Original Paper

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

Gen Kawamura,[†1] Shigeto Seno,[†1] Yoichi Takenaka[†1] and Hideo Matsuda[†1]

Chemical and biological activities of compounds provide valuable information for discovering new drugs. The compound fingerprint that is represented by structural information of the activities is used for candidates for investigating similarity. However, there are several problems with predicting accuracy from the requirement in the compound structural similarity. Although the amount of compound data is growing rapidly, the number of well-annotated compounds, e.g., those in the MDL Drug Data Report (MDDR) database, has not increased quickly. Since the compounds that are known to have some activities of a biological class of the target are rare in the drug discovery process, the accuracy of the prediction should be increased as the activity decreases or the false positive rate should be maintained 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 of partial dependence for 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. We estimate the prediction accuracy of compound datasets extracted from MDDR using the proposed 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),2)]. 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 [3)]. To identify these targets, systematic exploration of selected target families, without prior restriction to a specific therapeutic area, appears to be a promising method by which to improve the ligand identification process in drug discovery. In addition, as the progress of the High Throughput Screening (HTS) technology in combinatorial chemistry has increased the number of compounds to enable researchers to estimate their activities rapidly, research analyzing similarity-based identification of ligands requires more efficient and time saving methods.

In selecting a suitable compound as a ligand, large chemical databases and combinatorial libraries have become increasingly important in chemical research, and structural information need to be searched in an appropriate manner. Therefore, we need to consider the basis in compound activity: "Structurally similar molecules are expected to exhibit similar physical properties or, similar biological activities" [4)]. This hypothesis helps us to select compounds that have similar activities. Based on this hypothesis, substructure searching is used for the retrieval of all the compounds in a database that contain substructures with activities [5)]. The substructure or fragment searching has been improved to provide a valuable tool for accessing databases of compound structures using the measurement of a numerical distance among the structural queries [1),6)]. Similarity searching can retrieve compounds that are sorted in order of similarity rank. High-ranked compounds are likely to have similar biological properties to the query. In general, in order to identify such biological properties of compounds, a fingerprint of the chemical structure must be obtained and

†1 Graduate School of Information Science and Technology, Osaka University

the distance in the compound space must be calculated. The MACCS key and the Tanimoto coefficient have been proposed for these purposes [7),8)]. This representation of chemical structure as a string of binary bits and the above-mentioned distance allow a very efficient searching method based on biological similarity [7),9)].

On the other hand, evaluating the structural similarity, several well-known methods in statistics and machine learning algorithms have been applied. All of these methods (e.g., decision tree [10)], artificial neural networks [11),12)], partial least squares [13)] and support vector machine [11)]) 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. Additionally, applying machine learning technique still remains quite limited. One of the reasons for this comes from the fact that such tools do not possess appropriate features required for their successful use. For example, artificial neural networks and nonlinear support vector machine have high performance. However, artificial neural network is not efficient in dealing with high-dimensional data without dimension reduction or preselection of descriptors. Nonlinear support vector machine is capable of dealing with high-dimensional data but is not robust to the presence of a large number of irrelevant descriptors, thus requiring preselection of fingerprint. Decision tree is probably the closest to having the desired combination of features. It handles high-dimensional data well, has the ability to ignore irrelevant descriptors, and handles multiple class of activity. However, even decision tree usually has relatively low prediction accuracy, and does not provide a good score to perform compound similarity.

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 [14)] 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. Specifically, random forest is efficiently estimated to predict compound activity, and to classify biological cluster in compound subsets [15),16)]. Applying this method to similarity search, we can obtain efficient performance for searching compounds in some activities, without reoptimization of the fingerprint.

The remainder of this paper is organized as follows. In Section 2, we describe the proposed method and related algorithms, the Tanimoto coefficient, random forest classifier [14)], the proximity measures, and a linear discriminant analysis. Section 3 describes the datasets used in the present estimation in a drug database. Section 4 presents the obtained results and a discussion. Finally conclusions are presented in Section 5.

