2008－BIO－12 (6)
2008／3／4

化合物活性予測のためのTanimoto係数とRandom Forestの Proximity Measureの組合せ手法

河村 元[†] 瀬尾茂人[†]
竹中要一[†] 松田秀雄[†]

創薬プロセスでは化合物の化学的、生物学的な活性情報を用いた薬物探索が非常に重要である．化合物の活性を見つけるための構造類似性検索は、化合物の構造の有無をビットで表わしたフィンガープリントとTanimoto係数を用いた化合物類似尺度を用いて評価されることが多い．しかしながら、実際の化合物探索では少数の教師データを用いて精度を向上させながら大量のデータから活性化合物を見つけ出す手法が重要になってくる．そこで本研究では従来のTanimoto係数とRandam Forestの Proximity Measureを用いた化合物類似尺度を線形判別分析によって組み合わせる評価法を提案する．特に、proximity MeasureとTanimoto係数は学習手法と非学習手法という本質的に異った評価方法に基づいているので、これらの組合せによって活性化合物予測の精度が向上することが期待される．この手法を化合物データベースのいくつかのデータセットにおいて評価する．これらの結果から、提案手法が化合物活性の探索において有効であることが示される．

A Combination Method of the Tanimoto Coefficient and Proximity Measure of Random Forest for Compound Activity Prediction

Gen Kawamura,[†] Shigeto Seno,[†] Yoichi Takenaka[†]
and Hideo Matsuda[†]

Chemical and biological activities of compounds provide valuable information for discovering new drugs. Since the number of compounds that are known to have some activities of a biological class is small in the drug discovery process, the accuracy of the prediction should be increased in databases that have a large number of un-annotated compounds and a small number of annotated compounds of the biological activity. In this paper, we propose a new similarity scoring method composed of a combination of the Tanimoto coefficient and the proximity measure of random forest. The score contains two properties that are derived from unsupervised and supervised methods for predicting active compounds. Thus, the proposed method is expected to indicate compounds that have accurate activities. By evaluating the performance of the prediction compared with the two scores of the Tanimoto coefficient and the proximity measure, we demonstrate that the prediction result of the proposed scoring method is better than those of the two methods by using the Linear Discriminant Analysis (LDA) method. It is also shown that the proposed method can identify active compounds in datasets including several un-annotated compounds.

1. Introduction

A compound similarity and screening method have to meet important criteria in order to be used in current drug discovery and development.[1)]Specifically, the completion of the human genome project has a serious impact on the drug discovery process. As a consequence of its completion, the targets of a particular gene family have become available, and genomics methods are being developed to identify protein targets for novel drug candidates.To identify these targets, systematic exploration of selected target families, without prior restriction to a specific therapeutic area, appears to be a promising method to improve the ligand identification process in drug discovery.

On the other hand, evaluating the structural similarity, several well-known methods in statistics and machine learning algorithms have been ap-

† 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University

plied. All of these methods (e.g., artificial neural networks[2], partial least squares[3] and support vector machine[4]) have many successful merits in the structural similarity and screening methods.

However, there are several problems with accurate prediction that arise from the requirement in the compound structural similarity searching. One of the problems is the number of biologically annotated compounds is insufficient compared with the total number of compounds. In fact, although the amount of compound data is growing rapidly, the number of newly biological annotated compounds has not increased quickly. Such databases contain enormous numbers of un-annotated compounds and few of the annotated compounds of the biological activity.

In this paper, we propose a similarity searching and screening method to estimate some scores and distributions of variance by means of measures between the Tanimoto coefficient and proximity measure[5] and a method to combine the Tanimoto coefficient and the proximity measure in order to improve prediction accuracy. Here, the proximity measure is a new ensemble method called random forest in machine learning algorithms to measure the similarity with high-dimensional data by using decision trees.[6] Applying this method to similarity search, we can obtain efficient performance for searching compounds in some activities, without re-optimization of the fingerprint.

2. Method

In this section we present the proposed method, which is based on the MACCS key, the Tanimoto coefficient, random forest and proximity measures. In addition, we present a Linear Discriminant Analysis (LDA) to evaluate the prediction accuracy and to combine the scores between the Tanimoto coefficient and the proximity measure.

