Novel 3D Descriptors Using Excluded Volume 2: Application to Drug Classification

Yukio Tominaga*

Department of Chemistry I, Discovery Research Laboratories I, Dainippon Pharmaceutical Co., Ltd., Enoki 33-94, Suita/Osaka 564, Japan

Received May 12, 1998

We have developed novel 3D descriptors, van der Waals excluded volume of each molecule and probe, which do not require alignment rules, and are not significantly affected by the orientation of molecules. Each probe is constructed by the excluded volume of two spheres with different radii and the center identical to each molecule's center of gravity. We applied the descriptors to the classification of three types of chemotherapeutic drugs by using two pattern recognition methods, Soft Independent Modeling of Class Analogy (SIMCA) and the combination of genetic algorithms (GAs) and nonhierarchical clustering (NHC). The classification and prediction results that were independently compared through the two methods showed that the latter method (84.7% correctly classified and 52.0% correctly predicted) could give a better model than the former method (80.0% correctly classified and 48.3% correctly predicted).

INTRODUCTION

Since the development of high-throughput biological screening methods, pharmaceutical companies have been able to screen many thousands of compounds in a short time. The screening data of many thousands of compounds sometimes includes binary or categorical information, e.g. active or inactive. An efficient analysis of the screening results is important to find a new lead compound or to optimize that lead compound. To achieve an efficient analysis of the screening results, it is essential to develop the following two methods:[1] an efficient conversion from the chemical structure information to numerical values (descriptors of chemical structures) and[2] an analysis or pattern recognition method of a large categorical data set.

Comparative molecular field analysis (CoMFA)[1] provides one of the most powerful descriptors which includes three-dimensional (3D) information of a molecule. CoMFA has been successfully used in drug design and 3D-QSAR.[2−7] The success of CoMFA depends on the quality of the alignment rule or positioning of a molecular model within a fixed lattice. The alignment rule is the key input for CoMFA analysis, while the rule cannot be simply determined. To avoid this difficulty, we have proposed novel 3D descriptors, EV_{whole}, EV_{type}, and EV_{both}.[8] The descriptors do not significantly rely on the alignment rules. The descriptors are calculated by the following procedure. The probe constructed by the excluded volume of two spheres with different radii and the identical center corresponding to each molecule's center of gravity is defined (Figure 1), then descriptors, EV_{whole}, which are van der Waals excluded volume of each molecule and probe is calculated. The descriptors are not significantly affected by the orientation of the molecules. In CoMFA, on the other hand, if the orientation of the molecules is changed, interaction energies between the molecules and each lattice point is changed accordingly

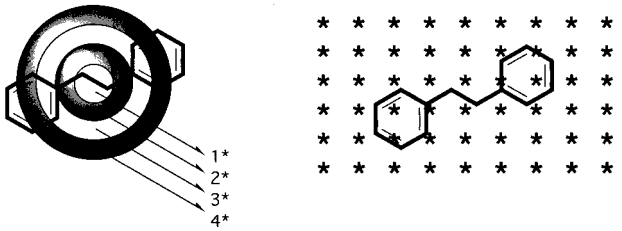

Figure 1. Schematic representation of the probe in this study (left) and CoMFA (right)*: layer numbers.

(Figure 1). The descriptors represent the expansion of molecular volume in 3D space. In addition to the descriptors, EV_{whole}, derived from the entire molecule, descriptors, EV_{type}, are calculated by determining the excluded volume between a specific type of atom of the molecule and probe. The atoms were classified into 15 types specified in SYBYL atom type notation:[9,10] C. Ar, C. 2, C. 3, all other types of carbon, O. 2, O. 3, all other types of oxygen, N. 2, N. 3, N. Ar, all other types nitrogen, phosphorus, sulfur, halogen, and all other types of atoms except hydrogen. The descriptors derived from this method represent the expansion of a specific type of atom in 3D space. As a result, the combination of EV_{whole} and EV_{type} becomes EV_{both}.

In our previous study, we applied these descriptors to the 3D-QSAR of corticosteroid-binding globulin (CBG) and demonstrated that these descriptors showed the superiority over the CoMFA.[8] The 3D-QSAR data set we used in the study, however, included merely 21 compounds.