2. Method

In this section we present the proposed method, which is based on the MACCS key, the Tanimoto coefficient, random forest, proximity measures, and their features. 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. This structure representation based on compound fragments is constructed by a string of keysets, which indicate whether a fragment of a specific substructure exists in the compound. Fragments of chemical structures can be coded in binary keys, which are presented as sequences of 0 s and 1 s (bitstrings). Here, 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 [17)]. To create these representations, the underlying technology is based on the general molecule per-

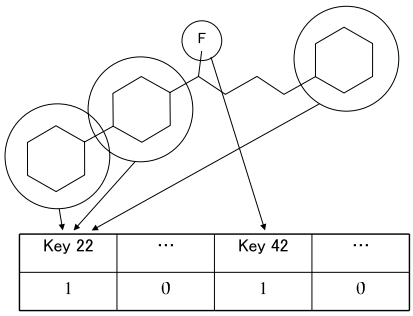

Fig. 1 Compound representation of the MACCS key.

ception algorithm, which perceives the number of atoms, bonds, and custom properties. The mapping of these properties into the binary keysets is performed under software control. For example, if compound include a Fluorine atom in their structure, the term of "Key 42" in the MACCS key denote 1. Also, an existence of 3 ring structures provides 1 to "Key 22" (as shown in **Fig. 1**).

2.2 Classifier Methods

In this study, we consider two 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

$$s = c/(a + b - c) \quad (1)$$

where a is the number of 1 s of the fingerprint of compound A, b is the number of 1 s of that of compound B, and c is the number of 1 s common to both A and B [6)]

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 addition, the small compounds in the same subset of an activity are likely to have few 1 s in a fingerprint because the Tanimoto coefficient, for example, does not take into account a common absence of features. For classification, the highest ranked compound will be selected as a class of the final subset. Since $c \leq min(a, b)$, low-similarity (and consequently high-dissimilarity) values will be obtained with small compounds.

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 [18)] concept and Ho's random subspace method [19)] to construct a collection of decision trees [14)]. Usually, in the study of Quantitative Structure-Activity Relationships (QSAR) [15),20)], random forest consists of B trees $\{T_1, ..., T_B\}$. For compound activity, a set of their class labels is

$$Y = \{C_l \mid l = 1, \ldots, m\} \quad (2)$$

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 [21),22)] 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 used 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 [15),16)]. 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 two advantages. This proximity measure is supervised because the proximity measure of random forest is created by the compounds depending on each dataset and database. With a small positive example with randomized unannotated compounds, the proximity measure can learn the scores between compounds efficiently. However, positive examples are required when the proximity measure is retrieved. On the other hand, since the unsupervised method of the Tanimoto coefficient cannot reflect the tendency, similarity searching sometimes selects irrelevant activity of the dataset. However, the score of the Tanimoto coefficient can adjust any dataset and make retrieval efficient.

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 [23)]. 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_j is the classification score for the respective case for the jth explaining variable, and M_j is the mean of variable x_j. 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 the present 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).

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

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

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).

$$Z_{i\,s}^{'}(tp) = \frac{s_i - M_s}{\sigma_s} \quad (i \in tp) \qquad (9)$$

$$Z_{i\,s}^{'}(fp) = \frac{s_i - M_s}{\sigma_s} \quad (i \in fp) \qquad (10)$$

$$Z_{i\,\hat{p}}^{'}(tp) = \frac{\hat{p}_i - M_{\hat{p}}}{\sigma_{\hat{p}}} \quad (i \in tp) \qquad (11)$$

$$Z_{i\,\hat{p}}^{'}(fp) = \frac{\hat{p}_i - M_{\hat{p}}}{\sigma_{\hat{p}}} \quad (i \in fp) \qquad (12)$$

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 $F_0^{'}$.

$$F_0^{'} = \frac{(M_{F_{tp}^{'}} + M_{F_{fp}^{'}})}{2} \qquad (13)$$

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 4.

