Boosting n-octanol/water Partition Coefficients Prediction with An Improved Gene Expression Programming Method

Cheng Yuan¹, Yuzhong Peng¹*, Chuyan Deng¹, Daoqing Gong¹ and Aiqing Cao¹

¹School of Computer & Information Engineering, Nanning Normal University, Nanning, Guangxi, 530024, China

*Corresponding author’s e-mail: jedison@163.com

Abstract. n-octanol/water partition coefficient (named logP) reflects the lipid solubility and aqueous solubility of the substance. Accurate and effective prediction of logP has great significance for drug development and monitoring human health, due to logP is related to the dissolution, absorption, distribution and transport of the drug in the human body. This study proposed an improved gene expression programming algorithm based on fuzzy control method with the feature of Morgan fingerprint to improve the logP prediction. Experimental results evaluated in terms of RMSE and MAE show that proposed method outperforms not only multicellular gene expression programming, but also the state-of-the-art methods including Back Propagation neural network, support vector regression, random forest regression and Gaussian process regression.

1. Introduction

The n-octanol/water partition coefficient named logP value is the ratio logarithm that is the partition coefficient of the substance in n-octanol and water [1]. logP describes the partition of matter in oil and water, which is regarded as a key role for molecular discovery activities in the fields of pesticides, materials science, food chemistry, and especially drug screening. The aqueous and lipid solubility are relating to the dissolution, absorption, distribution and transport of drugs in the body [2-3], therefore, logP is one of the key physicochemical parameters for determining the membrane permeability of drug-like compounds [4]. The n-octanol/water partition coefficient is also used for monitoring human health [5]. Many organic pollutants are enriched in adipose tissue ("oil") after entering the organism. Therefore, the bigger n-octanol/water partition coefficient is, the more containments are accumulated in fat that are hardly excreted in perspiration, urine, etc. [6]. In other words, it is similar to the extraction of organic matter, which is easier to extract by organic solvents and is not easier to stay in the water. Therefore, logP plays a vital role in the human metabolism and measure of organic solutes [7-9]. With the advent of more and more small-molecule data sets, it is urgent to develop an accurate and effective method for predicting logP on a large-scale.

Machine learning methods are widely used in the field of cheminformatics, such as the prediction of aqueous solubility [10], partition coefficient [11] and other physicochemical properties [12]. In addition, the linear regression is the simplest form of regression model. Regression models are used to predict continuous-value of different properties, including the physicochemical properties of the compounds. For x-values with multiple features, the linear method LASSO [13] is a better alternative as it estimates sparse coefficients. To nonlinear relationships, there are more sophisticated regression methods, such as random forest regression (RFR) [14], (stochastic) gradient tree boosting (GTB) [15], support vector regression (SVR) [16], neural network [17], nuclear ridge regression [18] and Gaussian
process [19]. Neural networks as well as RFR and SVR have recently been applied to predict aqueous solubilities [20]. Kernel ridge regression and Gaussian processes (GP) have been used in ML applications with quantum-mechanical data [21].

In this study, we developed a novel method for logP prediction. The method uses multicellular gene expression programming algorithm based on fuzzy control (MGEP-FC) with the feature of Morgan fingerprint to model and predict logP. We conducted experiment comparing with not only the state-of-the-art methods including SVR, RFR, GP and BP, but also multicellular gene expression programming algorithm (MCGEP) [22], for evaluating the proposed method performance in terms of RMSE and MAE. The experimental result shows that the proposed method outperforms the other comparative methods.

2. multicellular gene expression programming algorithm based on fuzzy control

GEP is a new heuristic random search algorithm that combines the advantages of GA and GP. It inherits the simple and fast character of GA fixed-length linear coding, and the characteristics of GP tree structure flexibility [23-24]. The GEP chromosome is represented by the fixed-length string and contain one or more genes. Each gene is composed of a head and a tail, which the head consists of the function set \( F \) (consists of operators or other elementary functions) and the terminal set \( T \) (consists of variable or constant). The tail is only composed by the terminal set \( T \). The relationship between the head \( h \) and the tail \( e \) is:

\[
e = h \times (n - 1) + 1
\]  
(1)

where \( n \) is the number of arguments of the function with the most arguments.

The MCGEP algorithm is a new algorithm which is proposed by introducing the homologous genes and cellular systems [21]. The homologous gene is different from the ordinary gene. Because its tail character represents each ordinary gene, and the head operator links the results of ordinary genes, which expands the individual expression space and search space. System coding is shown in Figure 1.

