J. of the Chosun Natural Science
Vol. 4, No. 3 (2011) pp. 197 – 201

Toward Proper 3D-QSAR Datasets for Parameter Evaluation

Seung Joo Cho[1,2†]

Abstract

3D-QSAR techniques including CoMFA have been used a lot for more than two decades now. For now, the perspective of 3D-QSAR has been changed. The realization of gorge activity cliffs and higher chance correlation with many independent variables (IVs) has changed the requirements. Some suggested the benchmarking datasets for 3D-QSAR. However, were they still useful for right reasons? Here, we propose the requirement of any general purpose 3D-QSAR benchmarking datasets for lead optimization, especially for feasibility test of any IVs. Specifically, we summarize the conceptual requirements for an ideal settings for 3D-QSAR especially CoMFA.

Key words : 3D-QSAR, CoMFA, Independent Variables(IVs)

1. Introduction: Changing Perspective of Physical Reality of QSAR in CADD

When a molecular modeler develops a QSAR (quantitative structure activity relationship) model, especially for lead optimization purpose, one implicitly assumes that structure-activity landscape is very smooth like Mount Huji (Figure 1a). Here the height corresponds to the biological activity, which horizontal axes indicate independent variables. Therefore they are three dimensional simplifications of multi-dimensional reality. This kind of smoothness has been implicitly assumed if not precisely spoken. At the same time, one usually hopes a linear model with an application domain which is big enough to suggest better chemical modifications. However, because of too many recent failures of QSAR models, it is getting more accepted that activity landscape is more like Bryce Cannon (Figure 1b), maybe much more rugged than that, even for simple cases[1]. In addition it was demonstrated that even with a dataset of very simple structure variation, adding additional factor could change the landscape completely[2]. This implies that to the very limited extent a linear QSAR models can apply. Somehow, in the circumstances, stable models are difficult to obtain. Since the model we develop for lead optimization in QSAR is usually a simple model, at least a simple functional form, the application domain should be of very limited extent. Another difficulty against accurate model building comes from the discrepancy between a very high accuracy of energy calculation (1~2 kcal/mol) and the state of the art of accuracy (~ 5 kcal/mol)[3]. Therefore, it's like we are groping in the dark with no light. Thus, regarding drastic simplifications of most 3D-QSAR calculations, getting a reliable model with general 3D-QSAR methodology would be very difficult, if not impossible. If this roughness/uncertainty holds true in general, the consequences are far reaching. For example, finding suitable independent variables would be also challenging. The landscapes are clearly dependent on the independent variables applied. Johnson argued that since there are too many available independent variables, even in cases where there is no innate cause, chance correlation could be obtained quite frequently [4]. Or, there could be multiple models. Therefore, what is possible in QSAR modeling might be, a model interpretation for the dataset at hand, rather than prediction to test set compounds.

2. Current Data Sets Involving 3D-QSAR

Since 3D-QSAR in CADD is critically related to molecular docking, it is worthwhile to know how the docking dataset is constructed. One of the popular dataset is DUD (directory of useful decoys)[5]. The database

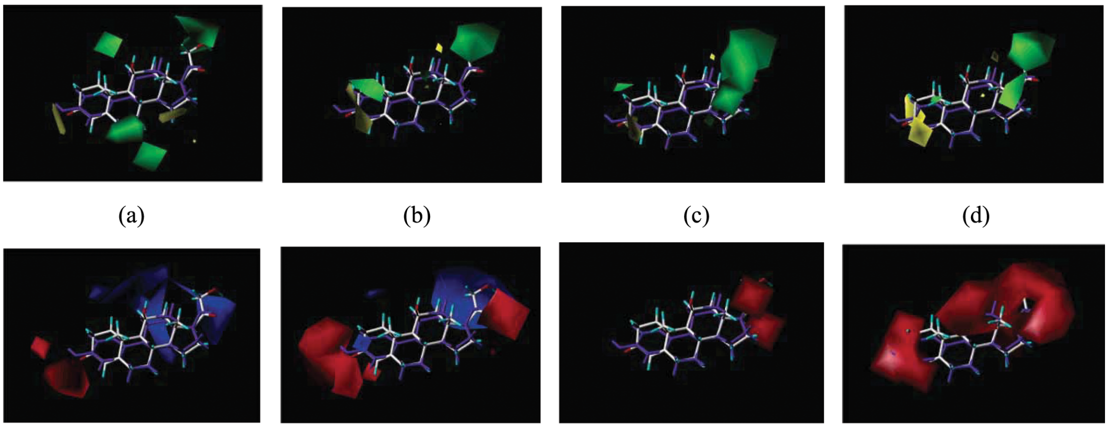
[1]Departments of Bio-New Drug Development
[2]Departments of Cellular · Molecular Medicine, College of Medicine, Chosun University, 375 Seosuk-dong, Dong-gu, Gwangju 501-759, Korea
†Corresponding author : chosj@chosun.ac.kr