3. Data Set

3.1 MDDR

The MDL Drug Data Report (MDDR) [24] is a licensed database that relates biological activities with drugs. The MDDR is a valuable resource and the activity labeling is more than adequate for the purpose for which it was intended, which is a human-readable database field. The information stored in the MDDR has been used to calculate similarities and to measure biological activities in a number of studies related to drug discovery [25]. Typically, a compound in the MDDR is assigned anywhere between 1 and 5 activity records. A compound record consists of an activity index (e.g., 1,100), a unique compound index (e.g., 80,003), and its MACCS key. The version of the MDDR that is used herein (2004.2) contains 153,000 compounds and 690 distinct activity indices.

3.2 Ligand Ontology

Ligand ontology, an annotation schema into ligand activities, has been proposed in a previous study [26]. The ontology is based on four major target classes: enzymes, G protein-coupled receptors, nuclear receptors, and ligand-gated ion channels, which are based on each database established by the EC [27], GPCRDB [28], NuclearDB [29] and LGICDB [30]. We use the results of the annotated information in the activity records of the MDDR to apply ligand ontology as a biological schema.

4. 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 [31].

4.1 Data set for Experiment

The compounds in the MDDR2004.2 database were sampled as data sets from the activity class associated with the target protein on the ligand ontology (shown in **Table 1**). Here we used Estrogen and Dopamine classes which have known activity as the target proteins of Estrogen and Dopamine receptors, respectively (shown in **Table 2**). The Dopamine ligand examples we selected were predicted all amine binding GPCR (G-protein coupled receptor) ligands, and their subsets (e.g., Dopamine D1 antagonist, Dopamine D3 antagonist) have similar biological property. Also, the estrogen ligand examples have several subsets and they were predicted as ligand of the nuclear receptors. It is suitable for evaluating the multiclass prediction and the similarity scoring in our study because they have the appropriate number of the subsets under the "estrogen-like" and "dopamine" parent classes and their biological subsets have the well-known activity [3].

Table 1 MDDR activity class and ligand ontology.

activity class	LO Parent
Dopamine (D1) Antagonist	dopamine
Dopamine (D3) Antagonist	dopamine
Dopamine (D4) Antagonist	dopamine
Estrogen	estrogen-like
Estrogen Receptor Modulator	estrogen-like
Antiestrogen	estrogen-like

LO : Ligand Ontology

Table 2 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	1,000

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 are regarded as positive examples and negative examples for predicting activity in our method, respectively. These data sets are 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).

4.2 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. **Table 3** and **Table 4** list the

Table 3 Success rate of estrogen.

method	success rate(%)
Proximity 1 nn	86.7
Proximity 3 nn	85.0
Proximity 10 nn	84.9
TC 1 nn	80.9

nn : Nearest Neighbor
TC : Tanimoto Coefficient

Table 4 Success rate of dopamine.

method	success rate(%)
Proximity 1 nn	89.5
Proximity 3 nn	89.1
Proximity 10 nn	89.5
TC 1 nn	84.4

nn : Nearest Neighbor
TC : Tanimoto Coefficient

results of the success rate, which is defined as follows, for each of the three types of k-nearest neighbors (k-nn) for the proximity measure and the Tanimoto coefficient.

$$Success\ Rate = \frac{tp + tn}{tp + tn + fp + fn} \quad (14)$$

where tp, fp, tn, and fn are the number of true positives, false positives, true negatives, and false negatives, respectively. Here, 1 nn, 3 nn, and 10 nn are the number of nearest neighbors extracted for each scored rank using proximity measure. These success rates denote precise values of prediction for activity classes. The data in Tables 3 and 4 indicate that the k-nn of the proximity measure performed better than the Tanimoto coefficient for these data sets for both the Estrogen and Dopamine target classes. These results also show that the distance of the proximity measure differs from that of the Tanimoto coefficient, and that the nearest neighbors of the proximity measure do not have a serious influence on the class discrimination.

