J Comput Aided Mol Des (2012) 26:39–43
DOI 10.1007/s10822-011-9511-4

PERSPECTIVE

The great descriptor melting pot: mixing descriptors for the common good of QSAR models

Yufeng J. Tseng · Anton J. Hopfinger · Emilio Xavier Esposito

Received: 15 November 2011 / Accepted: 2 December 2011 / Published online: 27 December 2011

Abstract The usefulness and utility of QSAR modeling depends heavily on the ability to estimate the values of molecular descriptors relevant to the endpoints of interest followed by an optimized selection of descriptors to form the best QSAR models from a representative set of the endpoints of interest. The performance of a QSAR model is directly related to its molecular descriptors. QSAR modeling, specifically model construction and optimization, has benefited from its ability to borrow from other unrelated fields, yet the molecular descriptors that form QSAR models have remained basically unchanged in both form and preferred usage. There are many types of endpoints that require multiple classes of descriptors (descriptors that encode 1D through multi-dimensional, 4D and above, content) needed to most fully capture the molecular features and interactions that contribute to the endpoint. The advantages of QSAR models constructed from multiple, and different, descriptor classes have been demonstrated in the exploration of markedly different, and principally biological systems and endpoints. Multiple examples of such QSAR applications using different descriptor sets are described and that examined. The take-home-message is that a major part of the future of QSAR analysis, and its application to modeling biological potency, ADME-Tox properties, general use in virtual screening applications, as well as its expanding use into new fields for building QSPR models, lies in developing strategies that combine and use 1D through *n*D molecular descriptors.

Keywords QSAR · Descriptors · QSPR · 4D-QSAR · 4D-fingerprint · ADME-tox

Y. J. Tseng
Department of Computer Science and Information Engineering, National Taiwan University, No. 1 Sec. 4, Roosevelt Road, Taipei 106, Taiwan

Y. J. Tseng
Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, No. 1 Sec. 4, Roosevelt Road, Taipei 106, Taiwan

A. J. Hopfinger
College of Pharmacy MSC09 5360 1, University of New Mexico, Albuquerque, NM 87131-0001, USA

A. J. Hopfinger · E. X. Esposito
The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest, IL 60045, USA

E. X. Esposito (✉)
exeResearch, LLC, 32 University Drive, East Lansing, MI 48823, USA
e-mail: emilio@exeResearch.com

Quantitative structure–activity relationships (QSARs) are used to describe and predict therapeutic biological potency, adsorption, distribution, metabolism, excretion, toxicology (ADME-Tox) properties and properties of materials, most notably polymers. Often non-biological potency applications are referred to as quantitative structure–property relationships (QSPRs). The usefulness of QSAR analysis as a cheminformatics tool depends heavily on the ability to calculate molecular descriptors relevant to the endpoints (activity and property measures) of interest, followed by the judicious selection of descriptors—or not—to form QSAR models (equations) from a representative set of the endpoints of interest (the training set). The power of cheminformatics, specifically in this case the QSAR paradigm, is being able to examine the resulting QSAR models and gain an understanding of why certain molecular descriptors are more important than others, and how these

descriptors permit reliable endpoint predictions. In turn, realizing this capability from a QSAR model is directly dependent upon the molecular descriptors used to construct the models (the trial descriptor pool).

Generally, scientists like to employ a paradigm, or methodology, that upon first use is not overly cumbersome to apply, and often provides interesting findings. In the case of QSAR analysis this approach translates into discovering and becoming familiar with a particular class of molecular descriptors and a model generation method that results in success. QSAR models can be created using multiple linear regression methods as in classic Hansch analysis [1–3], principal component analysis [4] (PCA), partial least squares [5, 6] (PLS), artificial neural networks [7–9] (NN), evolutionary algorithms including genetic algorithms [10, 11], support vector machines [12, 13] (SVM) or combinations of these model-building and optimization methods. Continuous or classification (discriminant analysis) models employ the same general protocol: assemble the dataset of compounds and endpoints, divide the dataset into a training set and a test set, calculate the molecular descriptors, construct and validate the models, analyze and test/validate the models, and implement the information gained from the models.