(Received : July 2, 2011, Revised : August 16, 2011, Accepted : August 31, 2011)

(a) Mount Huji

(b) Bryce Cannon National Park

Fig. 1. Smooth vs Rugged landscapes.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 2. A lot of changes in CoMFA Models which show reasonable statistics.
(a), (b), (c) and (d) show steric contour models while (e), (f), (g) and (h) depict electrostatic models. Fields shown in green and yellow indicate regions where steric bulk is positively and negatively correlated, respectively, with an increase in activity. Fields shown in red and blue indicate regions where negative charge is positively and negatively correlated, respectively, with an increase in activity.

decoys resemble the physical the real ligands so that enrichment is not simply a separation of trivial physical features. The decoys nevertheless should be chemically distinct from the ligands so that they are likely to be non-binders. DUD provides a more real world-like test for virtual screening. Sutherland et al.[6] collected 10 data sets including steroid data set. They studied CoMFA, CoMSIA, EVA, HQSAR, PLS, back-propagation neural network, genetic function approximation with 2D, 2.5D, and 3D descriptors. HQSAR showed similar performance as well as CoMFA. What is important is that they could not find any correspondence between cross-validated and test set predictive accuracy for four sets. Therefore using designed test sets would be very important. Peterson et al.[7] examined above dataset and tested 10 different CoMFA settings. They found no correlation between q2 and r^2_{pred} for the datasets tested. And even though q2 value is high enough, r^2_{pred} could have values 0.0~0.6 which indicates much lower predictive ability. Also with many CoMFA models which were determined reliable for corticosteroid data sets, the CoMFA maps show a lot of variance. This might reflect the greater than expected roughness of activity landscape Mittal et al. collected 40 various datasets [8] based on popularity, diversity and sufficient sample sizes. This might be enough sample sizes. Since SMILES structures are given as a supplementary materials, QSAR study using 2D descriptors may be performed. However, for 3D-QSAR methods such as CoMFA modeling, many calculation settings should be specified to test. This would be the consequence of the Bryce cannon-like landscape is the narrowness of application domains. It would be very hard to construct a useful application domain. As a result, again, there was no correlation between external validation (r^2_{ext}) and internal validation (q^2).

3. Dividing Training/Test Sets

Golbraikh et al.[9] argued 10 years ago that even though validation is quite important issue and q^2 has been extensively used for this purpose, q^2 is not really sufficient for validation of a model's predictive ability. It was already known for 3D QSAR, there is no correlation between the high LOO q^2 and the high predictive ability of a 3D-QSAR model. They used two-dimensional (2D) molecular descriptors and *k* nearest neighbors (*k*NN) QSAR method for the analysis of several datasets. Again no correlation between the values of q^2 for the training set and predictive ability for the test set was found for any of the datasets. Thus, the high value of LOO q^2 is the necessary condition not the sufficient one for the model validation. They insisted that the external validation is critical to build a reliable QSAR model. Therefore there have been studies on selection of training and test sets for the development of predictive QSAR models. It is impossible to predict a biological activity with absolute certainty. Therefore, finding a proper application domain is a crucial to the model's validity. Although a single QSAR model could be acceptable, the "partial" training set should represent the model developed from the entire population.

Golbraikh et al.[10] suggest that external validation using rationally selected training/test sets provides a means to establish a reliable QSAR model. We propose several approaches to the division of experimental datasets into training and test sets. They formulated a set of general criteria for the evaluation of predictive power of QSAR models. Ideally, the division into the training and test set must satisfy the following three conditions: (1) Test set should be close to training set, i.e., all representative compound-points of the test set in the multidimensional descriptor space must be close to those of the training set. (2) Vice versa, i.e., all representative points of the training set must be close to those of the test set. (3) The training set should represent entire data set. Thomas et al.[11] could obtain good external validation statistics were obtained when training and test sets were selected based on K-means clusters of factor scores of the descriptor space along with/without the biological activity values. Thus, if one wishes to validate a QSAR model, the points of the test set must be close to the points of the training set in the multidimensional descriptor space. Based on the results of the division of the training/test sets, they proposed a division method based on K-means-cluster could be utilized for building predictive QSAR models.

4. Guidelines to a proper collection of 3D-QSAR data sets

There are several characteristics which are required to any proper datasets.

4.1. Reproducibility

This is often ignored when many groups collect any datasets, but the dataset should be collected in a reproducible manner for any people. Also it is desirable that the collected data sets should be accessible via internet in order to reproduce the same data without much difficulty.