Figures 2 and **3** show the results for precision when the activity classes are 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. Usually, 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.

$$Precision = \frac{tp}{tp + fp} \quad (15)$$

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

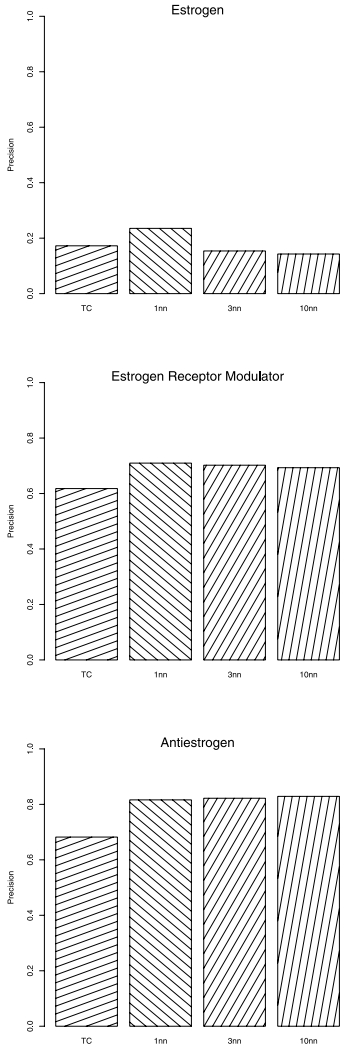

Fig. 2 Precision of proximity measure and Tanimoto coefficient for Estrogen. The bars labeled TC, 1 nn, 3 nn and 10 nn denote the precision of the Tanimoto coefficient and proximity measures for 1 nn, 3 nn, and 10 nn, respectively.

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[15], the proximity measure was already mentioned that it could show good performance of the hierarchical clustering. The Tanimoto coefficient, on the other hand, does not take into account the discrimi-

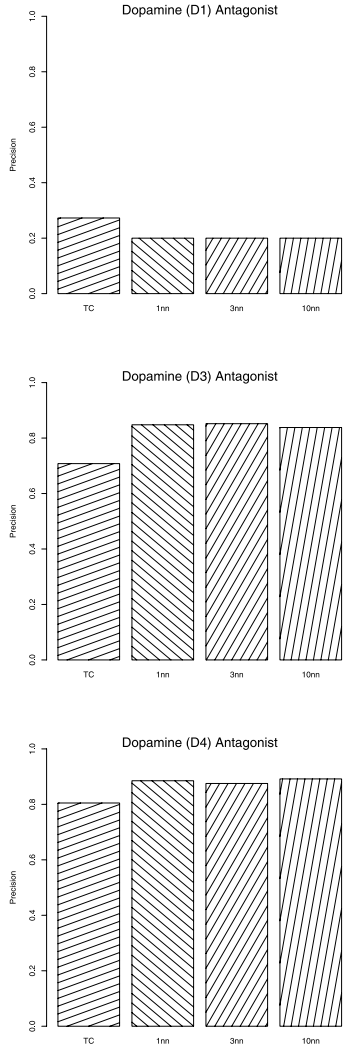

Fig. 3 Precision of proximity measure and Tanimoto coefficient for Dopamine. The bars labeled TC, 1 nn, 3 nn and 10 nn denote the precision of the Tanimoto coefficient and proximity measures for 1 nn, 3 nn, and 10 nn, respectively.