The task of QSAR model construction and optimization continues to borrow from the social sciences, computer science and statistics in order to better handle the ever larger, noisier and, at times, unbalanced datasets that are increasingly generated from experimental high-throughput screening (HTS) assays. However, the molecular descriptors that comprise QSAR models are typically unchanging. Moreover, minimal thought seems to have gone into the development of methods for the customized selection and exploration of trial descriptor sets for the building of QSAR models. Molecular descriptors belong to various classes based on the type of information they are derived from and contain. Descriptors that, for example, count the number of a specific chemical entity in a molecule (e.g., the number of aromatic rings or hydrogen bond acceptors) are often referred to as traditional 1D (dimensional). There are also 2D (sometimes called 2½D; a 3D-based molecular property represented as a single numerical value; e.g., the potential energy, volume, or molecular properties mapped to the molecule's surface) and 3D (molecular interaction fields based on compound-probe interaction isosurface contours; e.g., the hydrophobic volume defined between two interaction energies; GRIND [14] and VolSurf [15–17]) molecular descriptors. Gaining more popularity are *n*D or *multi*-dimensional descriptors that are molecular features extracted from an ensemble of conformations and molecular interactions: MI-QSAR analysis [18], 4D-QSAR analysis [19–23] and 4D-Fingerprints [24]. The conformations are usually an ensemble set taken from the molecular dynamics trajectories of the compounds with corresponding key features derived from intra- and, in some cases, inter- molecular interactions. Quantum mechanical properties, such as HOMO, LUMO, ionization potential and heat of formation energies (calculated with Spartan [25], GAMESS [26], CODESSA [27] or Gaussian [28] to name a few), can be considered 2½D, 3D or 4D descriptors, depending on one's point of view, but provide high-level information relating to intra-atomic interactions and electronic structure. Commonly used cheminformatics packages such as the Chemistry Development Kit [29, 30] (CDK), CODESSA [27], Dragon [31, 32], MOLCON [33], Molecular Operating Environment [28] (MOE), Pipeline Pilot [34], and SYBYL-X [35] provide scientists a multitude of molecular descriptors that span the 1D, 2D, and 3D descriptor classes.

It is quite common, and actually almost "standard operating procedure", for QSAR models to be constructed from a single class, or type, of molecular descriptors. Conversely, it is very is infrequent for 1D and 2D descriptors and 3D and 4D descriptors to be jointly used as a trial descriptor pool to build a QSAR model. There is no logical reason for keeping these descriptor classes segregated. Certainly one can appreciate situations, based upon the endpoint of interest, where multiple classes of descriptors are needed to adequately capture the molecular features and interactions that contribute to the endpoint of interest. For example, when the endpoint is an ED50 (an in vivo measure of biological potency), intuition tells us that there may be a transport/delivery component, as well as a ligand-receptor binding component, that jointly contribute to the expression of the ED50 value. The transport component is likely best treated with 2D and 2½D thermodynamic and size descriptors, whereas the ligand-receptor binding component may be best handled using pharmacophore descriptors derived from 4D-QSAR analysis. ADME transport properties, like cell penetration, may require 1D through 4D descriptors to capture both the magnitude and direction of the solute trajectory through the membrane.

The advantages of QSAR models constructed from multiple descriptor classes have been demonstrated in the exploration of different biological systems and endpoints. Recently we have shown that merging 1D through 4D molecular descriptor sets into a single trial descriptor pool leads to multi-class continuous and classification QSAR models that are superior for the prediction of hERG cardiotoxicity when compared to models from the corresponding segregated descriptor sets. [36, 37] The transport of organic compounds through lipid assemblies of the stratum corneum for transdermal drug delivery applications has also been modeled using multiple classes of descriptors. [38] The trial descriptor pool consisted of intermolecular interactions between a membrane (monolayer of DMPC) and the penetrant as well as intramolecular (1D

and 2D molecular descriptors; classic descriptors) features of only the penetrant. Two sets of QSAR models were constructed, one from the intramolecular-only trial descriptor pool and another from the intramolecular and intermolecular trial descriptor pool. The models constructed from the mixed-class descriptor pools resulted in remarkably better models ($r^2 = 0.80$, $q^2 = 0.77$) for the prediction of skin penetration than the intramolecular-only models ($r^2 = 0.56$, $q^2 = 0.51$) [38].

