g(d) and B3LYP/6-31g(d) which had lower correlation coefficient than the previous two at 0.761 and 0.730 respectively, while having higher root mean squared error at 4.624 and 4.773, respectively.

Refinement of the predictive model was performed using support vector machine, which is a more robust learning classifier, to yield significant improvements with correlation coefficients of 0.968 and 0.966, respectively, for models using descriptors derived from B3LYP/3-21g(d) and AM1 calculations. Likewise, the root mean squared error showed substantial decline to 1.122 and 1.247, respectively, for B3LYP/3-21g(d) and AM1 descriptors.

Predicting the imprinting factor of molecularly imprinted polymers

Molecular imprinting is a technology which enables the production of macromolecular matrices which can bind to template molecules of interest and function as artificial receptors, antibodies, and enzymes. These molecularly imprinted polymers (MIPs) are produced by polymerization of cross-linking monomers with the self-assembled template-monomer adducts. The template molecules are then extracted from the polymers to reveal complementary binding cavities that are specific to the original template molecule.

We have developed an approach to calculate the interaction strength of template molecules with its complementary functional monomers. This methodology essentially correlates the molecular properties of template-monomer adducts with its respective interaction strength in a quantitative manner via multivariate analysis. The molecular properties were derived from quantum chemical calculations to serve as quantitative description of the template molecules and functional monomers. Artificial neural network implementing the back-propagation algorithm was used as the multivariate analysis method.

The data sets used was comprised of two types of polymer: (i) irregularly-sized particles that was prepared by traditional bulk polymerization and (ii) uniformly-sized particles that was prepared by multi-step swelling or precipitation polymerization. The former yielded rather poor predictivity with correlation coefficient of 0.382 while the latter gave more robust results with correlation coefficient of 0.946. Reasons for such disparity in the predictive performance was attributed to the fact that the irregularly-sized MIPs had rather heterogeneous properties in terms of the (i) number of binding sites, (ii) distribution of the binding sites, (iii) size, and (iv) shape.

In the molecular imprinting literature, uniformly-sized MIPs has gained wide recognition for its larger surface area, monodispersity, and colloidal stability. Such fact was in line with the predictive performance of the devised QSAR model where uniformly-sized MIPs gave high predictive performance than the heterogeneous irregularly-sized MIPs.

Predicting GFP spectral properties

A practical example of QSAR/QSPR in action is modeling the spectral properties of Green Fluorescent Protein (GFP) from the Pacific Northwest jellyfish Aequorea victoria. Owing to its autofluorescent nature, GFP is an amazing protein which finds extensive applications in life sciences as reporters for gene expression, protein localization, protein-protein interaction, protein-lipid interaction, structural and behavioral determination of macromolecules and as analytical sensors. Much effort has been put forth to enhance the utility of such proteins by expanding the palette of colors which can be afforded by GFP and GFP-like proteins. The relationship between the structures of GFP chromophores and their respective spectral properties had been established in our previous study (Nantasenamat et al., 2007b).

In such investigation, the excitation and emission maximas of 19 GFP color variants and 29 synthetic GFP chromophores were modeled using multiple linear regression, partial least squares regression, and back-propagation neural network. Molecular descriptions of the GFP chromophores were used as independent variables and the spectral properties (e. g. excitation and emission maximas) were used as dependent variables.

For development of the QSPR model, molecular descriptors were derived from three software packages: (i) Spartan'04, (ii) E-Dragon, and (iii) RECON. Sparan'04 is a quantum chemical package which calculates the electronic properties of the chromophores. E-Dragon is an online version of the Dragon software package which can compute over 1,600 molecular descriptors spanning 20 categorical types. RECON is a software package used for deriving charge-based descriptors for the molecules of interest. Comparative assessment of the predictive performance for the QSPR model derived from the three software packages were carried out. Results indicated that the quantum chemical descriptors derived from Spartan'04 were most suitable for QSPR development as the selected descriptors could properly account for the substituent effects of the GFP chromophores.

In preliminary trials, the predictive performance of the QSPR model was relatively low for the data set comprising of 19 GFP color variants.Taking a closer look into the details of the QSPR model, it was found that the molecular structures did not reflect

the actual protonation state that was present in natural biological systems. The *p*-hydroxybenzylidene chromophores of GFP is present in 2 protonation forms, namely the protonated and deprotonated forms which are responsible for the major absorbance peak at 395 nm and the minor absorbance peak at 475 nm, respectively. The preliminary QSPR models were derived from GFP chromophores which were all drawn in the protonated form. This does not reflect the actual protonation states, therefore correction to the chromophore protonation state was performed by drawing chromophores with 395 nm absorbance peak in the protonated form and 475 nm absorbance peak in the deprotonated form. Consequently, the predictive performance of the QSPR model improved drastically from ($r_{excitation}$ = 0.3272, $RMS_{excitation}$ = 57.7310) and ($r_{emission}$ = 0.7209, $RMS_{emission}$ = 32.1526) to ($r_{excitation}$ = 0.9795, $RMS_{excitation}$ = 8.8237) and ($r_{emission}$ = 0.9067, $RMS_{emission}$ = 15.7614) for structures not taking the protonation state into consideration and for structures taking the protonation state into consideration, respectively.