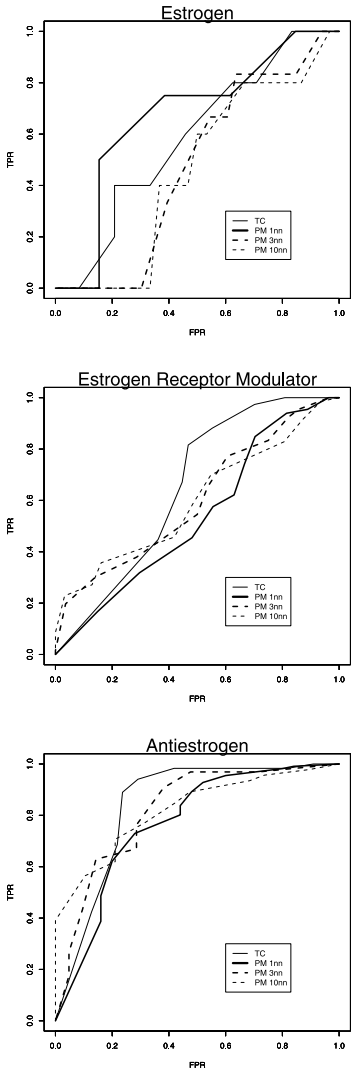

Fig. 4 ROC curves of the proximity measure and the Tanimoto coefficient for each Estrogen class. The solid, bold solid, bold dashed and dashed lines denote the ROC curves of the Tanimoto coefficient and proximity measures for 1 nn, 3 nn, and 10 nn, respectively.

nating power and treat all fingerprints equally, which resulted in lower performance. Also, from Figs. 2 and 3, 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.

Figures 4 and **5** present the results for the Receiver Operating Characteristic (ROC) curves comparing each of the activity classes. In the signal detection theory, the ROC curve is a graphical plot of *true positives vs. false positives*. The ROC analysis provides tools for selecting possibly optimal models and discard suboptimal models independently from (and prior to specifying) the cost context or the class distribution.

In our study of multi-class classifiers, to handle n classes, it is necessary to produce n different ROC curves, one for each class c_i. For the set of all classes C, the ROC curves plot the classification performance using the correctly predicted class c_i as the positive case P_i and all other predicted classes as the negative case

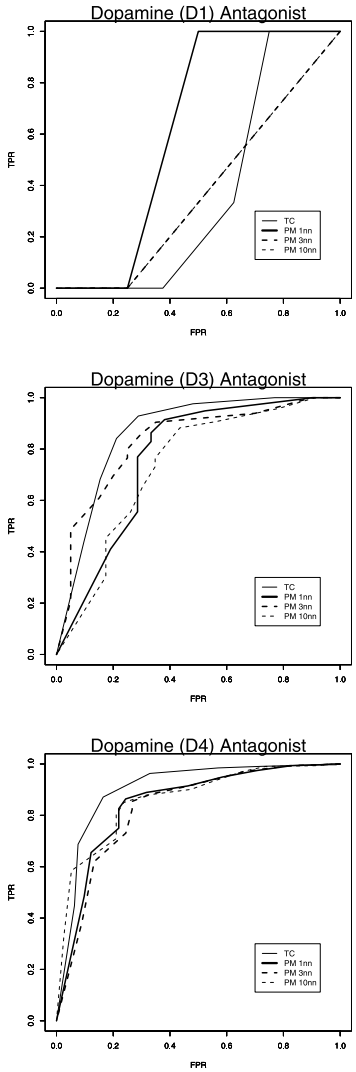

Fig. 5 ROC curves of the proximity measure and the Tanimoto coefficient for each Dopamine class. The solid, bold solid, bold dashed and dashed lines denote the ROC curves of the Tanimoto coefficient and proximity measures for 1 nn, 3 nn, and 10 nn, respectively.

N_i [32),33)]:

$$P_i = c_i \tag{16}$$

$$N_i = \bigcup_{j \neq i} c_j \in C \tag{17}$$

Figures 4 and 5 show that the scores based on the proximity measure provide relatively similar results to the Tanimoto coefficient in each class. In addition, for only a Dopamine (D1) antagonist, one obvious result of high degree of prediction in 1 nn of proximity measure is caused by the lower number of class subset to select similarity rankings. It can be seen that the proximity measure is comparable of the score rate except for a small number of the training data and validating data less than 100.

Note that the increases in the success rate and the accuracy of the proximity measure were caused by the deselection of randomized unannotated class. These data can condense the information contained in a set of candidate activity of the proximity measure, potentially resulting in better quality hit-lists than any of the hit-lists provided by the candidate rankings of the Tanimoto coefficient. 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 ROC curves show that the fusion-generated hit-lists might contain more accurate activities than either of the only candidate rankings by using the Tanimoto coefficient.