![Figure 1. Structure of MCGEP cellular system.](image)

As shown in Figure 1, the number in the first row is only used to indicate the gene position of each gene. The second line is the genotype coding string of the chromosome and each gene is separated by a red line. Lowercase letters represent corresponding integers, uppercase letters indicate the position in the constant array, the black part is a homologous gene.

In our previous work, based on MCGEP, we have proposed a function optimization algorithm called MGEP-FC that has a fast convergence and powerful global optimization capability [25]. It mainly adjusts the crossover rate, mutation rate and constant-set mutation rate of genetic operations according to the degree of concentration of individual fitness values in the population. By constructing the fuzzy membership function, it will calculate the crossover rate, the mutation rate and the constant-set mutation rate of the next generation in the population. The MGEP-FC algorithm calculate the rate \( D \) of the current population optimal fitness \( f_{best} \) and the average fitness \( f_{ave} \). The \( D \) value is determined by the following formula:

\[
D = \begin{cases} 
\frac{f_{min}}{f_{ave}}, & f_{best} < f_{ave} \\
\frac{f_{ave}}{f_{max}}, & f_{best} > f_{ave}
\end{cases}
\]  
(2)

When solving the minimum value of the function, the diversity of the population is determined by equation (2). Conversely, to obtain the maximum value of the function, equation (3) is used.
The MGEP-FC algorithm designs three fuzzy controllers to control the crossover rate, the mutation rate, and the constant-set mutation rate. The five fuzzy membership functions are described by using the five fuzzy linguistic variables that are \{LE, L, M, H, HE\}. The membership functions are constructed by using trigonometric and trapezoidal membership function, which are shown in Figure 2.

When the population diversity is lower, both the mutation rate of the algorithm and the diversity of the population are enhanced, so that the algorithm can jump out of the local optimum. The crossover rate of the algorithm is increased and the mutation rate is reduced to speed up the convergence of the algorithm, while the population diversity is good. According to this purpose, the fuzzy rules are shown as Table 1 and Table 2.

### Table 1. Fuzzy control rules of cross and mutation rate

| Population diversity (d) | LE | L | M | H | HE |
|--------------------------|----|---|---|---|----|
| **Cross rate**           | HE | H | M | L | LE |
| **Mutation rate**        | LE | L | M | H | HE |

### Table 2. Fuzzy control rules of constant-set mutation rate

| Number of iterations (N) | Population diversity (d) | LE | L | M | H | HE |
|--------------------------|---------------------------|----|---|---|---|----|
| LE                       | LE                        | LE | L | L | M |
| L                        | LE                        | L  | L | M | H |
| M                        | L                         | L  | M | H | H |
| H                        | L                         | M  | H | H | HE|
| HE                       | M                         | M  | H | HE| HE|

In this paper, based on MGEP-FC, we developed a novel regression analysis method called FCMGEP2logP for predicting \(n\)-octanol/water partition coefficient. In FCMGEP2logP, we mainly changed the function optimization process of MGEP-FC to a model training and testing process using...
samples, and used RMSE and MAE as the fitness function of FCMGEP2logP. This algorithm procedure of FCMGEP2logP is summarized as follows:

**Algorithm: FCMGEP2logP**

| Input: Dataset of Molecular Morgan fingerprint and logP data, parameters of FCMGEP2logP including population size N, Terminator set T, function set F, (Homeotic) gene head length and Maximum number of iterations. |
|---|
| **Output:** Prediction values of logP. |
| **Step1:** Use PCA to extract the feature of the Molecular Morgan fingerprints; |
| **Step2:** Encode the individual, and construct initial population; |
| **Step3:** Evaluate individual's fitness value; |
| **Step4:** Calculate the population diversity D; |
| **Step5:** Adjust the real constant-set mutation rate, mutation rate and cross rate using the fuzzy control method; |
| **Step6:** Execute multiple parallel genetic operations; |
| **Step7:** Evaluate new individual fitness values; |
| **Step8:** Construct a temporary population with all parent individuals and children individuals; |
| **Step9:** Select individuals from the temporary population to form the next population using the tournament and elite retention strategies; |
| **Step10:** If the termination condition met, select the optimal individual, otherwise return to step 3; |
| **Step11:** Translate the optimal individual into the prediction model; |
| **Step12:** Use the model and testing datasets to predict logP values. |

3. Experimental results and discussion

3.1. Experimental Data and Parameter Settings

In this study, FreeSolv database version 0.31 [26] is used for evaluating performance of our method comparing with four state-of-the-art prediction methods, including BP, SVM, RFR and GP. 75% data samples of the dataset are used as the training set, then the rest data samples are used as the test set. In our experiment, we firstly calculate the Morgan fingerprint by RDKit (an open source chemoinformatic tool) for each molecule sample. Then, we used the PCA method to transform the Morgan fingerprint into one-dimensional linear uncorrelated eigenvalues by orthogonal transformation. The eigenvalues are the input data of each prediction method.