2.1 Input variables

In the present study, we use the MACCS key, a fingerprint proposed by MDL, as the input variable of feature quantity of the compound structure. Fragments of chemical structures can be coded in binary keys, which are presented as sequences of 0s and 1s (bitstrings). 0 represents a fragment that does not exist in the structure; otherwise, the bit is 1, which indicates that the fragment exists. Specifically, this characteristic structure sequence, called the fingerprint, of the MACCS key has a length of 166 keysets.[7][8]

2.2 Classifier methods

We consider two classifier methods, the Tanimoto coefficient and proximity measures, to evaluate compound similarity using the MACCS key.

In order to measure the similarity between two compounds using the above described fingerprint, a number of similarity measures have been proposed. We consider a widely used similarity measure called the Tanimoto coefficient, which is defined by

$$s = c/(a + b - c) \tag{1}$$

where a is the number of 1s of the fingerprint of compound A, b is the number of 1s of that of compound B, and c is the number of 1s common to both A and B.[9] In a similarity search using this measure of the fragments that are represented by fingerprints, the compounds in the database are aggregated by biological activities, and it is thus appropriate to select similar compounds in data comparison of the coefficient.

In machine learning, random forest is a classifier that consists of several decision trees and outputs the class, which is the vote of the classes output by individual trees. This method combines Breiman's bagging[10] concept and Ho's random subspace method[11] to construct a collection of decision trees.[5]

Usually, in the study of Quantitative Structure-Activity Relationships (QSAR)[6], random forest consists of B trees $\{T_1, ..., T_B\}$. For compound activity, a set of their class labels is

$$Y = \{C_l \mid l = 1, \ldots, m\} \tag{2}$$

where C_l is a class label, and m is the total number of the classes. Each compound has a variable $X = \{x_1, ..., x_p\}$ which is a p-dimensional vector of compound descriptors or fingerprints associated with their structure. Here, we consider the training procedure of random forest for given data

$$D = \{(X_1, Y_1), ..., (X_n, Y_n)\}$$

where $X_i, i = 1, \ldots, n$, is a p-dimensional vector and Y_i is a class label. For above data D, the training procedure is as follows:

(1) Each tree is grown by bootstrap sampling. Each tree of size n is randomly drawn from the original data of n points and returns.

(2) For each bootstrap sample, the decision trees in the random forest are grown by the CART algorithm[12] to full length and are not pruned back. At each node of a tree, the random forest algorithm randomly selects m_{try} descriptors or fingerprints as input variables, and uses them to choose the best possible split. Generally, this algorithm is sufficiently robust for the selection of the number m_{try}, whose value is usually chosen as the square root of the total number of variables.

(3) The number of trees in the forest is grown until achieving a low error rate of convergence.

A bth desicion tree T_b for an compound with fingerprint of X outputs a class label $\hat{Y}_b(X) \in Y$ as its prediction. Thus, the ensemble of trees outputs the class labels $\{\hat{Y}_1(X), ..., \hat{Y}_B(X)\}$. The outputs of all trees are aggregated to decide one final prediction, $\hat{Y}$. For simple classification problems, $\hat{Y}$ is a class label predicted by the majority of trees. This voting rule is given by

$$\hat{Y} = \arg\max_{y \in Y} \sum_{b=1}^{B} I(\hat{Y}_b(X),\ y) \qquad (3)$$

where I is the following indicator function: $I(\kappa_1, \kappa_2) = 1$ if $\kappa_1 = \kappa_2$, and 0 otherwise.

In addition, in our classification analysis, we use the proximity measure of the above classifier trees to predict the similarity between two compounds in the fingerprint space. For the estimation of two compounds for the evaluation of the similarity by which to classify the assigned labels as each class, the proximity measure is defined as the probability of assigning two compounds to the same node of the ensemble trees. Although general researchers may be interested in the random forest voting classifier in order to determine the tree that is most relevant to the activity of interest, some studies have reported that the proximity measure of a random forest can be calculated between any pair of compounds in clustering analysis.[6] Given two compounds that have the variables X_1 and X_2, the proximity measure $\hat{p}$ is

$$\hat{p} = \frac{1}{B} \sum_{b=1}^{B} I(\hat{Y}_b(X_1),\ \hat{Y}_b(X_2)) \qquad (4)$$

More specifically, this measure of proximity has an advantage. This proximity measure is supervised because the proximity measure of random forest is created by the compounds depending on each dataset and database.