4.3 Combination of Scores

Figure 6 presents the results for the score distribution of the proximity measure and the Tanimoto coefficient obtained from the above experiments. This figure shows the dependency of the prediction results of each class. These scores by *proximity measure vs. Tanimoto coefficient* are plotted and can provide to consider.

Figures 7 and **8** show the results of precision plots before and after the consideration of the LDA for 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 combination 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

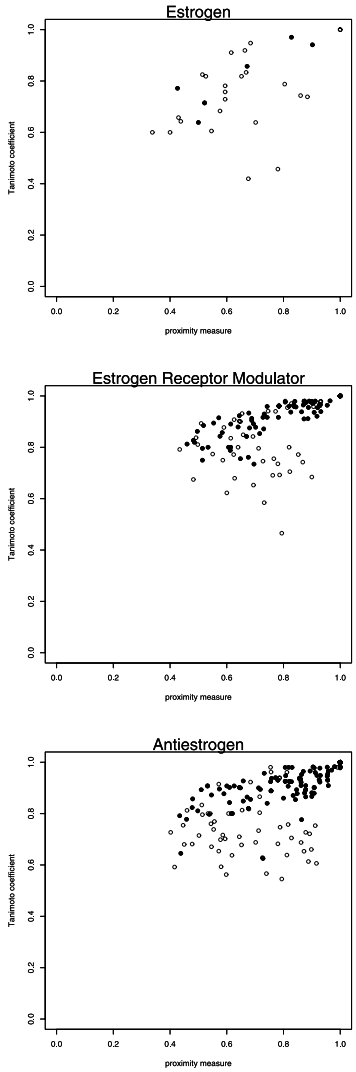

Fig. 6 Score distribution charts of the Tanimoto coefficient and the proximity measure for each Estrogen class. Horizontal and vertical lines represent the scores of the proximity measure and the Tanimoto coefficient, respectively. The black and white dots denote true positives and false positives, respectively.

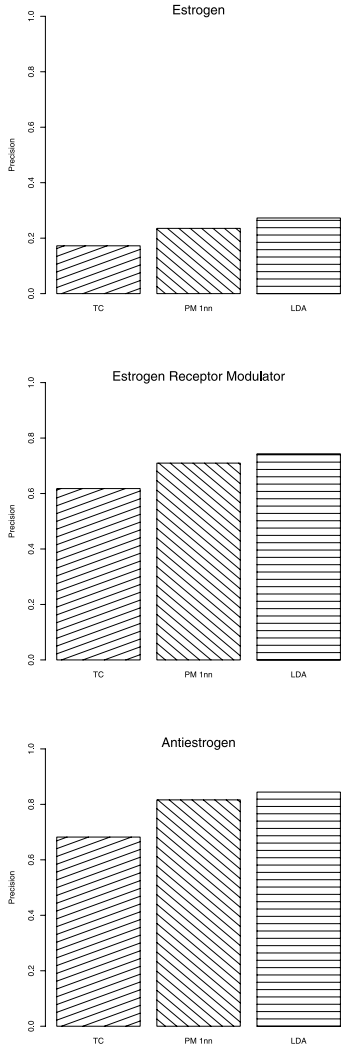

Fig. 7 Precision of the proximity measure and the Tanimoto coefficient for Estrogen. The bars labeled TC, PM 1 nn, and LDA denote the precision of the Tanimoto coefficient, the proximity measure for 1 nn, and their combination (LDA), respectively.

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. This procedure was used to test the hypothesis that there is a statistically significant difference in the scores of the active compounds retrieved by the two types of compound distance. The Tanimoto coefficient treats all bits of the MACCS key for the same importance, although the bits provides implicit information on the compound structure which depend on the activity. But, the Tanimoto coefficient can predict well on the information of simple chemical distance. On the other hand, the proximity measure has an important influence on individual class for each bit. Some compound classes are characterized by the multiple occurrences of structural features, such as a bit of the MACCS key, the weighted bits by the proximity measure can discriminate between active class.