In regards to the synthetic GFP chromophores, the absorbance spectra indicated that the compound is present in the protonated form. Such QSPR model gave satisfactory performance as the drawn structures accurately reflected those present in natural biological systems with correlation coefficient and root mean squared error for the excitation and the emission maxima of ($r_{excitation}$ = 0.9335, $RMS_{excitation}$ = 9.9095) and ($r_{emission}$ = 0.9626, $RMS_{emission}$ = 9.7508), respectively.

CONCLUSION

The past few decades have witnessed much advances in the development of computational models for the prediction of a wide span of biological and chemical activities that are beneficial for screening promising compounds with robust properties. In this review article, we have provided a brief introduction to the concepts of QSAR along with examples from our previous investigations on diverse biological and chemical systems. It should be noted that the applicability of QSAR models are only useful in the domains that they were trained and validated. As such, QSAR models spanning wider domains of molecular diversity have the benefit of being valid for wider spans of molecules. It is also interesting to note that there are many paths for researchers in the field of QSAR/QSPR in their quest of establishing relationships between structure and activities/properties. Such abstract nature holds the beauty of the field as there are endless possibilities in reaching the same destination of designing novel molecules with desirable properties.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge financial support from the Young Scholars Research Fellowship to C. Nantasenamat (No. MRG5080450) from the Thailand Research Fund and the governmental budget of Mahidol University (B.E. 2551-2555).

REFERENCES

Angeli C, Bak KL, Bakken V, et al. DALTON, a molecular electronic structure program. Release 2.0; 2005.

Bosse E, Roy J, Wark S. Concepts, models, and tools for information fusion. Norwood, MA: Artech House, Inc., 2007.

Chen N, Lu W, Yang J, Li G. Support vector machine in chemistry. Singapore: World Scientific Publishing, 2004.

Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press, 2000.

Cros AFA. Action de l'alcohol amylique sur l'organisme. Strasbourg, University of Strasbourg, Thesis, 1863.

Crum-Brown A, Fraser TR. On the connection between chemical constitution and physiological action. Pt 1. On the physiological action of the salts of the ammonium bases, derived from Strychnia, Brucia, Thebia, Codeia, Morphia, and Nicotia. T Roy Soc Edin 1868-1869;25:151-203.

Frisch MJ, Trucks GW, Schlegel HB, et al. Gaussian 03W, Revision C.02. Wallingford: Gaussian Inc., 2004.

Furusjö E, Svenson A, Rahmberg M, Andersson M. The importance of outlier detection and training set selection for reliable environmental QSAR predictions. Chemosphere 2006;63:99-108.

Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Anal Chim Acta 1986;185:1-17.

Goodman IR, Mahler RPS, Nguyen HT. Mathematics of data fusion. Dordrecht, Boston: Kluwer Academic Publishers, 1997.

Gordon MS, Schmidt MW. Advances in electronic structure theory: GAMESS a decade later. In: Dykstra CE, Frenking G, Kim KS, Scuseria GE (eds.): Theory and applications of computational chemistry: the first forty years (pp 1167-1189). Amsterdam: Elsevier, 2005.

Hall DL, McMullen SAH. Mathematical techniques in multisensor data fusion. Boston, MA: Artech House, Inc., 2004.

Hammett LP. Some relations between reaction rates and equilibrium constants. Chem Rev 1935;17:125-36.

Hammett LP. The effect of structure upon the reactions of organic compounds. Benzene derivatives. J Am Chem Soc 1937;59: 96-103.

Hansch C, Fujita T. p-σ-π analysis. A method for the correlation of biological activity and chemical structure. J Am Chem Soc 1964;86:1616-26.

Hansch C, Leo A. Exploring QSAR. Washington, DC: American Chemical Society, 1995.

Helguera AM, Combes RD, Gonzalez MP, Cordeiro MN. Applications of 2D descriptors in drug design: a DRAGON tale. Curr Top Med Chem 2008;8:1628-55.

Höskuldsson A. PLS regression methods. J Chemometr 1988;2:211-28.

Karelson M, Lobanov VS, Katritzky AR. Quantum-chemical descriptors in QSAR/ QSPR studies. Chem Rev 1996;96: 1027-44.

Karlström G, Lindh R, Malmqvist P-Å, et al. MOLCAS: a program package for computational chemistry. Comput Mater Sci 2003;28:222-39.