2.3 Combination of scores

To cope with the problems associated with the Tanimoto coefficient and the proximity measure of random forest, we propose a new similarity scoring system that considers their combination using Linear Discriminant Analysis (LDA). The LDA easily handles cases in which the class frequencies are unequal and their performances have been examined by randomly generated situations.[13] Given the score distributions of the Tanimoto coefficient and the proximity measure, we introduce the variable F_i in order to make the discriminant model.

$$F_i = \sum_{j=1}^{k} w_j Z_{ij} \quad (i = 1 \ldots n) \qquad (5)$$

w_j is the weight variable for the variable Z_{ij}, which is normalized by the original scores x_{ij}. x_{ij} denotes the classification score of the ith group on the jth explaining variable. Z_{ij} is given as

$$Z_{ij} = \frac{x_{ij} - M_j}{\sigma_j} \qquad (6)$$

$$(i = 1 \ldots n,\ j = 1 \ldots k)$$

where σ_j is the standard deviation, x_{ij} is the ith classification score for the respective case for the jth explaining variable, and M_j is the mean of variable x_{ij} for the jth variable. This method maximizes the ratio of the class variance in this specific data set to the class variance in any particular data set and guarantees maximal separatability from the distributions of several variables of the class. In this study, in order to increase the hit rate of similarity search, we estimate this method as a combination of two distributions in order to combine the scores of the Tanimoto coefficient and the proximity measure from Eqs.(1), (4), and (5).

$$F_i^{'}(tp) = w_s Z_{i\,s}^{'}(tp) + w_{\hat{p}} Z_{i\,\hat{p}}^{'}(tp) \qquad (7)$$

$$F_i^{'}(fp) = w_s Z_{i\,s}^{'}(fp) + w_{\hat{p}} Z_{i\,\hat{p}}^{'}(fp) \qquad (8)$$

Table 1 Classes and the number of Training sets

MDDR activity class	no. in class
Dopamine (D1) Antagonist	180
Dopamine (D3) Antagonist	280
Dopamine (D4) Antagonist	674
Estrogen	257
Estrogen Receptor Modulator	210
Antiestrogen	297
randomly selected	1000

For distributions of the true positives (tp) and the false positives (fp), $Z'_{i\,s}$ and $Z'_{i\,\hat{p}}$ are represented by Eq.(6).

$$Z'_{i\,s}(tp) = \frac{s_i - M_s}{\sigma_s} \quad (i \in tp) \tag{9}$$

$$Z'_{i\,s}(fp) = \frac{s_i - M_s}{\sigma_s} \quad (i \in fp) \tag{10}$$

$$Z'_{i\,\hat{p}}(tp) = \frac{\hat{p}_i - M_{\hat{p}}}{\sigma_{\hat{p}}} \quad (i \in tp) \tag{11}$$

$$Z'_{i\,\hat{p}}(fp) = \frac{\hat{p}_i - M_{\hat{p}}}{\sigma_{\hat{p}}} \quad (i \in fp) \tag{12}$$

Here, we can obtain the discriminant model F'_{tp} and F'_{fp} from the above equations. LDA determines the appropriate distribution functions F'_{tp} and F'_{fp} to combine each score of the Tanimoto coefficient and the proximity measure depending on the true positives and false positives from the base value F'_0.

$$F'_0 = \frac{(M_{F'_{tp}} + M_{F'_{fp}})}{2} \tag{13}$$

where $M_{F_{tp}}$ and $M_{F_{fp}}$ are the mean values of two distributions, F_{tp} and F_{fp}. The distribution function F_i can provide a classifier, which is classified as the base value F_0, that is expected to increase the accuracy of predicting the targets. The results are presented in Section 3.

