3D PHAMARCOPHOMORE MODELLING PART I

Pharmacophore modeling is powerful method to identify new potential drugs. Pharmacophore models are hypothesis on the 3D arrangement of structural properties such as hydrogen bond donor and acceptor properties, hydrophobic groups and aromatic rings of compounds that bind to the biological target. The pharmacophore concept assumes that structurally diverse molecules bind to their receptor site in a similar way, with their pharmacophoric elements interacting with the same functional groups of the receptor.

Validating 3D QSAR models

QSAR modeling produces predictive models derived from application of statistical tools correlating biological activity (including desirable therapeutic effect and undesirable side effects) or physico-chemical properties in QSPR models of chemicals (drugs/toxicants/environmental pollutants) with descriptors representative of molecular structure or properties. QSARs are being applied in many disciplines, for example: risk assessment, toxicity prediction, and regulatory decisions in addition to drug discovery and lead optimization. Obtaining a good quality QSAR model depends on many factors, such as the quality of input data, the choice of descriptors and statistical methods for modeling and for validation. Any QSAR modeling should

ultimately lead to statistically robust and predictive models capable of making accurate and reliable predictions of the modeled response of new compounds.

For validation of QSAR models, usually various strategies are adopted:

1. internal validation or cross-validation (actually, while extracting data, cross validation is a measure of model robustness, the more a model is robust (higher q2) the less data extraction perturb the original model);
2. external validation by splitting the available data set into training set for model development and prediction set for model predictivity check;
3. blind external validation by application of model on new external data and
4. data randomization or Y-scrambling for verifying the absence of chance correlation between the response and the modeling descriptors.

The success of any QSAR model depends on accuracy of the input data, selection of appropriate descriptors and statistical tools, and most importantly validation of the developed model. Validation is the process by which the reliability and relevance of a procedure are established for a specific purpose; for QSAR models validation must be mainly for robustness, prediction performances and applicability domain (AD) of the models.

Some validation methodologies can be problematic. For example, *leave one-out* cross-validation generally leads to an overestimation of predictive capacity. Even with external validation, it is difficult to determine whether the selection of training and test sets was manipulated to maximize the predictive capacity of the model being published.

Different aspects of validation of QSAR models that need attention includes methods of selection of training set compounds, setting training set size and impact of variable selection for training set models for determining the quality of prediction. Development of novel validation parameters for judging quality of QSAR models is also important.

QSARs (Quantitative Structure–Activity relationships) are based on the assumption that the structure of a molecule (i.e. its geometric, steric and electronic properties) must contain the features responsible for its physical, chemical, and biological properties, and on the ability to represent the chemical by one, or more, numerical descriptor(s). By QSAR models, the biological activity (or property, reactivity, etc.) of a new or untested chemical can be inferred from the molecular structure of similar compounds whose activities (properties, reactivities, etc.) have already been assessed. The QSPR (Quantitative Structure–Property relationship) acronymous is used when a property is modelled. Simply it means that "The structure of chemical compound influences its properties and bioactivity. QSAR in simplest terms, is a method for building computational or mathematical models which attempts to find a statistically significant correlation between structure and function using a chemometric technique. In terms of drug design, structure here refers to the properties or descriptors of the molecules, their substituents or interaction energy fields, function corresponds to an experimental biological/biochemical endpoint like binding affinity, activity, toxicity or rate constants. Various QSAR approaches have been developed gradually over a time span of more than a hundred years and served as a valuable predictive tool, particularly in the design of pharmaceuticals and agrochemicals. All one and two dimensional and related methods are commonly referred to as 'classical' QSAR methodologies. It is sometime used in more sense as a hansch analysis. The introduction of the Hansch model in 1964 enabled medicinal

chemists to formulate their hypothesis of structure activity relationships in quantitative terms and to check these hypotheses by means of statistical methods. From such QSAR, it is possible to elucidate the influence of various physiological properties on drug potency and to predict activity values for new compounds within certain limits. QSAR is essentially a computerised statistical method which tries to explain the observed variance in the biological effect of certain classes of compounds as a function of molecular changes caused by the substituents. It assumes that the potency of a certain biological activity exerted by a series of congeneric compounds is a function of various physicochemical parameters of the compounds. Once statistical analysis shows that certain physico-chemical properties are favourable to the concerned activity, the concerned activity can be optimized by choosing such substituents which would enhance such physicochemical properties. Description of the molecular structure, electronic orbital reactivity and the role of structural and steric components have been the subject of mathematical and statistical analysis. The ultimate object of such studies is to understand the forces governing the activity of a particular compound or a class of compounds. QSAR vary to an appreciable extent in depth and sophistication based on the nature of evaluation of structure or activity. A purposeful relation of structural variables must include steric factors, electronic features of component functional groups and the molecule as a whole.