Adding a known important molecular feature (e.g., log P, number of hydrogen bond donors, or total polar surface area, TPSA) to a trial descriptor pool consisting of molecular interaction fields *al a* CoMFA [39], or a set of 4D-QSAR grid cell occupancy descriptors, GCODs [20], can improve the predictive ability and usefulness of the resulting models. A thorough comparison of 4D-QSAR and CoMFA models for a steroidal dataset [40] provides insight to the importance of including other classes of molecular descriptors when creating QSAR models. The 4D-QSAR models were constructed with ($r^2 = 0.87$, $q^2 = 0.80$, 14 descriptors) and without ($r^2 = 0.85$, $q^2 = 0.76$, 14 descriptors) the calculated log P values. Hence, the inclusion of log P led to a better model as measured by both r^2 and q^2. Moreover, for this particular application, the 4D-QSAR models are superior to the CoMFA models (electrostatic only: $r^2 = 0.90$, $q^2 = 0.56$, 200 descriptors; electrostatics and sterics: $r^2 = 0.92$, $q^2 = 0.59$, 476 descriptors) based on the leave-one-out cross-validation and also test set average residuals of prediction [40]. Thus, it can be seen from this example that the class/type of descriptors used (4D GCODs versus CoMFA field descriptors) can lead to models of near identical quality with respect to r^2, but having markedly different predictive power.

If there were any reason for the segregation of classes of descriptors, it would be as a component to doing consensus modeling. One can advocate that the optimum way to do QSAR analysis involves building QSAR models for a common training set using segregated classes of descriptor sets, as well as various merged combinations of classes of descriptor sets. Comparisons of the resulting optimized QSAR models from these multiple descriptor pools should:

(a) permit identification of the best and unique set of QSAR models,
(b) provide a basis for consensus virtual screening,
(c) identify consistency, or lack thereof, in the features and properties, as portrayed by the descriptors of the QSAR models, responsible for the expression of the endpoint,
(d) establish a basis to rank the relative importance of the key descriptors from different classes in controlling the endpoint, and
(e) present a landscape of how different classes and types of descriptors are interacting with one another in expressing the endpoint.

Very little of this type of consensus modeling, using multiple classes and types of descriptors sets, has been, or is being, done. Investigators seem content to use their favorite set of descriptors and model-building technique to construct QSAR models, and pay little heed to other available QSAR modeling resources. In a study of cannabinoids [41] a single class of descriptors, namely ab initio quantum mechanical molecular descriptors, captured a set of physicochemical properties that accounted for the variance in the binding of cannabinoid ligands to cannabinoid receptor 2 (CB_2). However, this class of descriptors was unable to meaningfully describe the features for cannabinoid receptor 1 (CB_1). The authors propose that 3D modeling techniques and corresponding descriptors, such as CoMFA [39] or CoMSIA [42], are better suited to model the biological endpoints of CB_1 than quantum mechanical descriptors [41]. It would have been interesting to see this dataset explored with ab initio quantum mechanical descriptors along with 1D through 4D molecular descriptor, as individual and combined descriptor sets, to construct consensus QSAR models that would permit a more complete examination of the key molecular features.

An example of consensus and ensemble QSAR modeling, using multiple descriptor classes, is the study of a set of skin penetration enhancers [43]. A combination of classic intramolecular descriptors and 4D-fingerprints [24], based on the 4D-QSAR paradigm [20], were computed for multiple sets of skin penetration enhancers. Three types of QSAR models (a) classic descriptor models, (b) 4D-fingerprint models, and (c) classic and 4D-fingerprint models [43] were constructed for two different skin penetration enhancer training sets. The models for the first training set of 61 compounds had comparable results; the best classic model ($r^2 = 0.73$, $q^2 = 0.66$, 6 descriptors), the best 4D-Fingerprint model ($r^2 = 0.74$, $q^2 = 0.67$, 5 descriptors) and the best mixed-descriptor set model ($r^2 = 0.76$, $q^2 = 0.72$, 6 descriptors) [43]. While the models are statistically similar to one another, the descriptor make-up of the individual class models and mixed class model are significantly different from one another. The lone descriptor commonality is that the mixed class model shares a single descriptor with the classical model and also a single descriptor from the 4D-fingerprint model. The second training set of 44 skin penetration enhancers highlights the unique information contained within the multidimensional 4D-fingerprints not present in the classical molecular descriptors. Models constructed from the classic intramolecular descriptors for this dataset resulted in no significant QSAR models. However, models constructed from the mixed-class and the 4D-

fingerprints only trial descriptor pools yielded the exact same models composed only of 4D-fingerprints. Thus, in this specific application, the 4D-fingerprints out-performed the classic descriptors [43].

