Similarity Perception of Reactions Catalyzed by Oxidoreductases and Hydrolases Using Different Classification Methods

Xiaoying Hu,[†] Aixia Yan,*[,†] Tianwei Tan,[‡] Oliver Sacher,[§] and Johann Gasteiger[§,||]

State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering, P.O. Box 53, Beijing University of Chemical Technology, 15 BeiSanHuan East Road, Beijing 100029, People's Republic of China, Beijing Key Lab of Bioprocess Laboratory, College of Life Science and Technology, Beijing University of Chemical Technology, Beijing 100029, People's Republic of China, Molecular Networks GmbH, Henkestrasse 91, D-91052 Erlangen, Germany, and Universität Erlangen-Nürnberg, Computer-Chemie-Centrum and Institute of Organic Chemistry, Nägelsbachstrasse 25, D-91052 Erlangen, Germany

Received December 15, 2009

In this work, the perception of similarity of reactions catalyzed by hydrolases and oxidoreductases on the basis of the overall breaking and making of bonds of reactions is investigated. Six physicochemical properties for the reacting bond in the substrate of each enzymatic reaction were calculated to describe the characteristics of each reaction. The 311 reactions catalyzed by hydrolases (EC 3.b.c.d) and the 651 reactions catalyzed by oxidoreductases (EC 1.b.c.d) were classified by Kohonen's self-organizing neural network (KohNN), by a support vector machine (SVM), and by hierarchical clustering analysis (HCA). For the 311 reactions catalyzed by hydrolases, the classification accuracy of 95.8% by a KohNN and 97.7% by an SVM was achieved. For the 651 reactions catalyzed by oxidoreductases, the classification accuracy was 93.4% and 96.3% by a KohNN and a SVM, respectively. The similarities of reactions reflected by the physicochemical effects of reacting bonds were compared with the traditional Enzyme Commission (EC) classification system. The results of a KohNN and a SVM are similar to those of the EC classification system method. However, the perception of similarity of reactions by a KohNN and a SVM shows finer details of the enzymatic reactions and thus could provide a good basis for the comparison of enzymes.

1. INTRODUCTION

The Enzyme Commission (EC) classification system is widely used in chemistry and biology and is maintained by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB).[1] Enzymes are classified into six main classes; each main class is reclassified into several subclasses; a subclass is classified into subsubclasses. In this EC system, each enzyme has its unique EC number: EC a.b.c.d, where "a" refers to the six main classes of enzymes; "b" indicates the subclasses; "c" represents the subsubclasses; "d" is the serial number of the enzyme in its subsubclass. This method of enzyme classification is based on several criteria, such as reaction types, substrates, transferred groups, and acceptor groups.

As the EC classification system is based on a variety of criteria, inconsistencies may arise, and the EC number might not reflect information on the mechanism of a reaction. Furthermore, it is possible for an enzyme to obtain an EC number that later on is changed to another EC number. For example, the enzyme deoxyhypusine synthase was assigned the EC code (EC 1.1.1.249) in 1999, and then it was deleted in the same year, until in 2001, the EC code of this enzyme was revised to EC 2.5.1.46. So it is useful to classify enzyme-catalyzed reactions on the basis of other criteria, such as the breaking and making of bonds of reactions or the mechanism of reactions.

Recently, the similarity perception of enzyme-catalyzed reactions has become a topic of great interest, as the data on proteins with unknown function grow fast[2] and the interest in metabolic reactions is increasing. A series of research groups is working on the analysis and perception of similarity of enzyme-catalyzed reactions. Kotera et al.[3] introduced a method for identifying the reaction center of a reaction by matching the substrate onto the product and used this information to assign an RC (reaction classification) number, comparing this number to the EC system. Zhang and Aires-de-Sousa[4] computed a reaction Molmap, coding the reaction center, based on physicochemical descriptors of the reacting bonds. This information was used for an automatic classification of metabolic reactions.[5] The group of Gasteiger performed reaction classifications based on the physicochemical properties of the bonds or atoms at the reaction center.[6−10] O'Boyle et al.[11] investigated the reaction mechanisms for measuring enzyme similarity.

As enzymatic reactions are linked with proteins (enzymes), databases of reactions and enzymes have been constructed. Babbitt et al.[12] established the Structure Function Linkage Database (SFLD) that hierarchically classifies enzymes by linking the specific partial reactions with the conserved

* Corresponding author. Telephone: +86-10-64421335. Fax: +86-10-64416428. E-mail: aixia_yan@yahoo.com and yanax@mail.buct.edu.cn.
[†] State Key Laboratory of Chemical Resource Engineering, Beijing University of Chemical Technology.
[‡] Beijing Key Lab of Bioprocess Laboratory, Beijing University of Chemical Technology.
[§] Molecular Networks GmbH.
[||] Universität Erlangen-Nürnberg.

10.1021/ci9004833
Published on Web 06/01/2010

structural elements that mediate them. Mechanism, Annotation, and Classification in Enzymes (MACiE)[13] is a database documenting enzyme reaction mechanism. The Catalytic Site Atlas (CSA)[14] focuses on the active site and catalytic residues in enzymes. The RLCP classification by Nagano classifies the enzyme catalytic mechanisms at four levels including the basic reaction type, the ligand group, catalytic type and residues/cofactors. This is laid down in the Enzyme Catalytic-Mechanism Database (EzCatDB).[15]

This study builds on a publication by Sacher et al.[9] The reactions studied here were represented by the same descriptors as those used in the previous work.[9] In the present work, a larger data set for hydrolases (EC 3.b.c.d) and a new data set of oxidoreductases (EC 1.b.c.d) were analyzed. In the previous work,[9] only the classification method of Kohonen's self-organizing neural network (KohNN) was used; in this work, two additional classification methods, support vector machine (SVM) and hierarchical clustering analysis (HCA), were also applied. The results obtained from the three different methods were then compared with the EC classification system.[1] The results from KohNN and SVM could be characterized by a quantitative measure of the predictive power.

2. DATA SETS AND METHODS

2.1. BioPath. The data sets used in this work were taken from Biochemical Pathways database version 2.0 (BioPath).[16] This database was derived from the information of the well-known Biochemical Pathways wall chart[17] and the corresponding atlas.[18] Several important features of this database are unique, such as complete mass balance, availability of the three-dimensional (3D) structure of metabolites, annotations of atom-to-atom mappings, and marking of the reaction centers. The BioPath database allows the extraction of the reaction center of each reaction and, therefore, enables the calculation of physicochemical descriptors describing the reaction center, which is used for the studies described here. The data of the BioPath database can be accessed through the web-based retrieval system BioPath.Explore, which offers a variety of structure- and text-based search possibilities.[19]

2.2. Data Sets. *2.2.1. Data Sets of Hydrolases (EC 3.b.c.d).* Sacher, et al.[9] have investigated the similarity of reactions catalyzed by hydrolases comprising 135 reactions. In the present work, the data set of hydrolases was expanded to 311 reactions containing three more subclasses of hydrolases than in the previous publication:[9] EC 3.3.c.d (acting on ether bonds), EC 3.10.c.d (acting on sulfur−nitrogen bonds), and EC 3.11.c.d (acting on carbon−phosphorus bonds). In each of these reactions, only one bond was broken (besides the O−H bond of water). The structure of the data set EC 3.b.c.d is shown in Table 1. Within this data set, two subclasses EC 3.1.c.d and EC 3.5.c.d were selected for a more detailed investigation because they contained the largest number of reactions.

After extraction from BioPath, the data set was preprocessed before the calculation of physicochemical properties. Bonds to hydrogen atoms that were broken or made and bonds that only changed bond order were not considered. Some reactions contained generalized groups such as R. These R's were substituted by an H atom because the methods for calculating physicochemical effects need clearly defined molecules. Then six physicochemical descriptors (see Section 2.3) were calculated for all reacting bonds using the program PETRA version 4.[20]

Table 1. Subclasses and Subsubclasses of Hydrolases and the Number of Related Reactions[a]

class (reactions)	subclasses (reactions)	subsubclasses (reactions)
EC 3.b.c.d (311)	EC 3.1.c.d (112)	EC 3.1.1.d (39)
		EC 3.1.2.d (6)
		EC 3.1.3.d (43)
		EC 3.1.4.d (14)
		EC 3.1.5.d (1)
		EC 3.1.6.d (6)
		EC 3.1.7.d (1)
		EC 3.1.8.d (2)
	EC 3.2.c.d (36)	
	EC 3.3.c.d (4)	
	EC 3.4.c.d (3)	
	EC 3.5.c.d (92)	EC 3.5.1.d (35)
		EC 3.5.2.d (9)
		EC 3.5.3.d (11)
		EC 3.5.4.d (24)
		EC 3.5.5.d (4)
		EC 3.5.99.d (9)
	EC 3.6.c.d (39)	
	EC 3.7.c.d (12)	
	EC 3.8.c.d (10)	
	EC 3.10.c.d (1)	
	EC 3.11.c.d (2)	

[a] Subsubclasses of EC 3.1.c.d and EC 3.5.c.d were selected for a more detailed investigation.