Katritzky AR, Gordeeva EV. Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. J Chem Inf Comput Sci 1993;33:835-57.

Katritzky AR, Karelson M, Petrukhin R. CODESSA PRO, Florida, U.S.A., 2005.

Kendall RA, Aprà E, Bernholdt DE, et al. High performance computational chemistry: An overview of NWChem a distributed parallel application. Comput Phys Commun 2000;128:260-83.

Kim K. Outliers in SAR and QSAR: 2. Is a flexible binding site a possible source of outliers? J Comput Aid Mol Des 2007a;21: 421-35.

Kim K. Outliers in SAR and QSAR: Is unusual binding mode a possible source of outliers? J Comput Aid Mol Des 2007b;21: 63-86.

Labute P. A widely applicable set of descriptors. J Mol Graph Model 2000;18:464-77.

Meyer H. Zur Theorie der AIkoholnarkose. Arch Exp Path Pharm 1899;42:109-18.

Molecular Networks GmbH Computerchemie. ADRIANA.Code, Erlangen, Germany, 2008.

Nantasenamat C, Naenna T, Isarankura-Na-Ayudhya C, Prachayasittikul V. Recognition of DNA splice junction via machine learning approaches. Excli J 2005a;4:114-29.

Nantasenamat C, Naenna T, Isarankura Na Ayudhya C, Prachayasittikul V. Quantitative prediction of imprinting factor of molecularly imprinted polymers by artificial neural network. J Comput Aid Mol Des 2005b;19:509-24.

Nantasenamat C, Tantimongcolwat T, Naenna T, Isarankura-Na-Ayudhya C, Prachayasittikul V. Prediction of selectivity index of pentachlorophenol-imprinted polymers. Excli J 2006;5:150-63.

Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. Quantitative structure-imprinting factor relationship of molecularly imprinted polymers. Biosens Bioelectron 2007a;22:3309-17.

Nantasenamat C, Isarankura-Na-Ayudhya C, Tansila N, Naenna T, Prachayasittikul V. Prediction of GFP spectral properties using artificial neural network. J Comput Chem 2007b;28:1275-89.

Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. Prediction of bond dissociation enthalpy of antioxidant phenols by support vector machine. J Mol Graph Model 2008;27:188-96.

Overton CE. Studien über die Narkose. Jena: Fischer, 1901.

Randić M. The nature of chemical structure. J Math Chem 1990;4:157-84.

Randić M, Razinger M. On characterization of 3D molecular structure. In: Balaban AT (ed.): From chemical topology to three-dimensional geometry. New York: Plenum Press, 1997.

Richet MC. Note sur le rapport entre la toxicité et les propriétes physiques des corps. Compt Rend Soc Biol (Paris) 1893;45:775-6.

Schmidt MW, Baldridge KK, Boatz JA, et al. General atomic and molecular electronic structure system. J Comput Chem 1993;14: 1347-63.

Schrödinger, Inc. Jaguar, Version 7.5207; Portland, OR, 2008.

Shao Y, Molnar LF, Jung Y, et al. Advances in methods and algorithms in a modern quantum chemistry program package. Phys Chem Chem Phys 2006;8:3172-91.

Stewart J. MOPAC2009, Colorado, USA, 2009.

Sukumar N, Breneman CM. RECON, Version 5.5; New York, USA, 2002.

Taft RW. Separation of polar, steric and resonance effects in reactivity. In: Newman MS (ed.): Steric effects in organic chemistry (pp 556-675). New York: Wiley, 1956.

Talete srl. DRAGON, Milano, Italy, 2007.

Tetko IV, Gasteiger J, Todeschini R, et al. Virtual computational chemistry laboratory-design and description. J Comput Aid Mol Des 2005;19:453-63.

Todeschini R, Consonni V. Handbook of molecular descriptors, Vol. 11. Weinheim: Wiley-VCH, 2000.

Torra V. Information fusion in data mining. Secaucus, NJ: Springer-Verlag, 2003.

Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 2003;22:69-77.

Verma RP, Hansch C. An approach toward the problem of outliers in QSAR. Bioorg Med Chem 2005;13:4597-621.

Wang L. Support vector machines: theory and applications. New York: Springer-Verlag, 2005.

Wavefunction, Inc. Spartan'04, Irvine, California, USA, 2004.

Wold S, Trygg J, Berglund A, Antti H. Some recent developments in PLS modeling. Chemometr Intell Lab 2001;58:131-50.

Worachartcheewan A, Nantasenamat C, Naenna T, Isarankura-Na-Ayudhya C, Prachayasittikul V. Modeling the activity of furin inhibitors using artificial neural network. Eur J Med Chem 2009;44:1664-73.

Xue L, Bajorath J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. Comb Chem High Throughput Screening 2000;3:363-72.