An important part of the future of QSAR analysis, and its application to modeling biological potency, ADME-Tox properties, as well as its expanding use into new fields for building QSPR models, lies in developing strategies that combine and use 1D through *n*D molecular descriptors. The exploration for new, relevant and high information content descriptors is not over, especially the development of methods to generate custom 3D and 4D descriptors. However, we should endeavor to combine and optimize the use of sets of descriptor classes and types that are currently available when constructing QSAR models.

References

1. Hansch C, Fujita T (1964) ρ-σ-π analysis. A method for the correlation of biological activity and chemical structure. J Am Chem Soc 86:1616–1626
2. Hansch C, Lien EJ (1968) An analysis of the structure-activity relationship in the adrenergic blocking activity of the β-haloalkylamines. Biochem Pharmacol 17:709–720
3. Hansch C, Mahoney PP, Pujita T, Muir RM (1962) Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. Nature 194:178–180
4. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24: 417–441 and 498–520
5. Wold S, Sjöström M (1998) Chemometrics, present and future success. Chemom Intell Lab Syst 44:3–14
6. Wold S, Sjöström M, Ericksson L (1998) Partial least squares projections to latent structures (PLS) in chemistry. In: von Ragué Schleyer P (ed) Encyclopedia of computational chemistry vol. 3. John Wiley & Sons, Chichester, pp 2006–2021
7. Müller K-R, Mika S, Rätsch G, Tsuda K, Schölkopf B (2001) An introduction to kernel-based learning algorithms. IEEE Transac Neural Netw 12(2):181–201
8. So S-S, Karplus M (1996) Genetic neural networks for quantitative structure-activity relationships: improvements and application of benzodiazepine affinity for benzodiazepine/$GABA_A$ receptors. J Med Chem 39:5246–5256
9. Zupan J, Gasteiger J (1999) Neural networks in chemistry and drug design. Wiley-VCH, Weinheim
10. Holland JH (1975) Adaptation in artificial and natural systems. University of Michigan, Ann Arbor
11. Rogers D, Hopfinger AJ (1994) Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. J Chem Inf Comput Sci 34(4):854–866
12. Vapnik VN (1998) Statistical learning theory. Wiley, New York, p 736
13. Vapnik VN (2000) The Nature of statistical learning theory. Springer, New York, p 314
14. Pastor M, Cruciani G, McLay I, Pickett S, Clementi S (2000) GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. J Med Chem 43(17):3233–3243
15. Cruciani G, Crivori P, Carrupt P, Testa B (2000) Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. J Mol Struct: THEOCHEM 503(1–2):17–30
16. Cruciani G, Pastor M, Guba W (2000) VolSurf: a new tool for the pharmacokonetic optimization of lead compounds. Eur J Pharm Sci 11:S29–S39
17. Cruciani G, Pastor M, Mannhold R (2002) Suitability of molecular descriptors for database mining. A comparative analysis. J Med Chem 45(13):2685–2694
18. Kulkarni AS, Hopfinger AJ (1999) Membrane-interaction QSAR analysis: application to the estimation of eye irritation by organic compounds. Pharm Res 16:1244–1252
19. Hopfinger AJ, Reaka A, Venkatarangan P, Duca JS, Wang S (1999) Construction of a virtual high throughput screen by 4D-QSAR analysis: application to a combinatorial library of glucose inhibitors of glycogen phosphorylase b. J Chem Inf Comput Sci 39(6):1151–1160
20. Hopfinger AJ, Wang S, Tokarski JS, Jin B, Albuquerque M, Madhav PJ, Duraiswami C (1997) Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. J Am Chem Soc 119(43):10509–10524
21. Klein CDP, Hopfinger AJ (1998) Pharmacological activity and membrane interactions of antiarrhythmics: 4D-QSAR/QSPR analysis. Pharm Res 15(2):303–311
22. Krasowski MD, Hong X, Hopfinger AJ, Harrison NL (2002) 4D-QSAR analysis of a set of propofol analogues: mapping binding sites for an anesthetic phenol on the $GABA_A$ receptor. J Med Chem 45(15):3210–3221
23. Santos-Filho OA, Hopfinger AJ (2001) A search for sources of drug resistance by the 4D-QSAR analysis of a set of antimalarial dihydrofolate reductase inhibitors. J Comput Aided Mol Des 15(1):1–12
24. Senese CL, Duca J, Pan D, Hopfinger AJ, Tseng YJ (2004) 4D-fingerprints, universal QSAR and QSPR descriptors. J Chem Inf Comput Sci 44(5):1526–1539
25. Spartan, Wavefunction, Inc. 18401 Von Karman Avenue, Suite 370, Irvine, CA 92612 USA, Version '10, http://www.wavefun.com/
26. Schmidt MW, Baldridge KK, Boatz JA, Elbert ST, Gordon MS, Jensen JH, Koseki S, Matsunaga N, Nguyen KA, Su S, Windus TL, Dupuis M, Jr JAM (1993) General atomic and molecular electronic structure system. J Comput Chem 14(11):1347–1363
27. CODESSA Semichem Inc., 12456 W 62nd Terrace, Suite D, Shawnee, Kansas 66216 USA, http://www.semichem.com/codessa/
28. Molecular Operating Environment (MOE), Chemical Computing Group, Inc., 1010 Sherbrooke St. W, Suite 910, Montreal, Quebec, Canada H3A 2R7, http://www.chemcomp.com
29. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) The chemistry development kit (CDK): an open-source java library for chemo- and bioinformatics. J Chem Inf Comput Sci 43(2):493–500
30. Steinbeck C, Hoppe C, Kuhn S, Flores M, Guha R, Willighagen E (2006) Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics. Curr Pharm Des 12(17):2111–2120
31. Dragon TALETE srl, Via V. Pisani, 13–20124 Milano–Italy, http://www.talete.mi.it/products/dragon_description.htm
32. Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin VA, Radchenko EV, Zefirov NS, Makarenko AS, Tanchuk VY, Prokopenko VV (2005) Virtual computational chemistry laboratory–design and description. J Comput Aided Mol Des 19(6):453–463
33. Molconn, Hall Associates Consulting, 2 Davis Street, Quincy, Massachusetts 02170 USA, http://www.molconn.com