To apply the descriptors to the classification of a categorical data set which includes a relatively large number of compounds, we developed the combined method of genetic algorithms (GAs)[11−15] and nonhierarchical clustering (NHC)[15−17] that is appropriate for analyzing categorical data set. In this study, we used the combined method with EV_{both} to classify three types of chemotherapeutic drugs, i.e., antibacterials, antifungals, and antineoplastics, that are registered in the comprehensive medicinal chemistry (CMC) database.[18] The classification results of the combined method

* Tel.: +81-6-337-5898. Fax: +81-6-338-7656. E-mail: yukio-tominaga@dainippon-pharm.co.jp.

10.1021/ci980208s CCC: $15.00
Published on Web 10/02/1998

were compared to those of Soft Independent Modeling of Class Analogy (SIMCA),[19,20] a popular pattern recognition method.

METHOD

1. Data Set. Four-hundred fifty-two[452] antibacterial drugs, 102 antifungal drugs, and 407 antineoplastic drugs were selected from CMC database. These 2D chemical structures were converted into 3D structures by using Converter 95.0[21] within Insight II. These 3D structures were further optimized by MAXIMIN2 within SYBYL.[10]

2. System Used for Data Analysis. All calculations were carried out by using Indigo 2 running version 6.2 of the IRIX operating system.

3. Descriptors. Detailed calculation methods have been mentioned elsewhere,[8] so we describe the calculation conditions only. Iodide atom, I (with van der Waals radius of 2.05 Å), was used as the component of probe. We used 21 layers so EV_{whole}, EV_{type}, and EV_{both} consisted of 21, 315(21 × 15), and 336 (21 + 315) descriptors, respectively. EV_{both} was used in the analyses below.

4. SIMCA. SIMCA analysis was performed within SYBYL.[10] All descriptors were regularized to give them equivalent variance and means of zero. Optimum components of the principal component model for each category was determined by cross-validation.

5. Combined Method of GAs and NHC. First, a specific number of compounds are randomly selected as the seed points of NHC. NHC is then performed by using the seed points. Euclidean distance between a seed point and compound is used as the similarity measure. Once all compounds are assigned to the clusters containing the closest seed points, the number of compounds which belongs to the same category as the seed point is counted for each cluster. In other words, the number of correctly classified compounds are counted. To increase the number of correctly classified compounds, the selection of appropriate seed points for NHC is critical. GAs are used to optimize the combination of seed points for NHC. The number of correctly classified samples are used as the fitness function of GAs.

5.1. Implementation of GAs. We encoded the GAs as a FORTRAN program.

5.1.1. Coding. To classify between the seed point compounds and nonseed point compounds, we assigned a binary string, either one or zero, to each compound. The value of one implies the corresponding compound is the seed point for NHC.

5.1.2. Fitness Functions. The number of correctly classified compounds (see section 5) is used as the fitness function. If there are some singletons, the fitness score is set to zero because the model constructed with singletons only is invaluable.

5.1.3. Initial Population. A population of 1000 randomly selected combinations of strings is generated. Fifty strings with highest fitness scores, the number of correctly classified samples, are then selected as the initial population.

5.1.4. Selection. Three pairs of strings (parents) are selected using roulette wheel selection method. Using a roulette wheel which consists of portions corresponding to each member of the string, the selection probability of the strings is proportional to its fitness function.

Table 1. Results of SIMCA Analysis for Training Set

	no. of components	no. of drugs	no. of correctly classified drugs	ratio of correctly classified[a] (%)
antibacterial	28	452	371	82.0
antifungal	20	102	71	69.6
antineoplastics	27	407	327	80.3
total		961	769	80.0

[a] (Number of correctly classified drugs)/(number of drugs).

5.1.5. Crossover and Mutation. Three pairs of parents are crossovered. Each offspring is a subject to random point mutation. The mutation ratio is less than 1%.