2.2.2. Data Sets of Oxidoreductases (EC 1.b.c.d). All 651 reactions catalyzed by oxidoreductases (EC 1.b.c.d, catalyzing oxidoreduction reactions) were extracted from the BioPath database. The substrate that is oxidized is regarded as a hydrogen donor. The structure of data set EC 1.b.c.d is shown in Table 2. Within this data set, two subclasses EC 1.1.c.d and EC 1.14.c.d were selected for a more detailed investigation, again because of the size of these subclasses.

The data set of class EC 1.b.c.d contained 14 subclasses of oxidoreductases: EC 1.1.c.d (acting on a CH−OH group as donor), EC 1.2.c.d (acting on an aldehyde or oxo group as donor), EC 1.3.c.d (acting on a CH−CH group as donor), EC 1.4.c.d (acting on a $CH-NH_2$ group as donor), EC 1.5.c.d (acting on a CH−NH group as donor), EC 1.7.c.d (acting on other nitrogen atom containing compounds as donors), EC 1.8.c.d (acting on a sulfur group as donor), EC 1.11.c.d (acting on a peroxide as acceptor), EC 1.13.c.d (acting on single donors, with incorporation of an oxygen molecule), EC 1.14.c.d (acting on paired donors, with incorporation or reduction of an oxygen molecule), EC 1.17.c.d (acting on CH or CH_2 groups), EC 1.18.c.d (acting on iron−sulfur proteins as donors), EC 1.21.c.d (acting on X−H and Y−H to form an X−Y bond), and EC 1.97.c.d (oxidoreductases not belonging to other subclasses).

All reactions catalyzed by enzymes of subclass EC 1.1.c.d were extracted from all 651 reactions catalyzed by oxidoreductases for a more detailed analysis. This data set contained 242 reactions. Four subsubclasses were included in this data set: EC 1.1.1.d (with NAD^+ or $NADP^+$ as acceptor), EC 1.1.2.d (with a cytochrome as acceptor), EC 1.1.3.d (with oxygen as acceptor), and EC 1.1.99.d (with other acceptors). The number of each subsubclass is shown in Table 2. In this data set, there are two reactions catalyzed

Table 2. Subclasses and Subsubclasses of Oxidoreductases and the Number of Related Reactions[a]

class (reactions)	subclasses (reactions)	subsubclasses (reactions)
	EC 1.1.c.d (242)	EC 1.1.1.d (211)
		EC 1.1.2.d (3)
		EC 1.1.3.d (16)
		EC 1.1.99.d (12)
	EC 1.2.c.d (96)	
	EC 1.3.c.d (72)	
	EC 1.4.c.d (45)	
	EC 1.5.c.d (35)	
	EC 1.7.c.d (3)	
	EC 1.8.c.d (8)	
	EC 1.11.c.d (2)	
	EC 1.13.c.d (25)	
EC 1.b.c.d (651)		
	EC 1.14.c.d (92)	EC 1.14.11.d (12)
		EC 1.14.12.d (3)
		EC 1.14.13.d (46)
		EC 1.14.14.d (1)
		EC 1.14.15.d (5)
		EC 1.14.18.d (2)
		EC 1.14.21.d (7)
		EC 1.14.99.d (13)
		EC 1.14.-.- (3)
	EC 1.17.c.d (15)	
	EC 1.18.c.d (1)	
	EC 1.21.c.d (2)	
	EC 1.97.c.d (13)	

[a] Subsubclasses of EC 1.1.c.d and EC 1.14.c.d were selected for a more detailed investigation.

by an enzyme whose EC code was not fully assigned: EC 1.1.-.-. The reactions catalyzed by enzymes of these subsubclasses are shown in Figure 1.

All reactions catalyzed by enzymes of subclass EC 1.14.c.d were extracted from all 651 reactions catalyzed by oxidoreductases for a more detailed analysis. This data set contained 92 reactions. Eight subsubclasses were included in this data set: EC 1.14.11.d (with 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors), EC 1.14.12.d (with NADH or NADPH as one donor, and incorporation of two atoms of oxygen into one donor), EC 1.14.13.d (with NADH or NADPH as one donor, and incorporation of one atom of oxygen), EC 1.14.14.d (with reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen), EC 1.14.15.d (with reduced iron−sulfur protein as one donor, and incorporation of one atom of oxygen), EC 1.14.18.d (with another compound as one donor, and incorporation of one atom of oxygen), EC 1.14.21.d (with NADH or NADPH as one donor, and the other dehydrogenated), and EC 1.14.99.d (miscellaneous). Within this subclass, there are three reactions that do not have a fully assigned EC code: EC 1.14.-.-. The number of each subsubclass is shown in Table 2. The reactions catalyzed by enzymes of subsubclasses EC 1.14.11.d, EC 1.14.12.d, EC 1.14.13.d, EC 1.14.14.d, EC 1.14.15.d, EC 1.14.18.d and EC 1.14.21.d are shown in Figure 2.

After extraction from BioPath, the data set was preprocessed before the calculation of physicochemical properties. Some reactions contained generalized residues, R, as groups. These R's were substituted by an H atom; otherwise, the calculation of the physicochemical descriptors would fail as the calculation methods require exact atom information. In some cases the direction of a reaction was reversed in order to obtain a unique redox direction representing all reactions as oxidations. Making/breaking bonds in cofactors (NAD, O_2, etc.) and marked bonds to hydrogen atoms (except Csp2−H bonds, e.g., of aldehydes) were not considered. Then six physicochemical descriptors (see Section 2.3) were calculated for all reacting bonds using the program PETRA, version 4.[20] Each investigated reaction breaks one bond or changes the bond order.

2.3. Choice of Descriptors. To represent a reaction, six physicochemical descriptors were selected to represent each reacting bond on the substrate side, suited to describe the electronic character of the bonds taking part in the reaction. The choice of descriptors was based on the work previously reported[21] and modified for the needs of biochemical reactions. All descriptors were calculated by rapid empirical procedures implemented in the program PETRA.[20] The descriptors used in this study are the same as those in the previous publication.[9] The six descriptors are as follows: difference in partial atomic charges (sum of σ and π charges), Δq_{tot}; difference in σ-electronegativities, $\Delta\chi_\sigma$; difference in π-electronegativities, $\Delta\chi_\pi$; effective bond polarizability, α_b; delocalization stabilization of a negative charge, D^-; and delocalization stabilization of a positive charge, D^+.

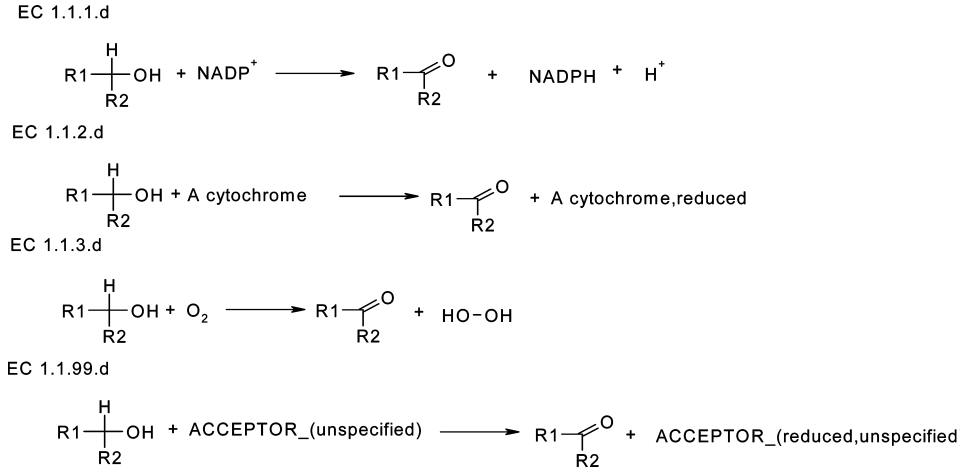

Figure 1. Reactions catalyzed by enzymes of subsubclasses of EC 1.1.c.d.

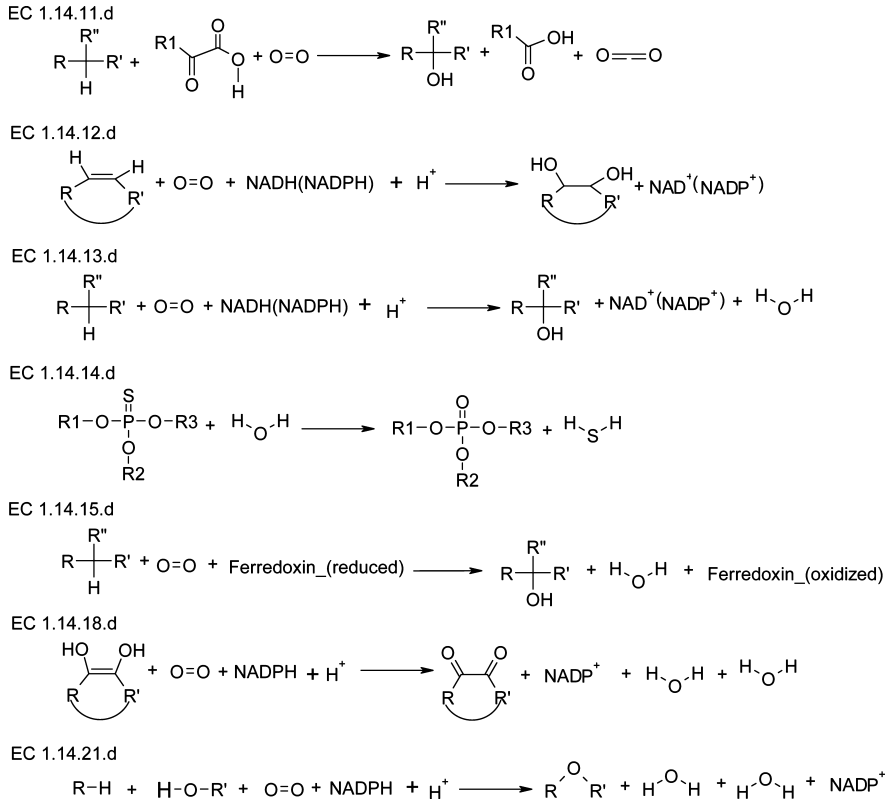

Figure 2. Reactions catalyzed by enzymes of subsubclasses EC 1.14.11.d, EC 1.14.12.d, EC 1.14.13.d, EC 1.14.14.d, EC 1.14.15.d, EC 1.14.18.d, and EC 1.14.21.d.