3. Results and Discussion

In this study, we used R, an open source statistical computing software from the R project for Statistical Computing, to perform data analysis.[14]

3.1 Data set for experiment

The compounds in the MDDR2004.2 database[15] were sampled as data sets from the activity class associated with the target protein. Here we used Estrogen and Dopamine classes which have known activity as the target proteins of Estrogen and Dopamine receptors, respectively (shown in Table 1).

All compounds of these data sets have activity classes that were selected as the target receptor. Thus, the other data, with the exception of the reference class, is randomly selected and is designated as belonging to the "other" class. The former and latter data sets were regarded as positive examples and negative examples for predicting activity in our method, respectively. These data sets were merged into a single set, and randomly split into two subsets as experimental data. The first half served as a candidate data set for similarity searching and the second half was used to form a reference set (training data).

3.2 Classifier

First, we discuss a similarity measure of the Tanimoto coefficient, so as to provide a cooperative line of performance for the similarity measure based on simple compound structural distance. Also, we show the results of proximity measure to consider different points between the Tanimoto coefficient and the proximity measure.

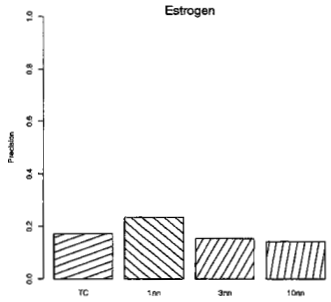
Figures 1 and 2 show the results for precision when the activity classes were predicted using each classifier.

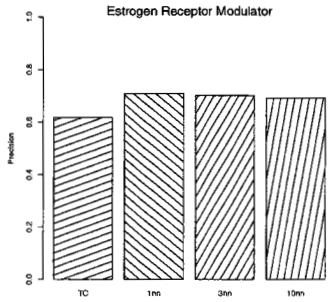
The most frequently used and basic measure for information retrieval effectiveness is precision. Precision is the fraction of the retrieved compounds that are relevant to successfully retrieval. Precision is usually measured as the ratio between the true positive rate predicted and the true positive rate of all of the predictions of each classifier.

$$Precision = \frac{tp}{tp + fp} \tag{14}$$

If all of the predicted classes are correct, this measurement can retrieve the compounds as a perfect classifier without any mistakes.

Based on this data, the proximity measures exhibit comparable or better precision than the Tanimoto coefficient. As mentioned previously, one focus of the present study for classification in drug discovery is a method by which to improve the true positive rate of similarity search by deselecting several un-annotated compounds. The reason for the slightly higher degrees of precision of classes remains unclear. However, in the previous study[6], the proximity measure was already mentioned that it could show good performance of the hierarchical





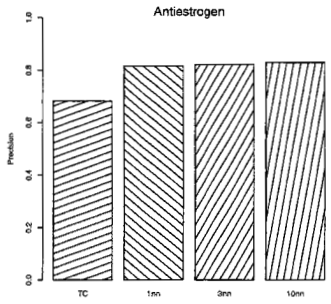
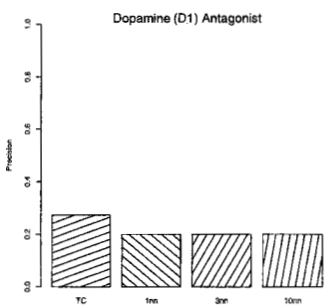
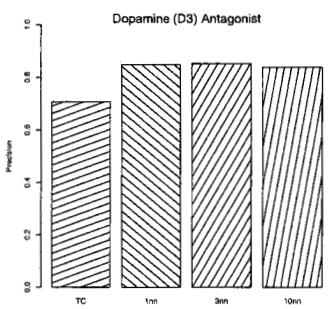


Figure 1 Precision of proximity measure and Tanimoto coefficient for Estrogen. The bars labeled TC, 1nn, 3nn and 10nn denote the precision of the Tanimoto coefficient and proximity measures for 1nn, 3nn, and 10nn, respectively.





