論 文 の 内 容 の 要 旨

論文題目　A Study on Chemical Structure Generation Based on Inverse Quantitative Structure-Property Relationship/Quantitative Structure-Activity Relationship

（定量的構造物性相関/定量的構造活性相関モデルの逆解析を利用した化学構造創出に関する研究）

氏　名　宮尾　知幸

1. Introduction

Quantitative structure-property relationship (QSPR) or quantitative structure-activity relationship (QSAR) is a way to find a quantitative relation between compounds and their corresponding property or activity in a statistical manner. Property or activity is usually numerical and, therefore, can be represented as an objective variable: y. To treat compound information numerically, a compound is usually translated into a set of chemical descriptors (**x**), which are abstract representations of a molecule. Therefore, a QSPR/QSAR model can be regarded as a regression model (y=f(**x**)) constructed with an experimental dataset. Inverse QSPR/QSAR-based molecular design is to generate chemical structures satisfying a specific y value through the backward analysis of a pre-constructed QSPR/QSAR model. In contrast to molecular design with QSPR/QSAR, such as virtual screening, inverse QSPR/QSAR analysis can generate chemical structures *de novo*. There have not been, however, methodologies for inverse QSPR/QSAR, which can be applied to practical applications: using various descriptors and nonlinear regression models, and considering applicability domains (ADs)[1]. ADs limit the chemical space in a way that, only inside ADs, predicted values produced by regression models should be trusted. ADs must be considered when applying QSPR/QSAR models to novel chemical structures.

2. Contributions of this thesis

Main contribution of this thesis is to develop a practical chemical structure generation system based on inverse QSPR/QSAR. In order to make the proposed system practical, several methodologies have been proposed and implemented by the author, namely:

1. To introduce a nonlinear regression methodology for capturing nonlinear relationship between **x** and y.
2. To develop a methodology for considering ADs with a Gaussian mixture model (GMM) as a probability density function (PDF). On the basis of the premise that inside the highly dense areas of training samples in chemical space reliability of predicted values by QSPR/QSAR models is high, the posterior PDF of **x** given a y value (p(**x**|y)) is expected to possess the degree of prediction reliability and the closeness to the y value.
3. To propose novel algorithms for chemical structure generation. Chemical structures are efficiently built by combining ring systems[2] and atom fragments by means of the canonical construction path method proposed by McKay[3].

4. To introduce and implement monotonous changing descriptors (MCDs) in the proposed inverse QSPR/QSAR system[4]. MCDs are descriptors whose values monotonically change by adding a building block to a growing structure. Since a wide range of descriptors can be categorized as a MCD, the construction of regression models with high predictability is expected.

Inverse QSPR/QSAR can be divided into two parts: obtaining **x** information from a y value, and constructing chemical structures from the **x** information. These parts are connected sequentially to form a system, which is the proposed chemical structure generation system based on inverse QSPR/QSAR analysis.

3. Structure generation

3.1 Structure generation algorithm

In order to recognize ring systems as graphs having fewer vertices than those of the ring systems, reduced colored graphs are introduced. A reduced colored graph has the same topology as that of the corresponding ring system. During structure generation, reduced graphs are used as elements instead of the original ring systems themselves. This treatment was expected to reduce the calculation cost including data storage. To construct structures by combing reduced graphs, the canonical construction path method proposed by McKay[3] is employed. Chemical structures can grow by adding a building block to them in every possible way without generating duplicate structures with the methodology.

3.2 Performance of structure generation

Efficiency of the proposed algorithm was compared with that of a simple fragment-combined-based generator developed by Arakawa *et al*[5]. It exhaustively combines building blocks until the number of used ones reaches a predetermined value. After the structure generation procedure, canonicalization operation and elimination of duplicate structures are conducted. As building blocks, 10 ring systems and 13 atom fragments were used (not shown here). The atom fragments were CH_3, CH_2, CH, C, NH_2, NH, N, OH, O, F, Cl, Br, and I. The number of combined building blocks was set from two to eight. For each number of building blocks, five trials were conducted in order to evaluate the calculation time statistically. The performance test was conducted on a Windows 10 personal computer with 3.33GHz Intel Xeon CPU and 16 GB RAM. For both generators, the relationship between calculation time and the number of generated structures seemed linear. For the fragment-combined-based generator, it took 1.39×10^{-3} to generate a structure, whereas the proposed generator took 3.83×10^{-6}. This result supports the efficiency of the proposed algorithm. It should be noted that both generators generated the same number of structures successfully.