In making these choices, all major electronic effects influencing reaction mechanisms, such as charge distribution, inductive, resonance, and polarizability effects were considered.

Before training, the input data (the selected descriptors) were scaled into a [0.1, 0.9] range through the formula:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \times 0.8 + 0.1 \qquad (1)$$

where x was the original value, x^* is the scaled value, and $x_{\min}$ and $x_{\max}$ are the corresponding minimum and maximum values of the descriptor variable, respectively.

2.4. Kohonen's Self-Organizing Neural Network (KohNN). In this study, Kohonen's self-organizing neural network (KohNN) was used as one method to investigate the similarity of enzyme-catalyzed reactions. It is a neural network model introduced by Kohonen for the construction of a nonlinear projection of objects from a high-dimensional space into a lower-dimensional space.[22] The similarity perception of objects is an essential feature of a KohNN. In a KohNN, the neurons are arranged in a 2D array to generate a 2D Kohonen map such that similarity in the data is preserved. In other words, if two input data vectors are similar, then they will be mapped into the same neuron or neurons near to each other in the Kohonen map. The KohNN network is composed of a 2D arrangement of neurons. Each neuron has a number of weights equal to the number of descriptors representing the reacting bonds (input vector). In the training process, the input vectors are presented to the network, and the neuron having weights most similar to the values of the descriptors of reacting bonds will receive the considered reaction. A weight adjustment process is then initiated.

Represented by the six descriptors as described in Section 2.3, the breaking or changing of each bond is an event in a 6D space, spanned by the six descriptors as coordinates for each bond. In order to determine the similarity among the reactions, each reaction is projected into a 2D plane using a KohNN. The software used for the generation of the Kohonen maps was SONNIA.[23,24] For comparison with the classical EC nomenclature, each neuron in the 2D map was colored by the reactions belonging to a specific EC subclass.

2.4. Support Vector Machine (SVM). A SVM[25] is a useful technique for data classification. A number of excellent introductions into SVM are available.[26−28] The SVM method originated as an implementation of Vapnik's structural risk minimization (SRM) principle from statistical learning theory. The input vectors for SVM are mapped into a higher (maybe infinite) dimensional space. Then, an SVM searches for a linear separating hyperplane with the maximal margin in this higher dimensional space so that an SVM can classify the data into several classes.

In this study, the LIBSVM developed by Chang and Lin[29] was used for SVM analysis. It supports multiclass classification by decomposing a multiclass problem into a number of binary problems. LIBSVM uses "one-against-one" approach to construct $k(k-1)/2$ classifiers. Each classifier trains data from two different classes. The penalty parameter C of SVM is selected by the user. The commonly used kernel is the radial basis function (RBF) kernel (eq 2). It is used to convert the data into a higher-dimensional space. The parameter C (eq 3) and γ were chosen by the autosearching program "grid" through a five-

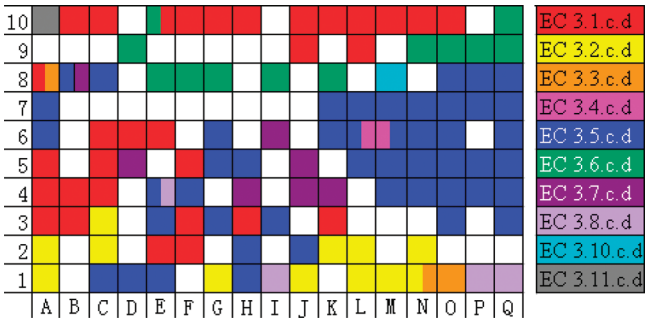

Figure 3. A rectangular Kohonen map of most frequent occupation for the 311 reactions of hydrolases EC 3.b.c.d. Each color represents 1 of 10 subclasses. Some neurons containing reactions catalyzed by enzymes of two different subclasses are marked in two colors. The neurons in white were not occupied by any reaction. Each of the other neurons is occupied by reactions catalyzed by hydrolases of the same subclass.

fold cross-validation method. Here, six physicochemical effects of reacting bonds are used as input vectors.

$$k(x, y) = \exp(-\gamma \|x - y\|^2) \tag{2}$$

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} W^T W + C \sum_{i=1}^{l} \xi_i \\ \text{subject to} \quad & y_i(W^T \varphi(X_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{3}$$

2.5. Hierarchical Clustering Analysis (HCA). Hierarchical clustering analysis (HCA) organizes cases based on the similarity or dissimilarity of selected characteristics.[30] This method starts with a set of distinct cases, each of which is considered as a separate cluster. Two clusters that are closest to each other according to some metric are agglomerated. This is repeated until all the cases belong to one hierarchically constructed cluster. Cases with similar characteristics will be clustered together as neighboring rows in a dendrogram.

In this work, the program PermutMatrix[31] was applied to investigate the similarity of the enzymatic reactions according to the six physicochemical properties of each reacting bond. After all the methods were tried, the Euclidean distance for dissimilarity and the average linkage method were selected in order to get the best results. Reactions with similar physicochemical properties were classified in a cluster. The results were compared with the EC classification system.

3. RESULTS AND DISCUSSION

3.1. Similarity Perception Results of Class EC 3.b.c.d (hydrolases) by Different Classification Methods. *3.1.1. Similarity Perception Results of EC 3.b.c.d.* First, a KohNN method was applied. All 311 reactions catalyzed by hydrolases (EC 3.b.c.d) were projected into a planar 17 × 10 rectangular Kohonen map. The resulting map is shown in Figure 3. The coloring of the neurons is on the basis of subclasses b (EC 3.b.c.d) for a comparison with the EC system. The map shows a clear separation of the individual subclasses, although the areas of the individual subclasses do not always form a single cluster. Nevertheless, the classification accuracy of KohNN is very high. A classification accuracy of 95.8% was achieved on the basis of the most frequent occupation of a neuron.

For this larger data set, using a KohNN provided results similar to those of the previous publications.[9] For the six subclasses EC 3.1.c.d, EC 3.2.c.d, EC 3.5.c.d, EC 3.6.c.d, EC 3.7.c.d, and EC 3.8.c.d, a good separation was achieved, quite similar to the first publication.[9] The three reactions of subclass EC 3.4.c.d, which were not separated from EC 3.5.c.d in the previous publication,[9] were also here projected into neurons together with some reactions of subclass EC 3.5.c.d.

In this work, three more subclasses (EC 3.3.c.d, EC 3.10.c.d, and EC 3.11.c.d) than in the previous publication[9] were investigated. There was only one reaction in subclass EC 3.10.c.d, and it was located in a neuron of its own; the two reactions of subclass EC 3.11.c.d were both projected into a single neuron, as shown in Figure 3. Two reactions of subclass EC 3.3.c.d (hydrolases acting on ether bonds, containing four reactions) were projected into a neuron together with some reactions of subclass EC 3.1.c.d (hydrolases acting on ester bonds). The other two reactions of EC 3.3.c.d were correctly classified. The two wrong-classified reactions, catalyzed by enzyme EC 3.3.2.5, are shown in Figure 4. According to the EC system, enzymes catalyzing reactions from subclass EC 3.3.c.d should act on ether bonds, however, here the enzyme acts on ester bonds. Thus, a misclassification of the EC system was discovered with our method.

A SVM was also used to classify the hydrolases. The data set of 311 reactions catalyzed by enzymes from EC 3.b.c.d was randomly divided into two sets: a training set containing 244 reactions and a test set containing 67 reactions. The two parameters of the SVM (C, γ) were selected using the autosearching program "grid" through five-fold cross-validation in LibSVM.[29] The training set was used to train a SVM model, the optimum parameters of $C = 128$ and $\gamma = 8$ were selected. The test set was used for prediction of subclasses of each reaction in the test set. With the training set, a classification accuracy of 97.5% was obtained, and the test set showed an accuracy of 98.5%. The similarity perception

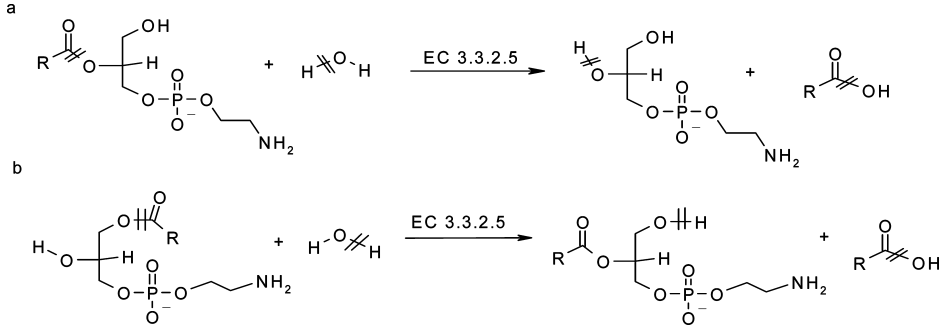

Figure 4. Two reactions catalyzed by enzyme EC 3.3.2.5, located in one neuron together with some reactions of EC 3.1.c.d.