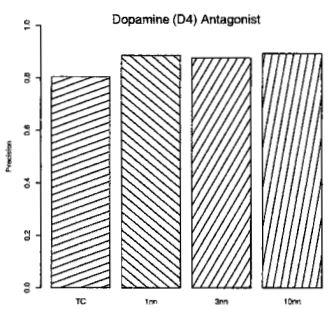


Figure 2 Precision of proximity measure and Tanimoto coefficient for Dopamine. Abbreviations are same as Figure 1.

clustering. The Tanimoto coefficient, on the other hand, does not take into account the discriminating power and treat all fingerprints equally, which resulted in lower performance. Also, from Figs. 1 and 2, the proximity measure can predict each class with accuracy. This investigation for the precision scores of the proximity measure and the Tanimoto coefficient would show that the proximity measure corresponds to the general similarity distance in the rate of score ranking.

Owing to its good performance in the proximity measure, these data provide not only a supervised quantitative value for the degree of resemblance between two compounds, but also their alignment without parameter tuning.

In addition, these results include the possibility of the accuracy rate of activities discovered when the two rankings of the Tanimoto coefficient and proximity measure are fused. The precision results

show that the fusion-generated hit-lists might contain more accurate activities than either of the only candidate rankings by using the Tanimoto coefficient.

3.3 Combination of scores

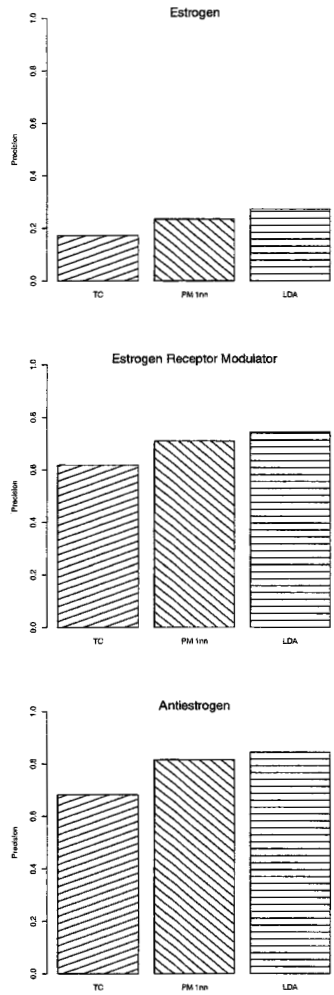


Figure 3 Precision of the proximity measure and the Tanimoto coefficient for Estrogen. The bars labeled TC, PM 1nn, and LDA denote the precision of the Tanimoto coefficient, the proximity measure for 1nn, and their combination (LDA), respectively.

Figures 3 and 4 show the results of precision plots before and after the consideration of the LDA for

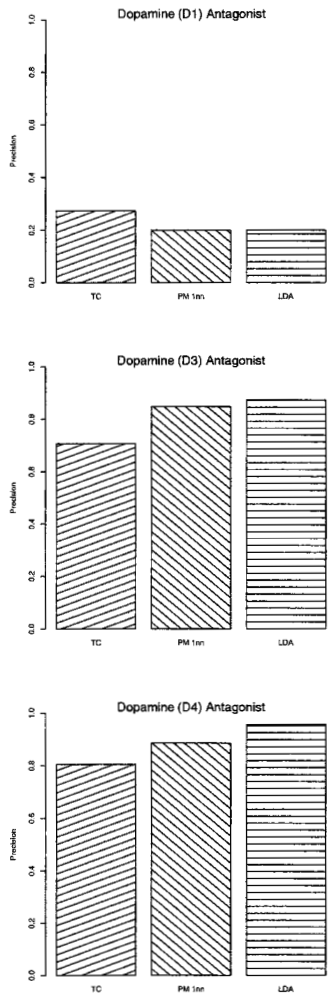


Figure 4 Precision of the proximity measure and the Tanimoto coefficient for Dopamine. Abbreviations are same as Figure 3.

estrogen and dopamine classes, respectively. The true positive rates increase by the combination of the Tanimoto coefficient and the proximity measure. In addition, the prediction accuracy of the Dopamine (D1) antagonist class decreases because of their difficulty. In only Dopamine (D1) antagonist class, our combination model of LDA cannot create well, caused the number of all training data is less than 100 and the training data to create LDA model is also less than 50. As a result, the combi-

nation method is sensitive to only the number of training data. The line with results of our study showing that more than 100 ligands which have a subset can be recognized more efficiently with our combination method defined by supervised and unsupervised. But, in this situation, even other way of only the Tanimoto coefficient or other supervised method cannot be expected very good accuracy of the prediction.