4. Inverse QSPR/QSAR analysis (from y to x)

4.1 Proposed methodology

In inverse QSPR/QSAR analysis, retrieving **x** information from a y value is also important since it determines generation conditions for the structure generation part. The author once proposed to use $p(\mathbf{x}|y)$ for considering ADs. $p(\mathbf{x}|y)$ can be derived in a closed form solution only when regression model is MLR and the prior distribution is a GMM. In order to overcome this limitation, for making inverse QSPR/QSAR practical, cluster-wise MLR (cMLR) with GMMs (GMMs/cMLR)[6] is introduced.

4.2 AD evaluation with the aqueous solubility dataset

Aqueous solubility dataset[7] was used for evaluating $p(\mathbf{x}|y)$ in various aspects. The dataset consists of 1,290 compounds annotated with measured aqueous solubility (S) at 20-25 degrees Celsius [mol/L]. Objective variable y is the logarithm of S (logS). After descriptor calculation, remaining 1,154 molecules were randomly divided into 900 training samples and the 254 test samples. Descriptors were, molecular weight (MW), number of hydrogen bond donor and acceptor (HBD and HBA) based on the Lipinski's rule[8], number of rings (CIC), topological polar surface area (TPSA), and number of rotatable bonds (nBR). Both MLR and GMMs/cMLR models were constructed. Seven Gaussians formed a $p(\mathbf{x})$ based on Bayesian information criterion. Predictability of the models is shown on Table 1. GMMs/cMLR shows higher predictability than MLR does. By combining the GMM/cMLR model ($p(y|\mathbf{x})$) and $p(\mathbf{x})$, $p(\mathbf{x}|y)$s were derived for various y values.

Table 1 Predictability of MLR and GMMs/cMLR

	R^2	RMSE	R_{pred}^2	$RMSE_{pred}$
MLR	0.73	1.06	0.72	1.13
GMMs/cMLR	0.85	0.79	0.85	0.82

RMSE: root mean squared error. R^2: R-squared. $RMSE_{pred}$: RMSE for test data. R^2_{pred}: R^2 for test data

Posterior PDFs should have higher density in the area where the desired property value is expected. Furthermore, based on the premise of AD, posterior PDFs are expected to inherit the prior PDF feature of training data density. This inference is natural based on Bayesian probability. In this respect, p(**x**) and p(**x**|y) were compared with each other in various y values. p(**x**) of the training dataset is plotted against p(**x**|y) with y values of -6, -4, 1 and 2 in **Figure 1**.

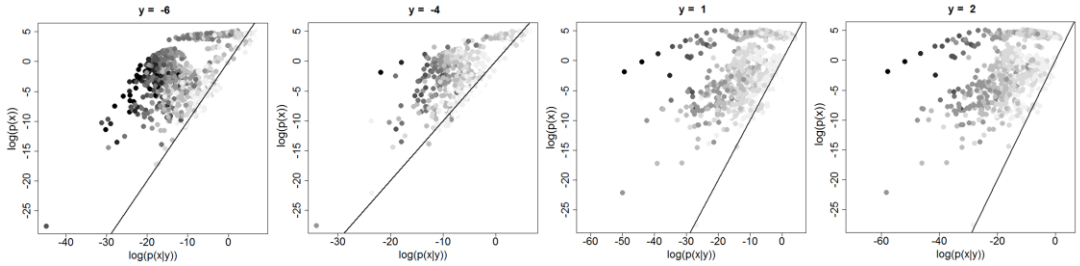

Figure 1 Logarithm of p(**x**) is plotted against logarithm of p(**x**|y) with various y values by GMMs/cMLR. Thick colors represent greater differences between measured y values and the target one; thin colors represent less differences.

From these plots, the former hypothesis was confirmed. The higher p(**x**|y) of a sample was, the lesser the absolute error between the measured and the target y became. Furthermore, it seemed that p(**x**|y) inherited the p(**x**) feature. Samples did not exceed the diagonal line radically even when their measured y values were close to the target one. These trends were also confirmed for test dataset. Therefore, p(**x**|y) could represent the likelihood that the coordinates exhibit the y value after taking ADs into account. In short, p(**x**|y) derived from p(**x**) and p(y|**x**) using the proposed methodology (i.e. GMMs/cMLR) had the preferable properties: inheriting p(**x**) and expressing the closeness degree to the target y value. It is, however, not straightforward to determine a proper threshold of p(**x**|y) for determining high dense regions since the scale of p(**x**|y) varies depending on y values.

5. Proposed system

An overview of the proposed chemical structure generation workflow is illustrated in **Figure 2**.