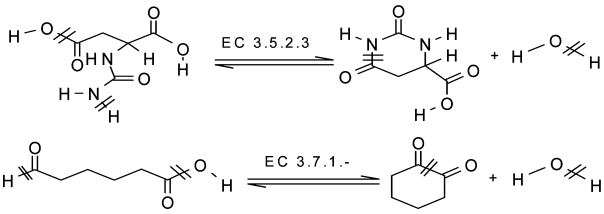

Figure 5. Comparison of the two reactions catalyzed by enzyme EC 3.5.2.3 and EC 3.7.1.-. These reactions have been stored in reverse reaction direction.

results of the SVM method were similar to those of the KohNN method.

Using the SVM method, a reaction catalyzed by enzyme EC 3.5.2.3 (acting on carbon−nitrogen bonds) was incorrectly predicted into subclass EC 3.1.c.d; while with the KohNN method, this reaction was projected into a neuron together with a reaction catalyzed by enzyme EC 3.7.1.- (acting on carbon−carbon bonds). A closer inspection of this relationship shows that by mistake both reactions have been input in reverse direction into BioPath eliminating a water molecule. As shown in Figure 5, the encoding of these reactions is, therefore, based on the breaking of a carboxylic acid bond. Similar physicochemical effects can be found in reactions that are catalyzed by hydrolases of class EC 3.1.c.d. Thus, by using our method, inconsistencies in the storage of the reaction direction of hydrolysis reactions became obvious. In the meantime, the reactions have been corrected in the database.

A HCA was also applied to classify hydrolases. The detailed results are shown in Figure S1 in the Supporting Information. In order to give an overview of the HCA results, Figure 6 shows the classification of the reactions of EC class EC 3.b.c.d into subclasses and subsubclasses. The results were quite similar to those using the KohNN and the SVM methods. This indicates that the selected six physicochemical effects are efficient in representing the reactions catalyzed by enzymes of class EC 3.b.c.d.

It must be realized that reactions of the individual subclasses are not completely clustered together. In fact, the hierarchical clustering in Figure 6 provides much more valuable information than a simple classification into subclasses. It shows finer details in the relationships between enzymatic reactions. Here, just a few results will be discussed.

The three enzyme-catalyzed reactions at the bottom of Figure 6 (EC 3.2.2.d, EC 3.5.4.d, and EC 3.5.99.1) are apparently quite different from the rest of the hydrolases. Further up the clustering of Figure 6, the similarity of quite a few reactions catalyzed by enzymes of subclasses EC 3.1.c.d and EC 3.6.c.d is indicated. Although reactions catalyzed by enzymes of subclass EC 3.5.c.d form a distinct cluster in the upper part of Figure 6, quite a few reactions catalyzed by enzymes of this subclass are distributed into several different clusters. Some reactions, such as those catalyzed by enzyme EC 3.5.5.d, EC 3.5.4.d and EC 3.5.1.99, stand out as being quite different from any other reaction.

This HCA based on physicochemical effects at the reaction site thus indicates relationships between enzyme-catalyzed reactions that cannot be covered by such a simple classification as the EC system. There is still a lot of interesting relationships and similarities to be discovered in the HCA

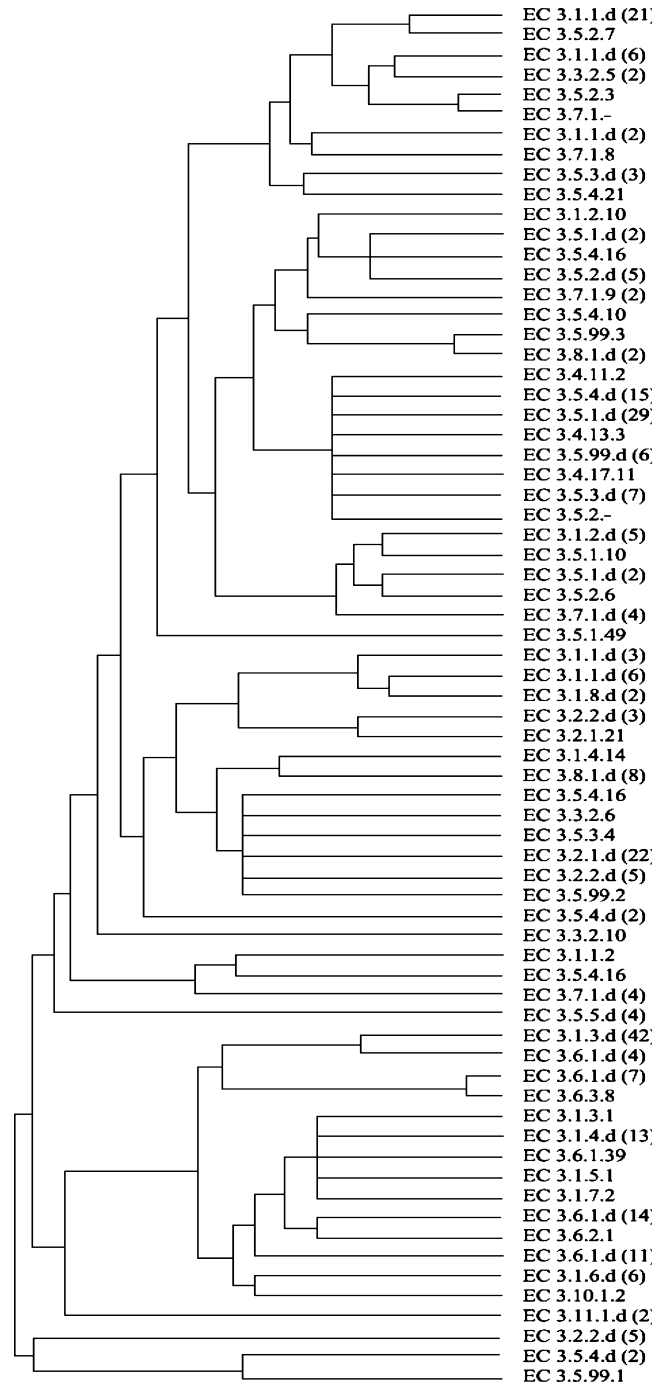

Figure 6. A rebuilt figure of HCA result of 311 reactions catalyzed by hydrolases from Figure S1 in the Supporting Information. The digit in the bracket indicates the number of reactions that are catalyzed by enzymes of the same subsubclass. These reactions were combined and represented by EC code with the fourth number "d".

of Figure 6. We, therefore, provide the data set that formed the basis of Figure 6 in the Supporting Information.

3.1.2. Similarity Perception Results of EC 3.1.c.d. All 112 reactions of subclass EC 3.1.c.d were explored based on the trained Kohonen map of Figure 3. The coloring of the neurons is based on the subsubclass c (EC 3.1.c.d), as shown in Figure 7. Most of the subsubclasses show a good separation, although the subsubclasses EC 3.1.1.d, EC 3.1.2.d, and EC 3.1.4.d do not always form a single cluster. On the basis of the most frequent occupation, a classification accuracy of about 99% was achieved.

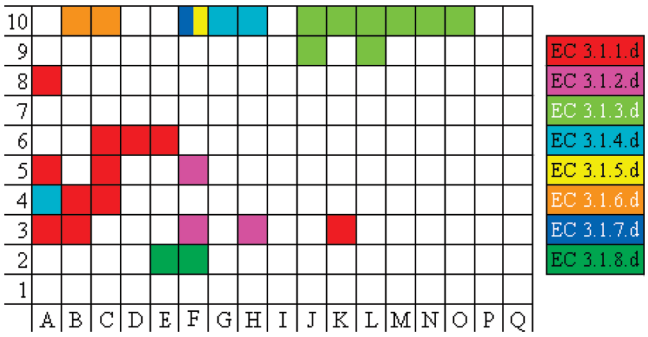

Figure 7. A rectangular Kohonen map for subclass EC 3.1.c.d. Coloring is based on the subsubclass c. The conflict neuron, which is occupied by reactions catalyzed by enzymes of different subsubclasses, is marked in two different colors, such as neuron F10. Each of the other colored neurons is occupied by reactions catalyzed by enzymes of the same subsubclass.