The four comparative prediction methods were implemented by the python toolkit Scikit-learn with the default parameters. The collective parameters of MCGEP and FCMGEP2logP algorithm are set as the follows: Population size, Common gene head length and Homeotic gene head length are set for 100, 15 and 4, respectively; Tournament scale =0.01* Population size, Function set = {+, -, *, /, Q, S, O}. The Mutation rate, Cross rate and constant set mutation rate of MCGEP are set for 0.15, 0.3 and 0.01, respectively, while those of FCMGEP2logP are adjusted adaptively using the initial value 0.1.

The performance of the algorithm was evaluated by calculating the RMSE and MAE with a 10-fold cross validation approach. The average and the standard deviation over the 10 RMSE or MAE values are reported.

3.2. Prediction results and discussion

The experimental results on different Morgan fingerprint are shown in Tables 3. It can be seen from the data in the result table that: (1) The proposed method has a better performance of the fitting and prediction comparing with five state-of-the-art QSPR methods both in terms of two evaluation metrics on four types of different Morgan fingerprint. The best performance on the 256-bit Morgan fingerprint; (2) In terms of MAE, FCMGEP2logP improved by 10.07%, 8.27%, 22.15%, 15.49% and 0.77% comparing with BP, SVR, RF and GP, respectively. In terms of RMSE, FCMGEP2logP improved by 13.64%, 12.77%, 27.78% and 0.71% comparing with BP, SVR, RF and GP, respectively; (3) The RMSE of some algorithms such as SVR, RFR will increase with the increase of Morgan fingerprint bit, which means that these algorithms are not suitable for fitting high-dimensional data. Because the Morgan fingerprint is a one-dimensional array, the higher for the dimension, the more sparsed for the
data, the more dispersed for the signatures, the worse for the eigenvalue obtained by PCA processing. The FCMGEP2logP can lighten the error, due to its strong expression and globle optimum-search ability, which can randomly search all possible solutions without directionality, and evolve towards the optimal direction.

| Bit | SVR  | GP   | RFR  | BP   | MCGEP | FCMGEP2logP |
|-----|------|------|------|------|-------|-------------|
| 32  | 1.439| 1.375| 1.616| 1.370| 1.379  | 1.361       |
| 64  | 1.449| 1.369| 1.612| 1.367| 1.380  | 1.357       |
| 128 | 1.461| 1.332| 1.706| 1.476| 1.322  | 1.313       |
| 256 | 1.461| 1.276| 1.765| 1.476| 1.313  | 1.275       |
|     |      |      |      |      |       |             |
| 32  | 1.024| 1.007| 1.158| 1.005| 1.010  | 0.982       |
| 64  | 1.042| 0.999| 1.106| 1.001| 1.016  | 0.986       |
| 128 | 1.040| 0.992| 1.291| 1.061| 0.999  | 0.981       |
| 256 | 1.040| 0.956| 1.226| 1.061| 0.990  | 0.954       |

4. Conclusions
The n-octanol/water partition coefficient prediction method based on FCMGEP2logP algorithm is proposed in this paper. Compared with other five state-of-the-art methods, the proposed algorithm achieved better prediction performance. Applying FCMGEP2logP to build QSPR model with Morgan fingerprint is an effective and promising methodology for molecular logP prediction on a large scale. This provides more prospects for the application of the FCMGEP2logP algorithm. In the future, we hope to explore the possibility of using FCMGEP2logP algorithm to predict more other properties of molecule well.

Acknowledgments
This work was supported by the Natural Science Foundation of Guangxi Zhuang Autonomous Region (Grants No.2017GXNSFAA198228), and the BAGUI Scholar Program of Guangxi Zhuang Autonomous Region of China.

