Investigations of Enzyme-Catalyzed Reactions Based on Physicochemical Descriptors Applied to Hydrolases

Oliver Sacher,[†] Martin Reitz,[‡] and Johann Gasteiger*[,†,‡]

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 August 8, 2008

The EC number system for the classification of enzymes uses different criteria such as reaction pattern, the nature of the substrate, the type of transferred groups or the type of acceptor group. These criteria are used with different emphasis for the various enzyme classes and thus do not contribute much to an understanding of the mechanisms of enzyme catalyzed reactions. To explore the reasons for bonds being broken in enzyme catalyzed metabolic reactions, we calculated physicochemical effects for the bonds reacting in the substrate of these enzymatic reactions. These descriptors allow the definition of similarities within these reactions and thus can serve as a method for the classification of enzyme reactions. To foster an understanding of the investigations performed here, we compare the similarities found on the basis of the physicochemical effects with the EC number classification. To allow a reasonable comparison we selected enzymatic reactions where the EC number system is largely built on criteria based on the reaction mechanism. This is true for hydrolysis reactions, falling into the domain of the EC class 3 (EC 3.b.c.d). The comparison is made by a Kohonen neural network based on an unsupervised learning algorithm. For these hydrolysis reactions, the similarity analysis on physicochemical effects produces results that are, by and large, similar to the EC number. However, this similarity analysis reveals finer details of the enzymatic reactions and thus can provide a better basis for the mechanistic comparison of enzymes.

1. INTRODUCTION

A widely accepted and used method for enzyme classification is the Enzyme Commission (EC) number, which was constituted in 1961, went through various versions and had its sixth edition in 1992. It is maintained by the *Nomenclature Committee of the International Union of Biochemistry and Molecular Biology* (NC-IUBMB).[1] This hierarchical classification system builds on a variety of criteria such as reaction patterns, substrates, transferred groups, and acceptor groups. In this system, a unique number is assigned to each enzyme, following a scheme consisting of four numbers separated by dots: a.b.c.d. The first number, a, gives the main class, consisting of six classes: oxidoreductases (1), transferases (2), hydrolases (3), lyases (4), isomerases (5), and ligases (6). These main classes are separated into several subclasses, b, again separated into sub-subclasses, c. The last number, d, is arbitrarily assigned to provide a unique number for each enzyme. For example: The enzyme glucose-6-phosphatase is classified as EC 3.1.3.9: hydrolases (a = 3), acting on ester bonds (b = 1), in this case a phosphoric monoester (c = 3).

Because of the diversity of criteria used at different levels of the classification scheme, the EC classification is not quite coherent as, depending on the EC class, the emphasis shifts between different criteria. For example in the second level b, the subdivision criteria for the oxidoreductases (EC 1) jump between the focus on electron donor or acceptor compounds, respectively. The transferases (EC 2) are subgrouped by the type of transferred group. The hydrolases (EC 3) are subgrouped by the type of bond that is hydrolyzed. Lyases (EC 4) are subgrouped by the type of bond which is broken. Isomerases (EC 5) focus on the type of isomerization, and the ligases (EC 6) are subgrouped by the bond which is formed during the reaction.

Clearly, the most important action of an enzyme is the catalysis of a reaction, an event that breaks and makes bonds. We therefore investigated enzyme-catalyzed reactions on the basis of the site where the reaction occurs, the reaction center. The reaction center was characterized by physicochemical effects calculated for the bonds, which take part in the reaction. This information can be used as input to data analysis methods. In our case, we used a self-organizing (Kohonen) neural network for grouping together all reactions considered similar on the basis of these properties. The results were then compared with the EC classification system.

In this paper, we restricted our studies to the class of hydrolases because these reactions have similar overall reaction patterns and the EC classification system also uses features based on the reaction type for the classification of this class. This provides an opportunity for the comparison of both classification systems.

The similarity analysis of enzyme catalyzed reactions can form the basis for searching for other enzymes that can catalyze a given reaction. Or, vice versa, for a given enzyme,

* To whom correspondence should be addressed. Tel: +49-9131-815668. Fax: +49-9131-815669. E-mail: gasteiger@molecular-networks.com. website: http://www.molecular-networks.com.
[†] Molecular Networks GmbH.
[‡] Universität Erlangen-Nürnberg.

10.1021/ci800277f CCC: $40.75
Published on Web 05/15/2009

different or novel reactions might be found that are catalyzed by the same enzyme.

The classification of enzymatic reactions is a field of great interest, as data on proteins with unknown function grow fast and the modeling of biochemical reaction networks depends on the knowledge of protein functions.[2] Izrailev and Farnum[3] used ligand-binding data of proteins with unknown function to classify them into the EC system by comparing the sets of processed ligands. Dobson and Doig[4] used several computable attributes from the crystal structure for this task. This bypasses the widely used method of protein function annotation by sequence comparison to other proteins with known function. Kotera et al.[5] 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 number (reaction classification) and compare this to the EC system. Zhang and Aires-de-Sousa[6] computed a Molmap for the reagent and product of each reaction using a set of physicochemical descriptors. From these Molmaps, a reaction Molmap can be computed, coding the reaction center. This information can then be used for an automatic classification of metabolic reactions.[7] Ruepp et al.[8] presented a classification system, FunCat, for the functional description of proteins, which is based on the pathway a protein is involved in and which is not restricted to enzymes and transporters as the EC system.