Table 3. Parameters of SVM Training for Four Selected Subclasses EC 3.1.c.d, EC 3.5.c.d, EC 1.1.c.d, and EC 1.14.c.d

	parameters of SVM		training set		test set	
subclass	*C*	Y	reactions	accuracy	reactions	accuracy
EC 3.1.c.d	8	8	89	100%	23	95.6%
EC 3.5.c.d	128	16	73	95.9%	19	94.7%
EC 1.1.c.d	2	64	192	89%	50	82%
EC 1.14.c.d	2	1024	72	98.6%	20	90%

Before using the SVM method, the 112 reactions were randomly divided into a training (89 reactions) and a test (23 reactions) set. The optimum parameters of SVM model are shown in Table 3. Classification accuracies of 100% for the training set and 95.6% for the test set were achieved.

With the HCA method, quite similar results were obtained.

3.1.3. Similarity Perception Results of EC 3.5.c.d. The 92 reactions of subclass EC 3.5.c.d were analyzed using the KohNN, SVM, and HCA methods. The resulting Kohonen map, which was obtained on the basis of Figure 3, is shown in Figure 8. The neurons were colored on the basis of subsubclass c (EC 3.5.c.d). A classification accuracy of 94.5% was achieved on the basis of the most frequent occupation of a neuron. Using an SVM, the 92 reactions were divided into a training (73 reactions) and a test (19 reactions) set. The parameters used for training are also shown in Table 3. Classification accuracies of 95.9% for the training set and 94.7% for the test set were achieved.

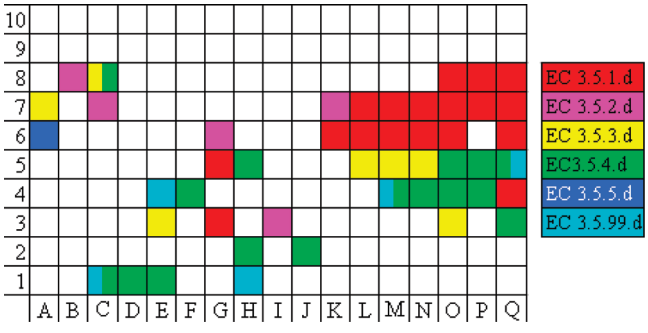

Figure 8. A rectangular Kohonen map for subclass EC 3.5.c.d. Coloring is based on the subsubclass c. The conflict neuron, which is occupied by reactions catalyzed by enzymes of different subsubclasses, is marked in two different colors, such as neuron C1. Each of the other colored neurons is occupied by reactions catalyzed by enzymes of the same subsubclass.

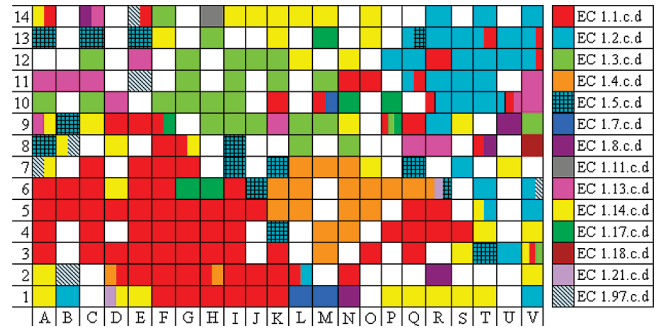

Figure 9. A rectangular Kohonen map of most frequent occupation for class EC 1.b.c.d. Neurons are colored based on the subclass b (EC 1.b.c.d). Some neurons containing reactions catalyzed by enzymes of two different subclasses are marked in two colors. Some neurons containing reactions catalyzed by enzymes of three different subclasses are marked in three colors. The white neurons are not occupied by any reaction. The other neurons are occupied by reactions catalyzed by enzymes of the same subclass.

3.2. Similarity perception results of class EC 1.b.c.d (oxidoreductases) by different classification methods.

3.2.1. Similarity perception results of EC 1.b.c.d. 651 reactions catalyzed by oxidoreductases (EC 1.b.c.d) were analyzed using KohNN. Six physicochemical properties corresponding to each reacting bond were used as input data. Based on these descriptors, the reactions were projected into a planar 22 × 14 rectangular Kohonen map. The resulting Kohonen map was produced indicating the most frequent occupation, as shown in Figure 9. The neurons were colored based on subclass b (EC 1.b.c.d) for a comparison with the EC system. For the subclass EC 1.4.c.d, the reactions are concentrated in a single area, whereas the reactions of all other subclasses, especially EC 1.14.c.d, are projected into several clusters. In this case, a classification accuracy of 93.4% was achieved on the basis of the most frequent occupation of a neuron.

It can be seen from Figure 9 that the subclasses were classified very well, especially the subclasses EC 1.1.c.d, EC 1.2.c.d, EC 1.3.c.d, and EC 1.4.c.d were clearly separated. Some reactions were projected into some conflict neurons. Here, some wrong-classified reactions will be discussed in detail.

Four reactions catalyzed by enzymes of subclass EC 1.1.c.d (acting on a CH−OH group as donor) are located in neurons with some reactions catalyzed by enzymes of subclass EC 1.4.c.d (acting on a CH−NH_2 group as donor). For example, the reaction catalyzed by EC 1.1.3.12 (the reaction "a" in Figure 10) is located in the same neuron as the reaction that is catalyzed by EC 1.4.3.5. According to the EC system, enzyme EC 1.4.3.5 should catalyze a reaction taking a CH−NH_2 group as donor. However, this reaction takes a CH−OH group as donor, as reaction "b" in Figure 10 indicates. Based on the analysis of the reaction center, these reactions a and b are quite similar and, therefore, should be grouped together. The reaction annotated by EC 1.4.3.5 should rather be grouped into subclass EC 1.1.c.d.

One reaction (a, in Figure 11) catalyzed by 1-pyrroline-5-carboxylate dehydrogenase is located in a neuron with some reactions of subclass EC 1.2.c.d (b, in Figure 11). As a multifunctional enzyme, it catalyzes most reactions having a CH−NH group as donor. Thus, for the EC system, this enzyme is giving the EC code EC 1.5.1.12 (acting on a CH−NH group as donor). Here, this enzyme also catalyzes

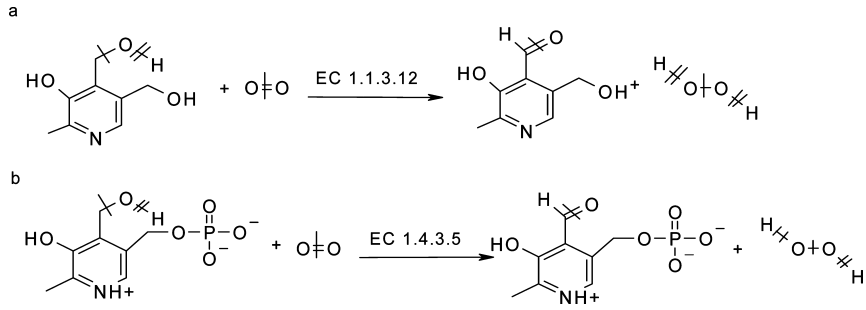

Figure 10. Comparison of the two reactions located in a single neuron. Both reactions have the same reaction center.

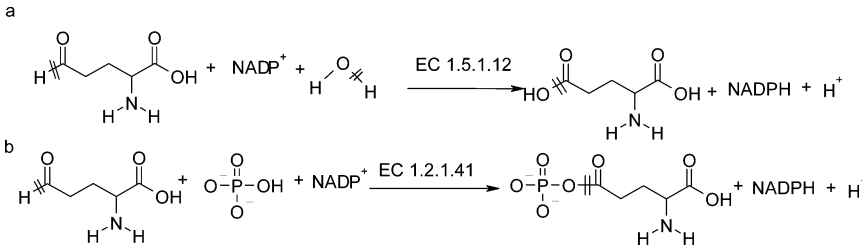

Figure 11. Comparison of two reactions located in a single neuron. Both reactions have an aldehyde group as donor.

a reaction (a, in Figure 11) taking an aldehyde group as donor, the same as those enzymes of subclass EC 1.2.c.d. Thus, the grouping together of these reactions is supported. For the type of multifunctional enzymes, the EC code may not be sufficient to represent the type of reactions they catalyzed. However, it can be found that the physicochemical descriptors work in classifying the reactions.

Several neurons in the Kohonen map were occupied by some reactions that belong to different subclasses. Three reactions catalyzed by enzyme EC 1.2.3.8, EC 1.1.1.107, and EC 1.13.11.30, respectively, are located in the same neuron. Although these reactions are catalyzed by enzymes of different subclasses of the EC system, all of them take an aldehyde group as donor, and thus the physicochemical effects of the reacting bonds in these reactions are very similar. The location of these reactions gives the suggestion that they should be rechecked in the EC system.

For a similarity perception by a SVM method, all 651 reactions catalyzed by oxidoreductases (EC 1.b.c.d) were randomly divided into a training (511 reactions) and a test (140 reactions) set. The training set was used to train an SVM model with the optimum parameters $C = 32$ and $\gamma = 256$, and the test set was used for the prediction of subclasses of each reaction in the test set. Classification accuracies of 98% for the training set and 90% for the test set were achieved.

Using an SVM, similar results to those by a KohNN were achieved. For example, one reaction catalyzed by the enzyme EC 1.5.1.12 was grouped into subclass EC 1.2.c.d (see Figure 11). Another reaction catalyzed by an enzyme EC 1.4.3.5 was predicted into subclass EC 1.1.c.d (see Figure 10). Thus, in both cases, the wrong classification of the EC code was also detected.