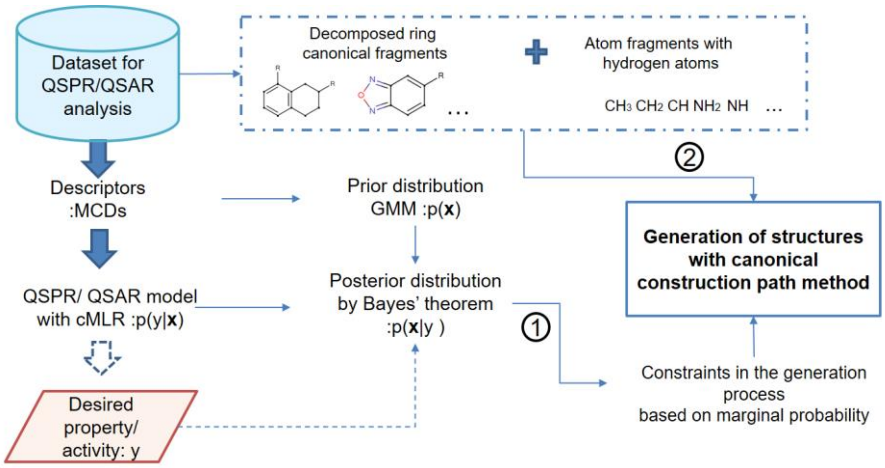

Figure 2. Overview of the proposed structure generation system workflow. MCDs: monotonous changing descriptors, cMLR: cluster-wise multiple linear regression, GMM: Gaussian mixture model.

Basically, ring systems are extracted from all the molecules in a dataset for constructing a QSPR/QSAR model. In addition to the ring systems, atom fragments are also used as building blocks for structure generation in order to generate diverse chemical structures (2). MCDs are used for constructing both a QSPR/QSAR model with cMLR and a prior PDF with a GMM. The input in inverse analysis is a specific y value. The posterior PDF of **x** given the y value (p(**x**|y)) can be derived as a closed-form solution by combining p(**x**) and p(y|**x**) as explained in the Inverse QSPR/QSAR analysis (from y to x) section. The **x** coordinates in descriptor space are determined based on p(**x**|y), followed by the transformation of the coordinates to constraints (1). To apply the proposed generation algorithm explained in the Structure generation section, constraints are set as the upper and lower bounds of MCDs, forming a hyper-rectangle. When determining the constraints based on p(**x**|y), the center of a Gaussian is focused upon. This implies that the ranges (i.e. constraints) for structure generation should be determined as narrowly as possible.

6. Conclusion

A chemical structure generation system based on inverse QSPR/QSAR analysis has been developed. Inverse QSPR/QSAR means that analyzing a pre-constructed QSPR/QSAR model inversely in order to obtain chemical structures exhibiting the property or the activity value that a designer expects. Against its simple definition, methodologies for inverse QSPR/QSAR was limited and hard to develop because of complicated mapping relations both between **x** and y, and between chemical structures and **x**. In this thesis, methodologies for tackling these challenges were proposed and demonstrated. In the structure generation part, several algorithms were proposed for producing chemical structures satisfying constraints. In the inverse QSPR/QSAR analysis part, which is obtaining **x** coordinates exhibiting a specific y value, GMMs/cMLR and using the posterior PDF were proposed for the enhancement of predictability and the consideration of ADs. Finally, these two methodologies were connected sequentially to form the proposed inverse QSPR/QSAR structure generation system.

7. References

(1) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicabilty Domain Estimation by Projection of the Training Set Descriptor Space: A Review. *Altern. Lab. Anim.* **2005**, *33*, 445–459.

(2) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(3) McKay, B. D. Isomorph-Free Exhaustive Generation. *J. Algorithms* **1998**, *26*, 306–324.

(4) Miyao, T.; Arakawa, M.; Funatsu, K. Exhaustive Structure Generation for Inverse-QSPR/QSAR. *Mol. Inform.* **2010**, 29, 111 –125.

(5) Arakawa, M.; Yamada, Y.; Funatsu, K. Development of the Computer Software. *J. Comput. Aided Chem.* **2005**, *6*, 90–96.

(6) Miyao, T.; Kaneko, H.; Funatsu, K. Inverse QSPR/QSAR Analysis for Chemical Structure Generation (from Y to X). *J. Chem. Inf. Model.* **2016**, *56*, 286–299.

(7) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266–275.

(8) Lipinski, C. A. Drug-like Properties and the Causes of Poor Solubility and Poor Permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235–249.