5.1.6. Exploration. The highest fitness score of an offspring is selected. Then the parent with similar strings to the offspring is selected. Tanimoto coefficient is used as the similarity measure (see eq 1).

$$\text{Tanimoto coefficient} = \{\sum(\text{parent's string})(\text{offspring's string})\}/\{\sum(\text{parent's string})^2 + \sum(\text{offspring's string})^2 - \sum(\text{parent's string})(\text{offspring's string}) \quad (1)$$

If the fitness score of the offspring is superior to that of the parent, the parent is replaced with the offspring. Sections 5.1.4−5.1.6 are repeated 5000 times to improve the fitness score of the population.

5.2. Selection of the Model and Prediction of New Samples. When exploration is finished, the model with maximum score in the final generation is selected. The prediction of a new compound is performed by using the model. Based on voronoi diagrams,[22] the category of a new compound is predicted as the category of the seed compound located closest to the new compound.

RESULTS AND DISCUSSIONS

SIMCA analysis was performed by using EV_{both}, 336 descriptors. An appropriate number of principal components for each category was determined by cross-validation. The number of principal components was 28, 20, and 27 for antibacterials, antifungals, and antineoplastics, respectively. The statistical results are shown in Table 1. In the SIMCA model, 82.0, 69.6, and 80.3% of antibacterials, antifungals, and antineoplastics, respectively, were correctly classified. That is, a total of 80.0% of the drugs was correctly classified.

The combined method of GAs and NHC was performed by using EV_{both}, 336 descriptors. The maximum, average, and minimum scores were monitored as function of generation. The results are shown in Figure 2. The maximum score increased significantly in the early generations, and then gradually saturated along with the increase of generations. The maximum score for each drug in the final generation is shown in Table 2. Sixty-seven, 13, and 62 seed points were selected from antibacterials, antifungals, and antineoplastics, respectively. A total of 142 seed points were selected. In the combined method model, 92.2, 57.8, and 83.0% of antibacterials, antifungals, and antineoplastics, respectively, were correctly classified. That is, a total of 84.7% of the drugs was correctly classified.

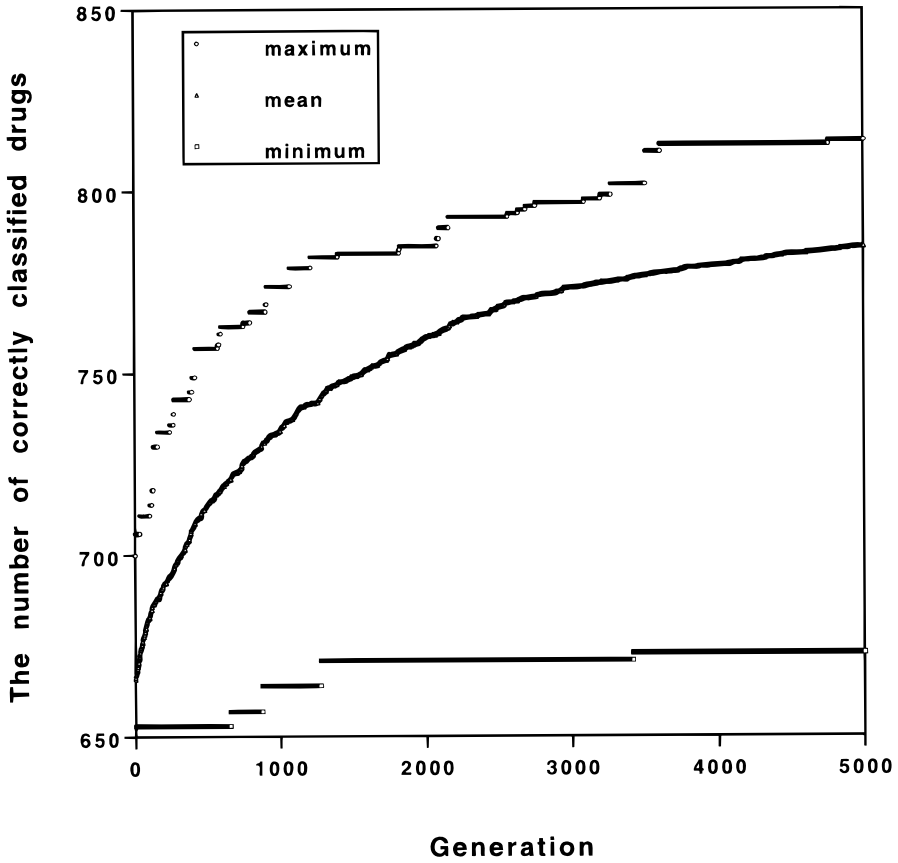

Figure 2. A graph of the maximum, average, and minimum scores as a function of generation.