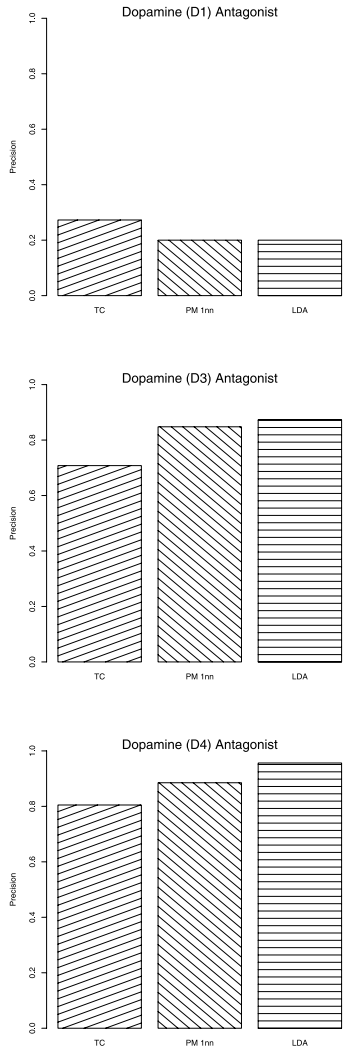

Fig. 8 Precision of the proximity measure and the Tanimoto coefficient for Dopamine. The bars labeled TC, PM 1 nn, and LDA denote the precision of the Tanimoto coefficient and proximity measures for 1 nn, and their combination (LDA).

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. The LDA with the proximity measure and the Tanimoto coefficient is efficient tool for the compound selection when the database includes the positive target compounds and the similar negative compounds.

5. 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. Experiments with the MDDR and the activity prediction of similarity searches demonstrated that even proximity-measure-based searching is comparable in effectiveness to Tanimoto-coefficient-based searching. The Tanimoto coefficient and the proximity measure identify active compounds from the experimental datasets including several unannotated 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) Raymond, J.W. and Willett, P.: Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases, *J. Comput-Aid. Mol. Des.*, Vol.16, pp.59–71 (2002).
3) Schuffenhauer, A., Floersheim, P. and Acklin, P.: Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins, *J. Chem.*