As shown above, for 651 reactions catalyzed by oxidoreductases (EC 1.b.c.d), both the KohNN and the SVM methods achieved an accuracy of over 90%, and the results of an SVM are similar to those by a KohNN method. This indicates that these six physicochemical properties of reacting bonds are effective in describing these reactions, robustly responding to different classification methods.

A HCA was also applied to analyze reactions of oxidoreductases (EC 1.b.c.d). The detailed results of HCA are shown in Figure S2 of the Supporting Information. The clustering of the 651 reactions into subclasses and subsubclasses is shown in Figure 12. Using an HCA, results similar to those by a KohNN and SVM method were achieved. However, the comparison with the KohNN results shows the advantage of using a 2D similarity perception against the 1D classification method of a HCA. On the one hand, the direction in a 2D KohNN map can indicate different kinds of similarities and, on the other hand, the distance can represent the degree of similarity. Nevertheless, as discussed with Figure 6, the HCA shows a much more detailed similarity in enzyme-catalyzed reactions than can be given by a simple classification into subclasses and subsubclasses, as put down in the EC classification method. Closer inspection of Figure 12 allows one to discover interesting relationships between reactions catalyzed by enzymes of different subclasses.

As a case in point, one discovery will be mentioned here. One reaction (reaction "a" in Figure 13), which is catalyzed by enzyme EC 1.1.3.24, was clustered into subclass EC 1.3.c.d (oxidoreductases acting on a CH−CH group as donor). This enzyme also catalyzes another reaction (reaction "b" in Figure 13), which was clustered into subclass EC 1.1.c.d. As shown in Figure 13, this enzyme takes both CH−CH and CH−OH group as donors, and therefore, the classification of these two reactions into different subclasses is warranted. On the basis of the EC system, the EC code of this enzyme now has been changed to EC 1.3.3.12. This provides proof that our classification method is powerful.

In the following, the similarity perception of subclasses EC 1.1.c.d and EC 1.14.c.d will be discussed in detail.

3.2.2. Similarity perception results of EC 1.1.c.d. All the 242 reactions of subclass EC 1.1.c.d (oxidoreductases acting on a CH−OH group as donor) were projected into the trained Kohonen map of Figure 9. The resulting map, as shown in Figure 14, was colored on the basis of the subsubclass c

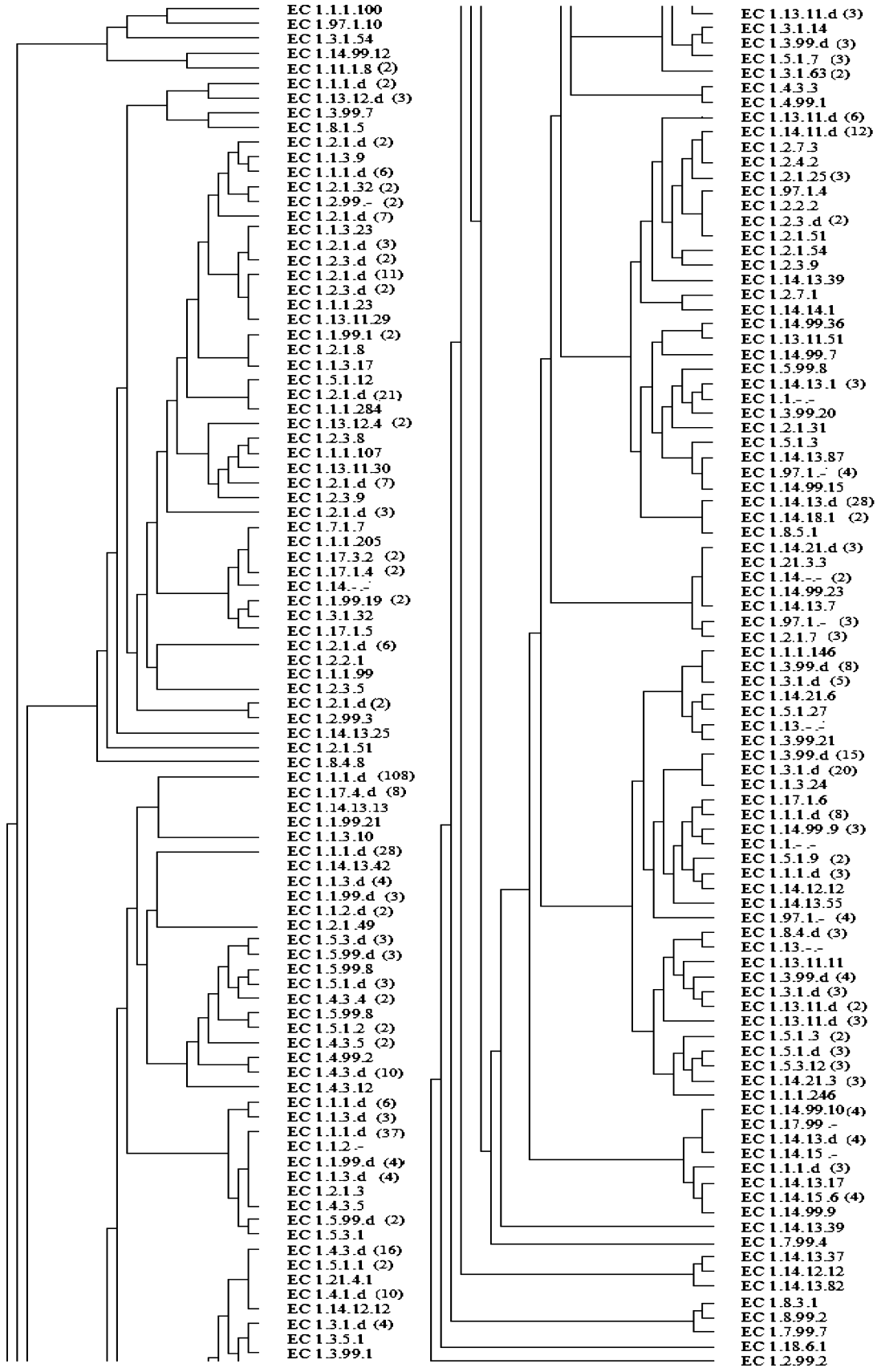

Figure 12. A rebuilt figure of HCA results of 651 reactions catalyzed by oxidoreductases from Figure S2 in the Supporting Information. The dendrogram is split into two parts, starting from the left side and continuing on the right side. The digit in the bracket indicates the number of reactions that are catalyzed by enzymes of the same subsubclass. These reactions were combined and represented by EC code with the fourth number "d".

(EC 1.1.c.d). A classification accuracy of 92.5% on the basis of the most frequent occupation was achieved.

There are two reactions catalyzed by an enzyme of which the EC code was not fully assigned: EC 1.1.-.-. Using a

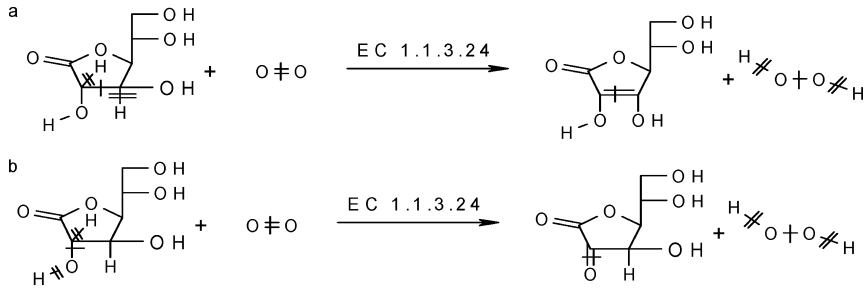

Figure 13. Two reactions catalyzed by enzyme EC 1.1.3.24. Reaction "a" was clustered into subclass EC 1.3.c.d, and reaction "b" was clustered into subclass EC 1.1.c.d.

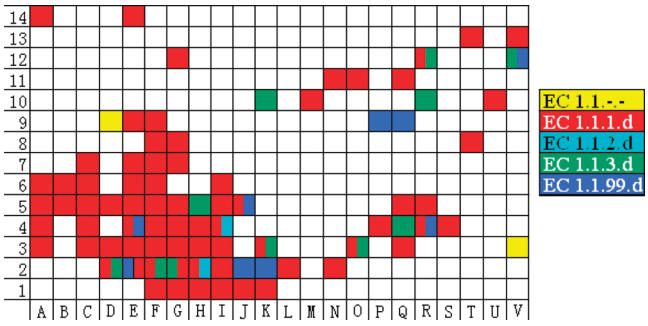

Figure 14. A rectangular Kohonen map for subclass EC 1.1.c.d. Neurons were colored based on the subsubclass c (EC 1.1.c.d). The conflict neuron, which is occupied by reactions catalyzed by enzymes of different subsubclasses, is marked in two different colors, such as neuron D2. Each of the other colored neurons is occupied by reactions catalyzed by enzymes of the same subsubclass.