Table 2. Results of the Combination GAs and NHC no. for Training Set

	no. of drugs	no. of correctly classified drugs	ratio of correctly classified[a] (%)
antibacterial	452	417	92.2
antifungal	102	59	57.8
antineoplastics	407	338	83.0
total	961	814	84.7

[a] (Number of correctly classified drugs)/(number of drugs).

Table 3. Results of SIMCA Analysis for Test Set of Compounds

	no. of compds	no. of correctly classified compounds	ratio of correctly classified[a] (%)
antibacterial	100	51	51.0
antifungal	100	35	35.0
antineoplastics	100	59	59.0
total	300	145	48.3

[a] (Number of correctly classified compounds)/(number of compounds).

To compare the predictiveness of SIMCA model to that of the combined method model, these models were used to predict the category for new compounds. One hundred compounds of each category were selected randomly from MDL drug data report (MDDR) database[9] which includes drug candidates. Three-dimensional (3D) structures of the selected drug candidates were then constructed and EV_{both} was assigned. The results of the prediction are shown in Tables 3 and 4. In the SIMCA model, 51.0, 35.0, and 59.0% of antibacterials, antifungals, and antineoplastics, respectively, were correctly predicted. That is, a total of 48.3% of the drug candidates were correctly predicted. In the combined method model, 69.0, 34.0, and 53.0% of antibacterials, antifungals, and antineoplastics, respectively, were correctly predicted. That is, a total of 52.0% of the drug candidates was correctly predicted.

Table 4. Results of the Combination of GAs and NHC no. for Test Set of Compounds

	no. of compds	no. of correctly classified compds	ratio of correctly classified[a] (%)
antibacterial	100	69	69.0
antifungal	100	34	34.0
antineoplastics	100	53	53.0
total	300	156	52.0

[a] (Number of correctly classified compounds)/(number of compounds).

When the results of the classification and prediction were compared for both models, the combined method of GAs and NHC gave a better model (84.7% correctly classified and 52.0% correctly predicted) than SIMCA (80.0% correctly classified and 48.3% correctly predicted). The main reason of the superiority of the combined method over SIMCA could be laid upon the different modeling strategies of the two methods. In SIMCA modeling, common features within each category is extracted by principal component analysis. In the combined method of GAs and NHC, on the other hand, a number of samples are selected from each category to represent the feature of each category.

Although we categorized chemotherapeutic drugs into three categories, even drugs which belong to the same category sometimes acted by different mechanisms. It is very important to consider subcategories when the asymmetric data set is analyzed. Subcategories were considered in the

combined method of GAs and NHC, and as a result the model gives better results compared to SIMCA.

CONCLUSIONS

EV_{both}, 336 descriptors, was successfully assigned to 961 chemotherapeutic drugs. Drug classification was performed by means of two pattern recognition methods, SIMCA and the combination of GAs and NHC. More than 80% of drugs was correctly classified in both methods. The combined method gave a better model because the model gave better results in both classification and prediction.

ACKNOWLEDGMENT

The author would like to thank the referees for their helpful and constructive comments.

REFERENCES AND NOTES

(1) Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(2) Raghavan, K.; Buolamwini, J. K.; Fesen, M. R.; Pommier, Y.; Kohn, K. W.; Weinstein, J. N. Three-Dimensional Quantitative Structure−Activity Relationship (QSAR) of HIV Integrase Inhibitors: A Comparative Molecular Field Analysis (CoMFA) Study. *J. Med. Chem.* **1995**, *38*, 890−897.