History of Q.S.A.R

It has been nearly 40 years since the quantitative structure-activity relationship (QSAR) paradigm first found its way into the practice of agrochemistry, pharmaceutical chemistry, toxicology, and eventually most facets of chemistry. Crum-Brown and Fraser (1868) expressed the idea that the physiological action of a substance in a certain biological system(Φ) was a function (f) of its chemical composition and constitution (C).

$$\Phi = f\,C \quad \text{Equation} \qquad [1]$$

Thus, an alteration in chemical constitution, ΔC, would be reflected by an alteration in biological activity ΔΦ. Richardson (1868) expressed the chemical structure as a function of solubility. Mills (1884) developed a QSPR model for the prediction of melting and boiling points in homologous series, results were accurate to better than one degree. Richet (1893) Correlated toxicities of a set of alcohols, ethers and ketones with aqueous solubility and showed that their cytotoxicities are inversely related to their corresponding water solubilities.

Overton and Meyer (1897, 1899) correlated partition coefficients of a group of organic compounds with their anesthetic potencies and concluded that narcotic (depressant) activity is dependent on the lipophilicity of the molecules. The seminal work of Hammett (1935, 1937) gave rise to the σ-ρ culture correlated the effect of the addition of a substituent on benzoic acid with the dissociation constant, postulated electronic sigma-rho constants and established the linear free energy relationship (LFER) principle. Hammett found that a linear relationship resulted when substitutions of different groups were made to aromatic compounds.

$$\log \frac{K}{K_0} = \rho \log \frac{K'}{K_0} = \rho \sigma$$

K_0 and $K_{0'}$ are equilibrium constants for unsubstituted compounds and K and K' are the equilibrium constants for substituted compounds. Hammett used benzoic acid as reference compound yielding the σ. To interpret this equation, if the linear relation defines $\rho > 1$, then the effect of the substitutions is greater than making the same substitutions on benzoic acid. The σ describes the properties of the substitution groups. If σ is positive, the group is electron withdrawing. If σ is negative, the group is electron donating. The magnitude of σ indicates the degree of these effects. In 1939, Ferguson correlated depressant action with the relative saturation of volatile compounds in their vehicle and introduced a thermodynamic generalization to the toxicity. Bell and Roblin (1942) Studied antibacterial activities of a series of sulfanilamides in terms of their ionizations. Albert (1948) examined the effects of ionization/electron distribution and steric access on the potencies of a multitude of aminoacridines. Taft (1952) Postulated a method for separating polar, steric, and resonance effects and introduced the first steric parameter, Es. Hansch and Muir (1962) Correlated the biological activities of plant growth regulators with Hammett constants and hydrophobicity. Using the octanol/water system, a whole series of partition coefficients were measured, and thus a new hydrophobic scale was introduced. The parameter π, which is the relative hydrophobicity of a substituent, was defined in a manner analogous to the definition of sigma.

$$\Pi_X = \log P_X - \log P_H \qquad \text{Equation [2]}$$

P_X and P_H represent the partition coefficients of a derivative and the parent molecule, respectively. The contributions of Hammett and Taft together laid the basis for the development of the QSAR paradigm by Hansch and Fujita (1964), which combined the

hydrophobic constants with Hammett's electronic constants to yield the linear Hansch equation and its many extended forms.

$$\text{Log } 1/C = a\,\sigma + b\,\pi + ck$$

There is a consensus among current predictive toxicologists that Corwin Hansch is the founder of modern QSAR. In has been illustrated that, in general, biological activity for a group of 'congeneric' chemicals can be described by a comprehensive model:

$$\text{Log } 1/C_{50} = a\,\pi + b\,\varepsilon + c\,S + d \qquad \text{Equation [3]}$$

In which C, the toxicant concentration at which an endpoint is manifested (e.g. 50% mortality or effect), is related to a hydrophobicity term, p, (this is a substituent constant denoting the difference in hydrophobicity between a parent compound and a substituted analog, it has been replaced with the more general molecular term the log of the 1-octanol/water partition coefficient, log K_{ow}), an electronic term, 1, (originally the Hammett substituent constant, s) and a steric term, S, (typically Taft's substituent constant, ES). Due to the curvilinear, or bilinear, relationship between log1/C50 and hydrophobicity normally found in single dose tests the quadratic π^2 term was later introduced to the model. Hansch (1969) Developed the parabolic Hansch equation for dealing with extended hydrophobicity ranges.