Inf. Comput. Sci., Vol.43, pp.391–405 (2003).
4) Brusle, M., Beck, B. and Schindler, T.: Descriptor Physical properties, and drug-likeness, *J. Med. Chem.*, Vol.45, pp.3345–3355 (2002).
5) Cummins, D.J., Andrews, C.W. and Bentley, J.A.: Molecular Diversity in Chemical Database: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds, *J. Chem. Inf. Comput. Sci.*, Vol.36, pp.750–763 (1996).
6) Willett, P. and Barnard, J.M.: Chemical Similarity Searching, *J. Chem. Inf. Comput. Sci.*, Vol.38, pp.983–996 (1998).
7) Xue, L., Godden, J.W. and Stahura, F.L.: Profile Scaling Increases the Similarity Search Performance of Molecular Fingerprints Containing Numerical Descriptors and Structural Keys, *J. Chem. Inf. Compt. Sci.*, Vol.43, pp. 1218–1225 (2003).
8) Godden, J.W., Xue, L. and Bajorath, J.: Combinatorial Preferences Affect Molecular Similarity/Diversity Calculations Using Binary Fingerprints and Tanimoto Coefficients, *J. Chem. Inf. Compt. Sci.*, Vol.40, pp.136–166 (2000).
9) Sadym, A. and Lagunin, A.: Prediction of biological activity spectra via the Internet, *SAR and QSAR Environmental Research*, Vol.14, pp.339–347 (2002).
10) Rusinko, A., Farmen, M.W., Lambert, C.G. and Brown, P.L.: Analysis of a large structure/biological activity data set using recursive partitioning, *J. Chem. Inf. Comput. Sci.*, Vol.39, pp.1017–1026 (1999).
11) 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).
12) 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).
13) 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).
14) Breiman, L.: Random forests, *Machine Learning*, Vol.45, pp.5–32 (2001).
15) 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).
16) Li, S. and Fedorowicz, A.: Application for the Random Forest Method in Studies of Local Lymph Node Assay Based Skin Sensitization Data, *J. Chem. Inf. Model*, Vol.45, pp.952–964 (2005).
17) 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).
18) Breiman, L.: Bagging predictors, *Machine Learning*, Vol.24, pp.123–140 (1996).
19) 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).
20) Svetnik, V. and Wang, T.: Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling, *J. Chem. Inf. Model.*, Vol.45, pp.786–799 (2005).
21) Breiman, L. and Friedman, J.H.: *Classification and Regression Trees* (1984).
22) Hyafil, L. and Rivest, R.L.: Constructing Optimal Binary Decision Trees is NP-Complete, *Inf. Process. Lett.*, Vol.5, No.1, pp. 15–17 (1976).
23) Balakrishnama, S. and Ganapathiraju, A.: *Linear Discriminant Analysis — A Brief Tutorial* (1998).
24) MDL ISIS/HOST software, MDL Information Systems, Inc. San Leandro, CA: MDL Drug Data Report Version 2004.2. http://www.mdli.com
25) Sheridan, R.P. and Shpungin, J.: Calculating Similarities between Biological Activities in MDL Drug Data Report Database, *J. Chem. Inf. Comput. Sci.*, Vol.44, pp.727–740 (2004).
26) Schuffehauer, A. and Zimmermann, J.: An Ontology for Pharmaceutical Ligands and Its Application for in Silico Screening and Library Design, *J. Chem. Inf. Comput. Sci.*, Vol.42, pp. 947–955 (2002).
27) Enzyme Nomenclature: EC. http://www.chem.qmw.ac.uk/iubmb/enzyme/
28) Information System for G protein-coupled receptors: GPCRDB. http://www.gpcr.org/7tm
29) An Information System for Nuclear Receptors: NucleaRDB. http://receptors.ucsf.edu/NR/
30) The Ligand Gated Ion Channel Database: LGICDB. http://www.pasteur.fr/recherche/banques/LGIC/LGIC.html
31) R: the R Development Core Team. http://www.r-project.org
32) Fawcett, T.: *ROC Graphs: Notes and Practical Considerations for Researchers* (2004).
33) Nakas, C.T. and Yiannoustsos, C.T.: Ordered

multiple-class ROC analysis with continuous measurements, *Statistics in medicine*, Vol.23, pp.3437–3449 (2004).

(Received October 16, 2007)
(Accepted November 29, 2007)

(Communicated by *Susumu Goto*)

Gen Kawamura received his Bachelor of Science degree in Applied Physics from Fukuoka University in 2000, and Master of Science degree in Fundamental Material Physics from Kyushu University in 2002. He is currently a student of Ph.D. course at Graduate School of Information Science and Technology, Osaka University since 2005. His research interests include machine learning to chemoinformatics field and Grid computing of database management to life sicences.

Shigeto Seno is an Assistant Professor of the Graduate School of Information Science and Technology, Osaka University. He received his B.E., M.E. and Ph.D. degrees from Osaka University in 2001, 2003 and 2006 respectively. He is a member of IEEE and IPSJ.

Yoichi Takenaka received the M.E. and Ph.D. in 1997, and 2000 from Osaka University, respectively. He worked for Osaka University from 2000 to 2002 as assistant professor, and now he is associate professor at Graduate School of Information Science and Technology, Osaka University. His research interests include Bioinformatics, DNA computing, and Neural Networks.

Hideo Matsuda is Professor of the Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University. He received his B.S., MEng., and Ph.D. degrees from Kobe University in 1982, 1984 and 1987, respectively. His research interests include computational analysis of genomic sequences, integrated biological databases, and data grid technology. He is a member of JSBi, ISCB, IEEE CS and ACM.