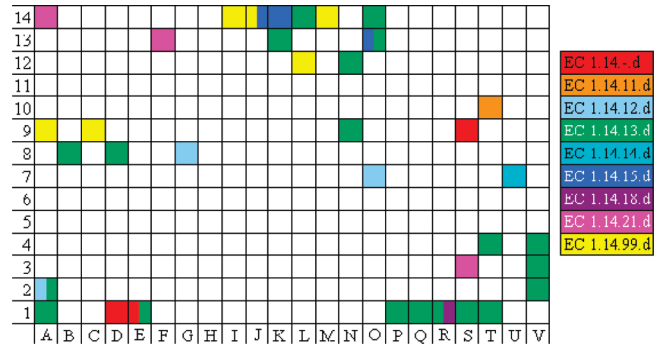

Figure 15. A rectangular Kohonen map for subclass EC 1.14.c.d. The coloring of the neurons is based on subsubclass c (EC 1.14.c.d). The conflict neuron, which is occupied by reactions catalyzed by enzymes of different subsubclasses, is marked in two different colors, such as neuron A2. Each of the other colored neurons is occupied by reactions catalyzed by enzymes of the same subsubclass.

KohNN method, the two reactions were projected into two neurons D9 and V3 separately, as shown in Figure 14. Thus, it is difficult to guess the third number of the EC code.

Subsubclasses of most subclasses of EC 1.b.c.d were defined in the EC system on the basis of the respective acceptors. For example, enzymes of EC 1.1.1.d take NAD^+ or $NADP^+$ as acceptors; enzymes of EC 1.1.2.d take a cytochrome or a protein as acceptors, enzymes of EC 1.1.3.d take oxygen as acceptors, etc. As shown in Figure 14, reactions of subsubclass EC 1.1.2.d were projected into neurons together with some reactions of subsubclass EC 1.1.1.d. Some reactions of subsubclass EC 1.1.3.d and EC 1.1.99.d were also located in neurons together with some reactions catalyzed by EC 1.1.1.d. Although these reactions have different kinds of acceptors, the reacting bonds of the substrates are ever similar, and thus they have similar properties. Distinguishing between subsubclasses of EC 1.1.c.d would require a description of the acceptor, e.g., by physicochemical data, such as the reduction potential.

Using an SVM, the 242 reactions were divided into a training (192 reactions) and a test (50 reactions) set. The parameters used for training are also shown in Table 3. Classification accuracies of 89% for the training set and 82% for the test set were achieved. Using an SVM, most reactions of subsubclasses EC 1.1.2.d, EC 1.1.3.d, and EC 1.1.99.d were predicted into subsubclass EC 1.1.1.d. Using a HCA, similar results were achieved.

3.2.3. Similarity Perception Results of EC 1.14.c.d. The 92 reactions of subclasses EC 1.14.c.d (oxidoreductases acting on paired donors, with incorporation or reduction of molecular oxygen) were projected into the trained Kohonen map of Figure 9. The resulting map of subclasses is shown in Figure 15, whose neurons were colored on the basis of subsubclass c (EC 1.14.c.d). A classification accuracy of 91.3% was achieved on the basis of the most frequent occupation.

From Figure 15, it can be seen that reactions of different subsubclasses were clearly separated, although some reactions were projected into conflict neurons.

There are only two reactions catalyzed by enzymes of subsubclass EC 1.14.18.d. Both reactions were projected into the same neuron R1 together with five reactions catalyzed by an enzyme of subclass EC 1.14.13.d (with NADH or NADPH as one donor). On the basis of the EC classification system, enzymes of subsubclass EC 1.14.18.d act on paired donors, with another compound as one donor. Here both reactions (a and b in Figure 16) also take NADPH as one donor. The similarity of the reaction centers of these reactions supports their grouping together.

There are three subsubclasses taking NADH or NADPH as one donor: EC 1.14.12.d, EC 1.14.13.d, and EC 1.14.21.d. The other donors of these three subsubclasses are different: subsubclass EC 1.14.12.d takes two atoms of oxygen as the other donor, subsubclass EC 1.14.13.d takes one atom of oxygen as the other donor, and the other donor of subsubclass EC 1.14.21.d is dehydrogenated. As shown in Figure 15, most of the reactions catalyzed by enzymes of these three subsubclasses were separated clearly, except one reaction. The reaction catalyzed by naphthalene 1,2-dioxygenase (EC 1.14.12.12) was projected into the neuron A2 together with another reaction catalyzed by an enzyme of subsubclass EC 1.14.13.d. It was found that naphthalene 1,2-dioxygenase is a multifunctional enzyme catalyzing several different reactions. Here, this reaction, where a hydroxyl group is oxidized to a carbonyl group, is quite different from the other reactions catalyzed by this enzyme,

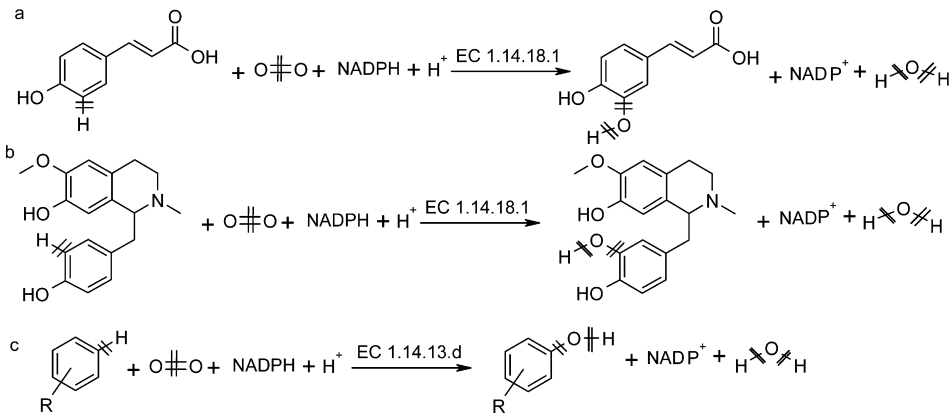

Figure 16. Reactions catalyzed by enzymes of subsubclass EC 1.14.13.d and EC 1.14.18.d.

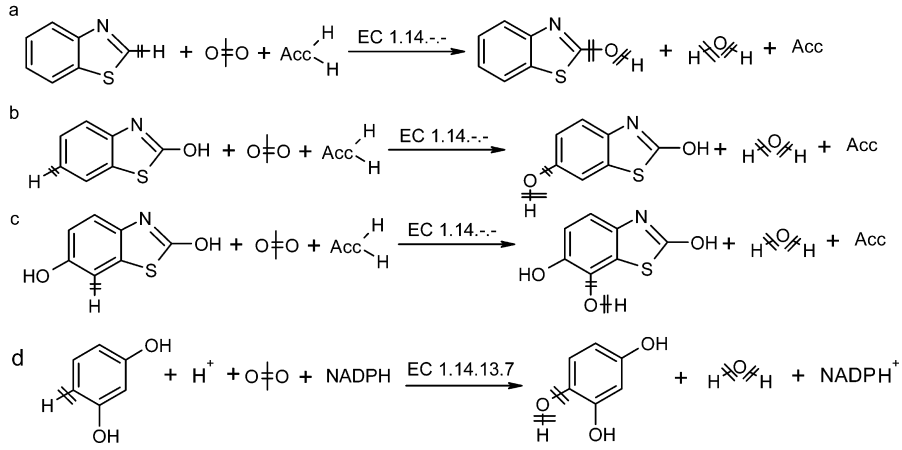

Figure 17. Comparison of reactions catalyzed by enzyme EC 1.14.-.- to a reaction catalyzed by enzyme EC 1.14.13.7.

where a carbon−carbon double bond is oxidized to diol. Thus, the separation of this reaction was supported.

In this data set, there are three reactions catalyzed by enzymes having an EC code that is not fully assigned: EC 1.14.-.-. Two of the three reactions were projected into two neurons separately, as shown in Figure 15. The other reaction (b, in Figure 17) was projected into neuron E1 together with another reaction (d, in Figure 17) catalyzed by enzyme EC 1.14.13.7 of subsubclass EC 1.14.13.d. As shown in Figure 17, these reactions have a very similar reaction center, although one donor of reaction a, b, and c is not specified.

For an SVM analysis, the 92 reactions were divided into a training (72 reactions) and a test (20 reactions) set, and the parameters used for training are also shown in Table 3. Classification accuracies of 98.6% for the training set and 90% for the test set were achieved. Most reactions were classified correctly, except three reactions. Two of the three reactions are catalyzed by enzyme EC 1.14.18.1, and they were classified into the subsubclass of reactions catalyzed by enzymes of EC 1.14.13.d. The other reaction, catalyzed by enzyme EC 1.14.-.- (b, in Figure 17), was classified into subclass EC 1.14.13.d. These results are similar to those achieved using a KohNN method. Using a HCA, similar results were also achieved.

3.2.4. Similarity Perception Results of Reactions Taking NAD$^+$ or NADP$^+$ as Acceptors. Enzymes of EC 1.b.1.d take either NAD$^+$ or NADP$^+$ as acceptors. Most reactions catalyzed by enzymes taking the same acceptor were projected close to each other, although some reactions taking different acceptors were located in the same neurons. Some examples follow: A reaction catalyzed by EC 1.1.1.10 (taking NADP$^+$ as acceptor) was projected into the same neuron as the reaction catalyzed by EC 1.1.1.13 (taking NAD$^+$ as acceptor). A reaction catalyzed by EC 1.3.1.2 (taking NADP$^+$ as acceptor) was projected into the same neuron as the reaction catalyzed by EC 1.3.1.1 (taking NAD$^+$ as acceptor). Apparently, reactions that are quite similar based on the reaction center but have different acceptors can only be distinguished if bonds in the acceptors are also included in the analysis.