In our study, the three types of searches of LDA produce relatively better performance for retrieved accuracy with respect to the number of active compounds retrieved. These combination can provide the results of our study showing that the tendency of distribution for the true positives and false positives of the Tanimoto coefficient and the proximity measure generated the improvement of the retrieval accuracy.

4. Conclusion

Fingerprint-based structural representation and the Tanimoto coefficient are very widely used for similarity searching and virtual screening of chemical databases. Although both are efficient and effective for prediction, the fingerprint and the Tanimoto coefficient exhibit several undesirable characteristics, and there is continuing interest in alternative approaches. We have described the methods of the proximity measure on similarity search and a method combining the different distances on fingerprint space and have succeeded in efficient similarity searching of large chemical databases. We have shown that such searches are effective for improving the degree of predicted accuracy. The Tanimoto coefficient and the proximity measure identified active compounds from the experimental datasets including several un-annotated compounds. The results of the proposed method and compound activity analyses revealed a useful method of obtaining similarity scores, and these observations could be rationalized considering some inherent features in the calculation of chemical structures.

Acknowledgments This study was supported by the Ministry of Education, Culture, Sports, Science and Technology Japan (MEXT) through the Science Grid NAREGI project, and through a Grant-in-Aid for Scientific Research on Priority Areas "Information Explosion". The authors would also like to thank Leo Breiman and for his ensemble learning, Random Forest, and *R* software.

References

1) Bender, A. and Glen, R. C.: Molecular similarity: a key technique in molecular informatics, *Org. Biomol. Chem.*, Vol. 1, pp. 3204–3218 (2004).
2) Kauffman, G. W. and Jurs, P. C.: QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitor using topologically based numerical descriptors, *J. Chem. Inf. Comput. Sci.*, Vol. 41, pp. 1533–1560 (2001).
3) Sheridan, R. P., Nachbar, R. B. and Bush, B. L.: Extending the trend vector: the trend matrix and sample-based partial least squares, *J. Comput-Aided Mol. Des.*, Vol. 8, pp. 323–340 (1994).
4) Doniger, S., Hofmann, T. and Yeh, J.: Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms, *J. Comput. Biol.*, Vol. 9, pp. 849–864 (2002).
5) Breiman, L.: Random forests, *Machine Learning*, Vol. 45, pp. 5–32 (2001).
6) Svetnik, V. and Liaw, A.: Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling, *J. Chem. Inf. Comput. Sci*, Vol. 43, pp. 1947–1958 (2003).
7) Gasteiger, J. and Engel, T.: *Chemoinformatics*, WILEY-VCH (2003).
8) Durant, J. L., Leland, B. A. and Henry, D. R.: Reoptimization of MDL Keys for Use in Drug Discovery, *J. Chem. Inf. Comput. Sci.*, Vol. 42, pp. 1273–1280 (2002).
9) Willett, P. and Barnard, J. M.: Chemical Similarity Searching, *J. Chem. Inf. Comput. Sci.*, Vol. 38, pp. 983–996 (1998).
10) Breiman, L.: Bagging predictors, *Machine Learning*, Vol. 24, pp. 123–140 (1996).
11) Ho, T. K.: The random subspace method for constructing decision forest, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 20, pp. 832–844 (1998).
12) Breiman, L. and Friedman, J. H.: *Classification and Regression Trees* (1984).
13) Balakrishnama, S. and Ganapathiraju, A.: *Linear Discriminant Analysis - A Brief Tutorial* (1998).

14) R: the R Development Core Team. *http://www.r-project.org.*
15) MDL ISIS/HOST software, MDL Information Systems, Inc.: MDL Drug Data Report Version 2004.2. *http://www.mdli.com.*