(3) Tsventan, G.; Gantchev, H. A.; Johan, E. L. Quantitative Structure−Activity Relationships/Comparative Molecular Field Analysis (QSAR/CoMFA) for Receptor-Binding Properties of Halogenated Estradiol Derivatives. *J. Med. Chem.* **1994**, *37*, 4164−4176.

(4) Waller, C. L.; Oprea, T. I.; Giolitti, A.; Marshall, G. R. Three-Dimensional QSAR of Human Immunodeficiency Virus (I) Protease Inhibitors. 1. A CoMFA Study Employing Experimentally-Determined Alignment Rules. *J. Med. Chem.* **1993**, *36*, 4152−4160.

(5) Waller, C. L.; Marshall, G. R. Three-Dimensional Quantitative Structure−Activity Relationship of Angiotensin-Converting Enzyme and the Rmolysin Inhibitors. II. A Comparison of CoMFA Models Incorporating Molecular Orbital Fields and Desolvation Free Energies Based on Active-Analogue and Complementary-Receptor-Field Alignment Rules. *J. Med. Chem.* **1993**, *36*, 2390−2403.

(6) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3D-QSAR of Angiotensin-Converting Enzyme and Thermolysin Inhibitors: a Comparison of CoMFA Models Based on Deduced and Experimentally Determined Active Site Geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372−5384.

(7) Atul, A.; Philip, P. P.; Ethan, W. T.; Hong, B. L.; Torsten, D.; Margareta, H.; Youhua, Y.; Georgina, L.; David, L. N.; John, W. R.; Arnold, R. M. Three-Dimensional Quantitative Structure−Activity Relationships of 5-HT Receptor Binding Data for Tetrahydropyridinylindole Derivatives: A Comparison of the Hansch and CoMFA Methods. *J. Med. Chem.* **1993**, *36*, 4006−4014.

(8) Tominaga, Y.; Fujiwara, I. Novel 3D Descriptors Using Excluded Volume: Application to 3D Quantitative Structure−Activity Relationships. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1158−1161.

(9) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, F. R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of 'Molecular Diversity' Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(10) SYBYL Molecular Modeling Software, version 6.3; Tripos Associates, Inc.: St. Louis, MO 63144.

(11) Hibbert, D. B. Genetic algorithms in chemistry. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 277−293.

(12) Lucasius, C. B.; Kateman, G. Understanding and Using Genetic Algorithms Part 1. Concepts, Properties, and Context. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 1−33.

(13) Lucasius, C. B.; Kateman, G. Understanding and Using Genetic Algorithms Part 2. Representation, Configuration, and Hybridization. *Chemom. Intell. Lab. Syst.* **1994**, *25*, 99−145.

(14) Tominaga, Y. Representative Subset Selection Using Genetic Algorithms. *Chemom. Intell. Lab. Syst.* in press.

(15) Tominaga, Y. Data Structure Comparison Using Box Counting Analysis. *J. Chem. Inf. Comput. Sci.* in press.

(16) Hartigan, J. A. *Clustering Algorithms*; John Wiley & Sons: New York, 1975.

(17) Tominaga, Y.; Fujiwara, I. Data Structure Comparison Using Fractal Analysis. *Chemom. Intell. Lab. Syst.* **1997**, *39*, 187−193.

(18) MDL Information Systems, Inc.: 14600 Catalina Street, San Leandro, CA 94577.

(19) Dunn III, W. J.; Wold, S.; Martin, Y. C. Structure−Activity Study of β-Adrenergic Agents Using the SIMCA Method of Pattern Recognition. *J. Med. Chem.* **1978**, *21*, 922−930.

(20) Wold, S.; Dunn III, W. J. Multivariate Quantitative Structure−Activity Relationships (QSAR): Conditions for Their Applicability. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 6−13.

(21) Converter 95.0; MSI: 9685 Scranton Road, San Diego, CA 92121-2777 U.S.A.

(22) Aurenhammer, F. Voronoi Diagrams − A Survey of a Fundamental Geometric Data Structure. *ACM Computing Surveys* **1991**, *23*, 347−405.

CI980208S