$$\text{Log } 1/C = -a\,(\log P)^2 + b.\log P + c\,\sigma + k$$

Free and Wilson (1964) formulated an additive model, where the activity is discretized as a simple sum of contributions from different substituents.

$$BA = a_i x_i + u$$

BA is the biological activity, u is the average contribution of the parent molecule, and a_i is the contribution of each structural feature; x_i denotes the presence $X_i = 1$ or absence $Xi = 0$ of a particular structural fragment. Fujita and Ban (1971) simplified the Free-Wilson equation estimating the activity for the non-substituted compound of the series and postulated Fujita-Ban equation that used the logarithm of activity, which brought the activity parameter in line with other free energy-related terms.

$$\text{Log } BA = G_i X_i + u$$

In this equation, u is defined as the calculated biological activity value of the unsubstituted parent compound of a particular series. G_i represents the biological activity contribution of the substituents, whereas X_i is ascribed with a value of one when the substituent is present or zero when it is absent. Kubinyi (1976) Investigated the transport of drugs *via* aqueous and lipoidal compartment systems and further refined the parabolic equation of Hansch to develop a superior bilinear (non-linear) QSAR model.

$$\text{Log } 1/C = \text{a. log P- b.log } (\beta.\ P + 1) + k$$

Hansch and Gao (1997) Developed comparative QSAR (C-QSAR), incorporated in the CQSAR program. Heritage and Lowis in 1997 Developed Hologram QSAR (HQSAR), where the structures are converted into all possible fragments, which are assigned specific integers, and then hashed into a fingerprint to form the molecular hologram. The bin occupancies of these holograms are used as the QSAR descriptors, encoding the chemical and topological information of molecules. Cho and workers (1998) Developed Inverse QSAR, which seeks to find values for the molecular descriptors that possess a desired activity/property value. In other words, it consists of finding the optimum sets of descriptor values best matching a target activity and then generating a focused library of candidate structures from the solution set of descriptor values. Labute (1999) Developed Binary QSAR to handle binary activity measurements from high throughput screening (*e.g.*, pass/fail or active/inactive), and molecular descriptor vectors as input. A probability distribution for actives and inactives is then determined based on Bayes' Theorem.

QSAR Theory

The overall goals of QSAR retain their original essence and remain focused on the predictive ability of the approach and its receptiveness to mechanistic interpretation. Rigorous analysis and fine-tuning of independent variables have led to an expansion in development of molecular and atom-based descriptors, as well as descriptors derived from quantum chemical calculations and spectroscopy. It is now possible not only to develop a model for a system but also to compare models from a biological database and to draw analogies with models from a physical organic database. This process is dubbed model mining and it provides a sophisticated approach to the study of chemical-biological interactions. All QSAR analyses are based on the assumption of linear additive contributions of the different structural properties or features of a compound to its biological activity, provided that there are no

nonlinear dependences of transport or binding on certain physicochemical properties. This simple assumption is proven by some dedicated investigations, for example the scoring function of the de novo drug design program LUDI (eqn.1); in addition, the results of many Free Wilson and Hansch analyses support this concept.

$$\Delta G_{binding} = \Delta G_0 + \Delta G_{hb} + \Delta G_{ionic} + \Delta G_{lipo} + \Delta G_{rot} \qquad (1)$$

Overall loss of translational and rotational entropy,

$\Delta G_{binding} = +5.4 \text{KJ mol}^{-1}$

Ideal neutral hydrogen bond, $\Delta G_{hb} = -4.7 \text{ KJ mol}^{-1}$

Ideal ionic interaction, $\Delta G_{ionic} = -8.3 \text{ KJ mol}^{-1}$

Lipophilic contact, $\Delta G_{lipo} = -0.17 \text{ J mol}^{-1} \text{Å}^{-2}$

Entropy loss per rotatable bond of the ligand, $\Delta G_{rot} = +1.4 \text{ KJ mol}^{-1}$