34. Pipeline Pilot, Accelrys, Inc., 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA, http://accelrys.com/products/pipeline-pilot/
35. SYBYL-X, Tripos Inc., 1699 South Hanley Road, Saint Louis, Missouri 63144, USA, http://www.tripos.com
36. Shen M-y, B-H Su, Esposito EX, Hopfinger AJ, Tseng YJ (2011) A comprehensive SVM binary hERG classification model based on extensive but biased endpoint hERG data sets. Chem Res Toxicol 24(6):934–949
37. Su B-H, Shen M-y, Esposito EX, Hopfinger AJ, Tseng YJ (2010) In silico binary classification QSAR models based on 4D-fingerprints and MOE descriptors for prediction of hERG blockage. J Chem Inf Model 50(7):1304–1318
38. Santos-Filho OA, Hopfinger AJ, Zheng T (2004) Characterization of skin penetration processes of organic molecules using molecular similarity and QSAR analysis. Molecular Pharmaceutics 1(6):466–476
39. Cramer RD III, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1 Effect of shape on binding of steriods to carrier proteins. J Am Chem Soc 110(18): 5959–5967
40. Ravi M, Hopfinger AJ, Hormann RE, Dinan L (2001) 4D-QSAR analysis of a set of ecdysteroids and a comparison to CoMFA Modeling. J Chem Inf Comput Sci 41(6):1587–1604
41. Ferreira AM, Krishnamurthy M, Moore BM II, Finkelstein D, Bashford D (2009) Quantitative structure–activity relationship (QSAR) for a series of novel cannabinoid derivatives using descriptors derived from semi-empirical quantum-chemical calculations. Bioorg Med Chem 17(6):2598–2606
42. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146
43. Iyer M, Zheng T, Hopfinger AJ, Tseng YJ (2007) QSAR analyses of skin penetration enhancers. J Chem Inf Model 47(3):1130–1149