Some enzymes can take both NAD$^+$ and NADP$^+$ as acceptors. For example, enzyme EC 1.1.1.1 can catalyze two reactions taking NAD$^+$ (or NADP$^+$) as acceptors. Both reactions were projected into the same neuron. Thus, it is difficult to separate the reactions catalyzed by enzymes taking NAD$^+$ or NADP$^+$ as acceptors.

4. CONCLUSIONS

In this work, the similarity of reactions catalyzed by hydrolases (EC 3.b.c.d) and oxidoreductases (EC 1.b.c.d) were investigated using the Kohonen's self-organizing neural network (KohNN), the support vector machine (SVM), and the hierarchical clustering analysis (HCA) methods, while representing the reactions with physicochemical effects of the reacting bonds. Similar and good similarity perception results were achieved using all these three different methods. Using a KohNN method, the similarity of the reactions catalyzed by enzymes of different subclasses or subsubclasses is represented in a Kohonen map, thus, it is easy to get the information of the similarity of the reactions. Using a SVM, classification results with high accuracy

were achieved. These results were similar to those achieved using a KohNN, and the KohNN method was supported. Using a HCA method, the hierarchical relationships between the reactions were given in the dendrogram, which is another representation of the similarity of the enzymatic reactions.

The similar results achieved using these three methods indicate that the selected six physicochemical properties represented the reactions very well. The similarity perception results on the basis of physicochemical properties are also coherent with those of the EC classification system. This is true for hydrolases (EC 3.b.c.d) because with this class of enzyme the EC code considers the nature of the reaction center. And it is also true for oxidoreductases (EC 1.b.c.d) because in this case the EC code considers the nature of the substrate and thus implicitly the reacting bonds. However, the methods presented here provide a much more detailed analysis of the similarity of enzyme-catalyzed reactions than can ever be given by a simple classification system, such as the EC code. In particular, the investigation of reactions based on the physicochemical effects of the reacting bonds shows similarities and differences in the reactions that go beyond the phenomenological classification of the EC system.

The method also allows the identification of EC numbers that are wrongly or, at least, very unfortunately assigned.

It is clear that the method presented here will much less correspond with the EC code in those cases where the EC code is less concerned, either explicitly or implicitly, with the nature of the reacting bonds. However, we do believe that a more detailed analysis of enzyme-catalyzed reactions should focus on the very nature of the reacting bonds and the physicochemical effects that make these bonds reactive.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (20605003 and 20975011) and the Chinese Universities Scientific Fund (ZZ0911) of Beijing University of Chemical Technology.

Supporting Information Available: Similarity perception results of reactions catalyzed by hydrolases (EC 3.b.c.d, 311 reactions, as shown in Figure S1) and oxidoreductases (EC 1.b.c.d, 651 reactions, as shown in Figure S2) by Hierarchical Clustering Analysis (HCA). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes by the Reactions they Catalyze. Enzyme Nomenclature; Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB): London, U.K.; http://www.chem.qmul.ac.uk/iubmb/enzyme/. Accessed March 1, 2010.
(2) Arita, M. The metabolic world of Escherichia coli is not small. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 1543–1547.
(3) Kotera, M.; Okuno, Y.; Hattori, M.; Goto, S.; Kanehisa, M. Computational Assignment of the EC Numbers for Genomic-scale Analysis of Enzymatic Reactions. *J. Am. Chem. Soc.* **2004**, *126*, 16487–16498.
(4) Zhang, Q. Y.; Aires-de-Sousa, J. Structure-based classification of chemical reactions without assignment of reaction centers. *J. Chem. Inf. Model.* **2005**, *45*, 1775–1783.
(5) Latino, D. A. R. S.; Zhang, Q. Y.; Aires-de-Sousa, J. Genome-scale classification of metabolic reactions and assignment of EC numbers with self-organizing maps. *Bioinformatics* **2008**, *24*, 2236–2244.
(6) Chen, L.; Gasteiger, J.; Rose, J. R. Automatic Extraction of Chemical Knowledge from Organic Reaction Data: Addition of Carbon-Hydrogen Bonds to Carbon-Carbon Double Bonds. *J. Org. Chem.* **1995**, *60*, 8002–8014.
(7) Rose, J. R.; Gasteiger, J. HORACE: An Automatic System for the Hierarchical Classification of Chemical Reactions. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 74–90.
(8) Chen, L.; Gasteiger, J. Knowledge Discovery in Reaction Databases: Landscaping Organic Reactions by a Self-Organizing Neural Network. *J. Am. Chem. Soc.* **1997**, *119*, 4033–4042.
(9) Sacher, O.; Reitz, M.; Gasteiger, J. Investigations of Enzyme-Catalyzed Reactions Based on Physicochemical Descriptors Applied to Hydrolases. *J. Chem. Inf. Model.* **2009**, *49*, 1525–1534.
(10) Satoh, H.; Sacher, O.; Nakata, T.; Chen, L.; Gasteiger, J.; Funatsu, K. Classification of Organic Reactions: Similarity of Reactions Based on Changes in the Electronic Features of Oxygen Atoms at the Reaction Sites. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 210–219.
(11) O'Boyle, N. M.; Holliday, G. L.; Almonacid, D. E.; Mitchell, J. B. O. Using Reaction Mechanism to Measure Enzyme Similarity. *J. Mol. Biol.* **2007**, *368*, 1484–1499.
(12) Pegg, S. C. H.; Brown, S. D.; Ojha, S.; Seffernick, J.; Meng, E. C.; Morris, J. H.; Chang, P. J.; Huang, C. C.; Ferrin, T. E.; Babbitt, P. C. Leveraging Enzyme Structure-Function Relationships for Functional Inference and Experimental Design: The Structure-Function Linkage Database. *Biochemistry* **2006**, *45*, 2545–2555.
(13) Holliday, G. L.; Almonacid, D. E.; Bartlett, G. J.; O'Boyle, N. M.; Torrance, J. W.; Murray-Rust, P.; Mitchell, J. B. O.; Thornton, J. M. MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. *Nucleic Acids Res.* **2007**, *35*, D515–D520.
(14) Porter, C. T.; Bartlett, G. J.; Thornton, J. M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* **2004**, *32*, D129–D133.
(15) Nagano, N. EzCatDB: The Enzyme Catalytic-Mechanism Database. *Nucleic Acids Res.* **2005**, *33*, D407–D412.
(16) Reitz, M.; Sacher, O.; Tarkhov, A.; Truembach, D.; Gasteiger, J. Enabling the exploration of biochemical pathways. *Org. Biomol. Chem.* **2004**, *2*, 3226–3237.
(17) *Biochemical Pathways Wall Chart*, Michal, G., Ed.; Boehringer Mannheim (now Roche): Mannheim, Germany, 1993. It can also be accessed online at the Swiss Institute of Bioinformatics: Lausanne, Switzerland; http://www.expasy.org/tools/pathways/. Accessed March 1, 2010.
(18) Michal, G. *Biochemical Pathways - An Atlas of Biochemistry and Molecular Biology*; Spektrum Akademischer Verlag: Heidelberg, Germany, 1999.
(19) *BioPath.Explore*, version 1.1; Molecular Networks GmbH: Erlangen, Germany; http://www.molecular-networks.com. Accessed March 1, 2010.
(20) Gasteiger, J. Empirical Methods for the Calculation of Physicochemical Data of Organic Compounds. In *Physical Property Prediction in Organic Chemistry*; Jochum, C.; Hicks, M. G.; Sunkel, J., Eds.; Springer: Heidelberg, Germany, 1988, pp 119−138.
(21) Sacher, O. PhD Dissertation, University of Erlangen-Nuernberg: Erlangen, Germany, 2001.
(22) Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982**, *43*, 59–69.
(23) *SONNIA*, version 4.2; Molecular Networks GmbH: Erlangen, Germany; http://www.molecular-networks.com. Accessed March 1, 2010.
(24) Zupan, J., Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed; Wiley-VCH: Weinheim, Germany, 1999.
(25) Boser, B. E.; Guyon, I.; Vapnik, V. A training algorithm for optimal margin classifiers. In *Proceedings ACM Workshop on Computational Learning Theory*; Haussler, D., Ed.; ACM: New York, 1992; pp 144−152.
(26) Vapnik, V.; Chapelle, O. Bounds on error expectation for support vector machines. *Neural Comput. Appl.* **2000**, *12*, 2013–2036.
(27) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.
(28) Burges, C. J. C. A tutorial on Support Vector Machines for pattern recognition. *Data Min. Knowl. Discov.* **1998**, *2*, 121–167.
(29) Chang, C. C.; Lin C. J. *LIBSVM:a library for support vector machine*; Department of Computer Science and Information Engineering, National Taiwan University: Taipei, Taiwan; http://www.csie.ntu.edu.tw/. Accessed March 1, 2010.
(30) Olson, C. F. Parallel algorithms for hierarchical clustering. *Parallel Comput.* **1995**, *21*, 1313–1325.
(31) Caraux, G.; Pinloche, S. Permutmatrix: A Graphical Environment to Arrange Gene Expression Profiles in Optimal Linear Order. *Bioinformatics* **2005**, *21*, 1280–1281.

CI9004833