Eqn.1 correlates the free energy of binding, $\Delta G_{binding}$, with a constant term, ΔG_0, that describes the loss of overall translational and rotational degrees of freedom and ΔG_{hb}, ΔG_{ionic} and ΔG_{lipo}, which are structure-derived energy terms for neutral and charged hydrogen bond interactions and hydrophobic interactions between the ligand and the protein; ΔG_{rot} describes the loss of internal rotational degrees of freedom of the ligand. Eqn1 1 holds for a wide range of energy values: the $\Delta G_{binding}$ of 45 different ligand-protein complexes ranges from -9 to -76 KJ mol-1, which corresponds to binding constants between 2.5×10^{-2} M and 4×10^{-14} M; its standard deviation of 7.9 KJ mol-1 corresponds to a mean error of about 1.4 log units in the prediction of ligand binding constants from the mathematical model. Because of the extra thermodynamic relationship between free energies ΔG and equilibrium constants K (eqn.2) or rate constants k (k_{on} = association constant, k_{off} = dissociation constant of ligand-receptor complex formation), the logarithms of such values can be correlated with binding affinities.

$$\Delta G = -2.303 \text{ RT log } K = -2.303 \text{ RT log } k_{on} / k_{off} \qquad (2)$$

Logarithms of molar concentrations C that produce a certain biological effect can be correlated with molecular features or with physiological properties that are also free-energy-related equilibrium constants; normally the logarithms of inverse concentrations, log1/C, are used to obtain larger values for the more active analogs.

References

1. Kim, K.H., Greco, G. and Novellino, E., *A critical review of recent CoMFA applications*, In Kubinyi, H., Folkers, G., and Martin, Y.C., (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 257–316.
2. Dunn III, W.J. and Hopfinger, A.J., *3D QSAR of flexible molecules using tensor representation*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 167–182.

3. Hahn, M. and Rogers, D., *Receptor surface models*, in Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 117–134.
4. Heritage, T.W., Ferguson, A.M., Turner, D.B. and Willett, P., *EVA — a novel theoretical descriptor for QSAR studies*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 381–398.
5. Klebe, G., *Comparative molecular similarity indices analysis — CoMSIA*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 87–104.
6. Walters, D.E., *Genetically evolved receptor models (GERM) as a 3D QSAR tool*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 159–166.
7. Wade, R.C., Ortiz, A.R. and Gago, F., *Comparative binding energy analysis*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 19–34.
8. Holloway, M.K., *A priori predicition of ligand affinity by energy minimization*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 63–84.
9. Todeschini, R. and Gramatica, P., *New 3D molecular descriptors: The WHIM theory and QSAR applications*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 355–380.
10. Silverman, B.D., Platt, D.E., Pitman, M. and Rigoutsos, I., *Comparative molecular moment analysis (COMMA)*, in Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 183–196.
11. Jain, A.N., Koile, K. and Chapman, D., *Compass: Predicting biological activities from molecular surface properties — performance comparisons on a steroid benchmark*, J. Med. Chem., 37 (1994) 2315–2327.
12. Martin, Y.C., Kim, K.-H. and Lin, C.T., *Comparative molecular field analysis: CoMFA*, In Charton, M. (Ed.) Advances in quantitative structure property relationships, JAI Press, Greenwich, CT, 1996, pp. 1–52.
13. Greco, G., Novellino, E. and Martin, Y.C., *Approaches to 3D-QSAR*, In Martin, Y.C. and Willett, P. (Eds.) Designing bioactive molecules: Three-dimensional techniques and applications, America Chemical Society, Washington, DC, 1997 (in press).
14. Ajay and Murcko, M.A., *Computational methods to predict binding free-energy in ligand–receptor complexes*, J. Med. Chem., 38 (1995) 4953–4967.
15. Kollman, P.A., *Advances and continuing challenges in achieving realistic and predictive*

simulations of the properties of organic and biological molecules, Acc. Chem. Res., 29 (1996) 461–469.
16. Bush, B.L. and Nachbar Jr., R.B., *Sample-distance partial least-squares — PLS optimized for many variables, with application to CoMFA*, J. Comput.-Aided Mol. Design, 7 (1993) 587–619.
17. Burger, A., *Medical chemistry — the first century*, Med. Chem. Res., 4 (1994) 3–15.
18. Willett, P., *Similarity and clustering techniques in chemical information systems*, Research Studies Press, Letchworth, 1987.
19. Hodgkin, E.E. and Richards, W.G., *Molecular similarity based on electrostatic potential and electric field*, Int. J. Quantum Chem., 14 (1987) 105–110.
20. Kier, L.B., *Molecular orbital theory in drug research*, Academic Press, New York, 1971, p. 258.