4.2. Simplicity

The calculation settings need to be as simple and standardized as possible. For example, data sets need to be collected from enzyme binding assay with the same binding mode, not from the cell based assays which could possess complex interactions. One of the setting options that affect CoMFA models are alignment methods. Since for some targets, there is no x-ray structure. Therefore, to cover broad range of targets, ligand based alignment may be more useful. Other settings need to be used CoMFA default values. Then from this settings, one can observe the effect of any change of settings in a systematic way.

4.3. Interpretation

In the literature, there is not serious concern with the problem of interpretation. Whenever the q2 is higher than 0.5 then with the model, the interpretation has been performed with the CoMFA maps. However, the interaction of steric and electrostatic effects has never been studied and there could be a significant impact in this map. Separate modeling with each of the factor followed by direct interpretation would be a useful practice to ensure if the factor is really in play.

4.4 Predictivity

When the model has both internal/external predictive abilities, then we say this model is a good model with predictive ability. Validation through dividing training/ test sets would be essential for this. This is critically related to the issue of application domain. Test set would be just about the application domain in practice. To define an application domain would be very hard if not impossible. There seems to be no consensus which partition methods would be best. This could be dependent on the nature of data set. A generally acceptable method for this partitioning would be a very important breakthrough in this area.

5. Conclusion

QSAR is only suitable to study cause and effect relationships under one condition, there should exist a known linear relationship, i.e., linear relationship between independent variables and dependent variables is a prerequisite and can never be inferred from statistical results. So, the causality should be known before the experiments. The linearity assumption makes QSAR a quite delicate tool. Moreover, it is very difficult to prove linearity with statistical means. Nonlinear behavior of complex biological process cannot be ruled out in principle. This brings an important question. Is it possible to set up a benchmarking collection of data sets which will show high enough q^2 and r^2_{pred}? with clearly known mechanical reasons? For example, CoMFA has two sets of parameters, namely electrostatic and steric parameters. When a ligand binds to a receptor, it's very difficult to point out which one of the factors will be important quantitatively for the observed binding. In this rugged landscape, this seems to be an impossible task. However, parameters should be evaluated based on the correlation between activities and descriptors used in modeling. By clarifying the selection of methods and settings it will enable more researchers to make use of the valuable benefits that QSAR can bring into lead optimization-like research projects. An important step toward improving this situation is to make available a collection of data sets for any parameter validation. The requirements are (i) representative of CADD in lead optimization, (ii) diversity, (iii) sufficient sample size, (iv) mechanistic interpretation, (v) correlation between q^2 and r^2_{pred}.

Acknowledgements

This work was supported by Korea Research Foundation (KRF)of Korea government (MEST). Contract grant number: 2009-0076553.

Reference

[1] G. M. Maggiora, “On Outliers and Activity Cliffs: Why QSAR Often Disappoints” J. Chem. Inf. Model., Vol. 46, No. 4, 1535, 2006.

[2] C. G. Gadhe, G. Kothandan, and S. J. Cho “A Large Change of Activity Landscape: Evaluation of Partial Charge Schemes on the Mutagenicity of Mutagen X Analogs” J. Comput. Aided Mol. Des. Submitted.

[3] J. Manchester and R. Czerminski, “SAMFA: Simplifying Molecular Description for 3D-QSAR” J. Chem. Inf. Model., Vol. 48, p. 1167, 2009.

[4] S. R. Johnson, “The Trouble with QSAR (or How I Learned to Stop Worrying and Embrace Fallacy)” J. Chem. Inf. Model, Vol. 48, p. 25, 2008.

[5] N. Huang, B. K. Shoichet, and J. J. Irwin, “Benchmarking Sets for Molecular Docking”, J. Med. Chem. Vol. 49, p. 6789, 2006.

[6] J. J. Sutherland, L. A. O’Brien, and D. F. Weaver, “A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships”, Vol. 47, p. 5541, 2004.

[7] S. D. Peterson, W. Schaal, and A. Karlen, “Improved CoMFA Modeling by Optimization Settings” J. Chem. Inf. Model., Vol. 46, p. 355, 2006.

[8] R. R. Mittal, R. A. McKinnon, and M. J. Sorich, “Comparison Data Sets for Benchmarking QSAR Methodologies in Lead Optimization” J. Chem. Inf. Model., Vol. 49, p. 1810, 2009.

[9] A. Golbraikh and A. Tropsha, “Beware of q^2!”, J. Mol. Graph. Mod., Vol. 20, p. 269, 2002.

[10] A. Golbraikh, M. Shen, Z. Xiao, Y. D. Xiao, K. H. Lee, and A. Tropsha, “Rational selection of training and test sets for the development of validated QSAR models”, Vol. 17, p. 241-253, 2003.

[11] J. T. Lonard and K. Roy, “On Selection of Training and Test Sets for the Development of Predictive QSAR models”, QSAR Comb. Sci. Vol. 25, p. 235,