References
[1] Long, X X, Niu, J F, Shi, S Q. (2006) Research on quantitative structure-property relationships for n-octanol/water partition coefficients of phthalic acid esters[J]. Huan jing ke xue= Huanjing kexue., 11: 2318-2322.
[2] Kundi Varun, Ho Junming. (2019) Predicting Octanol-Water Partition Coefficients: Are Quantum Mechanical Implicit Solvent Models Better than Empirical Fragment-Based Methods? The journal of physical chemistry. B., 31: 6081-6822.
[3] Mannhold, R., Poda, G. I., Ostermann, C., Tetko, I. V. (2009) Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of LogP Methods on More Than 96,000 Compounds. J. Pharm. Sci., 98: 861–893.
[4] Daina, A., Michielin, O., Zoete, V. (2014) iLogP: A Simple, Robust, and Efficient Description of n-Octanol/Water Partition Coefficient for Drug Design Using the GB/SA Approach. J. Chem. Inf. Model., 54: 3284–3301.
[5] Lipinski, C. A., Lombardo, F., Dominy, B. W., Feeney, P. J. (2001) Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. Adv. Drug Delivery Rev., 46: 3–26.
[6] Bennion, B. J., Be, N. A., Mcnerney, M. W., Lao, V., Carlson, E. M., Valdez, C. A., Malfatti, M. A., Enright, H. A., Nguyen, T. H., Lightstone, F. C., Carpenter, T. S. (2017) Predicting a
Drug’s Membrane Permeability: A Computational Model Validated With in Vitro Permeability Assay Data. J. Phys. Chem. B., 121: 5228–5237.

[7] Ogata, K., Hatakeyama, M., Nakamura, S. (2018) Effect of Atomic Charges on Octanol-Water Partition Coefficient Using Alchemical Free Energy Calculation. Molecules., 23: 425.

[8] Genheden, S. (2017) Solvation Free Energies and Partition Coefficients with the Coarse-Grained and Hybrid All-Atom/CoarseGrained Martini Models. J. Comput. Aided Mol. Des., 31: 867–876.

[9] Bannan, C. C., Calabró, G., Kyu, D. Y., Mobley, D. L. (2016) Calculating partition coefficients of small molecules in octanol/water and cyclohexane/water. Journal of chemical theory and computation, 12: 4015-4024.

[10] Wang, J., Hou, T. (2011) Recent Advances on Aqueous Solubility Prediction. J. Comb. Chem. High Throughput Screening., 14: 328–338.

[11] Eros, D., Kövesdi, I., Orfi, L., Takacs-Novák, K., Acsády, G., Ke´ri, G. (2002) Reliability of logP Predictions Based on Calculated Molecular Descriptors: A Critical Review. Curr. Med. Chem., 9: 1819–1829.

[12] Hughes, L. D., Palmer, D. S., Nigsch, F., Mitchell, J. B. O. (2008) Why are Some Properties More Difficult to Predict than Others? A Study of QSAR Models of Solubility, Melting Point and log P. J. Chem. Inf. Model., 48: 220–232.

[13] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. J. Royal Stat. Soc. B., 58: 267–288.

[14] Breiman, L. (2001) Random forests. Machine Learning., 45: 5–32.

[15] Friedman, J. H., Hastie, T., Tibshirani, R. (2001) The Elements of Statistical Learning. Springer, Berlin. Vol. 1.

[16] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., Vapnik, V. (1997) Support Vector Regression Machines. In: Advances in Neural Information Processing Systems. pp. 155–161.

[17] Specht, D. F. (1991) A General Regression Neural Network. IEEE Transactions on Neural Networks., 2: 568–576.

[18] Hastie, T., Tibshirani, R., Friedman, J. H. (2001) Springer Series in Statistics. Springer, New York.

[19] Rasmussen, C. E., Williams, C. K. I. (2006) Gaussian Processes for Machine Learning. MIT Press: Cambridge, MA.

[20] Lusci, A., Pollastri, G., Baldi, P. (2013) Deep Architecture and Deep Learning in Cheminformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. J. Chem. Inf. Model., 53: 1563–1575

[21] Rupp, M., Tkatchenko, A., Müller, K.-R., von Lilienfeld, O. A. (2012) Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. Phys. Rev. Lett., 108: 058301.

[22] Peng, Y.Z., Yuan C.A., Chen W.J., et al. (2010) Multicellular gene expression programming algorithm for function optimization. Control Theory & Applications., 27:1585-1589.

[23] Ferreira C. (2001) Gene expression programming:a new adaptive algorithm for solving problems.Complex Systems., 13: 87-129.

[24] Zhong J, Feng L, Ong Y S. (2017) Gene Expression Programming: A Survey. IEEE Computational Intelligence Magazine., 12: 54-72.

[25] Deng C.Y., Peng, Y.Z., Li, H., Gong, D.Q., Zhang, H., Liu, Z. (2019) A self-adaptive multicellular gep algorithm based on fuzzy control for function optimization. arXiv:2634575.

[26] Mobley, D. L., Guthrie, J. P. (2014) FreeSolv: A Database of Experimental and Calculated Hydration Free Energies, with Input Files. J. Comput. -Aided Mol. Des., 28: 711–720.