In contrast to our representation of the overall reaction, a series of research groups are working on the analysis and classification of the reaction mechanism of enzyme-catalyzed reactions: The research group of Babbitt et al. established the Structure Function Linkage Database (SFLD) that hierarchically classifies enzymes by linking the specific partial reactions with the conserved structural elements that mediate them.[9] MACiE[10] is a database that documents enzyme reaction mechanisms whereas the Catalytic Site Atlas (CSA)[11] is focusing on the active side 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 and is part of the EzCatDB database.[12] Especially for hydrolases, a hierarchical classification based on the catalytic site was reported by Gariev.[13] When our investigations had been terminated, a paper by O'Boyle et al.[14] appeared that investigated the reaction mechanisms for measuring enzyme similarity. Two approaches were presented, a fingerprint-based approach that incorporates information on each mechanistic step and an approach based on bond formation, cleavage, and changes in bond order.

Clearly, the kind of reactions that can be catalyzed by a given enzyme are influenced by two dominant factors, the shape of the active site and the physicochemical effects that are operative at this site. In this investigation, we concentrate on the elementary steps of a reaction, the bonds broken in a reaction and how this process can be characterized by physicochemical effects. This will then form the basis for the similarity analysis of enzymatic reactions. Thus, we neglected the size and the shape of the active site because often this information is not available.

2. MATERIALS AND METHODS

2.1. BioPath. The reactions that were used in this study were taken from the *BioPath* database which has been presented in detail in ref 15. This database was compiled from the information of the well-known *Biochemical Pathways* wall chart[16] and the corresponding atlas.[17] Clearly, more enzymatic reactions have been accumulated in recent years, but we confined ourselves to the content of this highly valuable source of information because we deemed it sufficient to explain the methodology. In this database, all compounds and reactions are stored as connection tables giving access to each atom and bond of a molecule. Reactions were input in a stoichiometrically correct manner and enriched by information such as enzyme name with EC number, regulators, reactant, and product name, as well as by information on the compartment where the reaction occurs. Important for the present application is the fact that all atoms of the substrates are mapped onto the corresponding atoms of the products by atom-atom mapping numbers and that all bonds which are broken, made, or altered during the reaction are marked. This information was derived by carefully analyzing each of the reactions by expert chemists. This feature of the *BioPath* database allows the extraction of the reaction center of each reaction and 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.[18] The *BioPath* database is publicly available at http://www.molecular-networks.com/biopath/index.html.

2.2. Data Sets. The *BioPath* database, version 1.0, consisting of 1545 reactions, was used as a basis for extracting reactions that are catalyzed by enzymes of class EC 3.b.c.d. This data set contains reactions that are catalyzed by hydrolases with subclasses ranging from EC 3.1.c.d to EC 3.8.c.d, except for EC 3.3.c.d, where no representative reactions were contained in the database. Three subclasses within this data set having the largest number of reaction instances were selected for a more detailed investigation: EC 3.1.c.d, EC 3.2.c.d, and EC 3.5.c.d.

After extraction from *BioPath*, the data set was preprocessed before the calculation of physicochemical properties. Reactions containing labels, R, for residues had the R's substituted by an H atom; otherwise the calculation of the physicochemical descriptors would fail as the methods require exact atom information for the calculation. Also, reactions occurring at macromolecules, for example, RNA, were removed, as this study is focused on small molecules. Furthermore, reversible reactions, where the hydrolyzed product was coded on the substrate side were inverted to match with the EC nomenclature for hydrolases, where the hydrolyzed compound is always considered on the product side.

After cleanup, six physicochemical descriptors (see section 2.3) were calculated for all reacting bonds using the program *PETRA*.[19] The reactions investigated were breaking one or two bonds in the substrate. The bond in the water molecule was not considered as this bond is the same in all hydrolysis reactions and thus provides no specific information for classification.

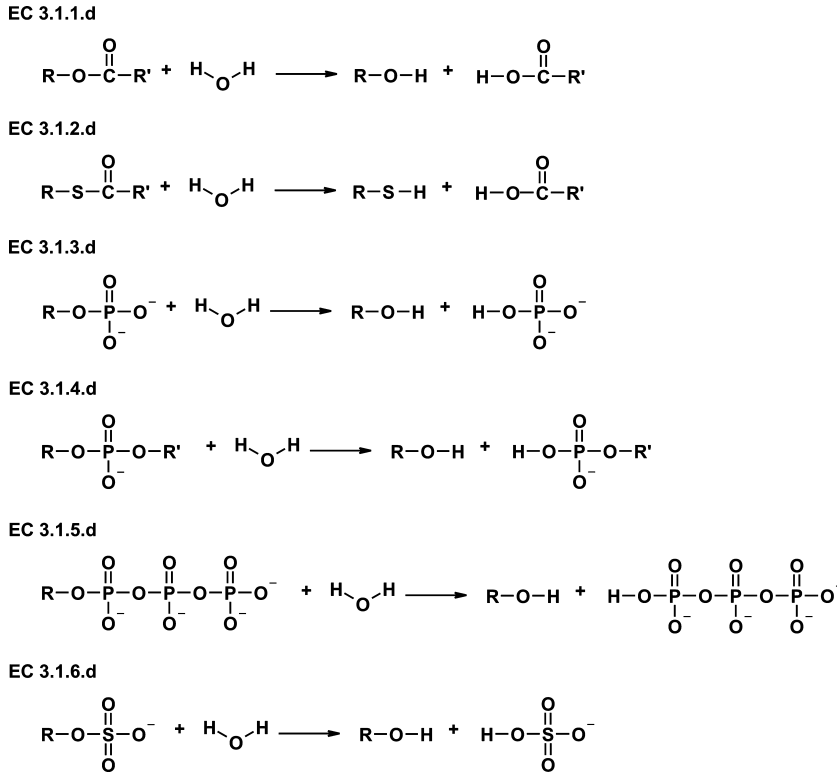

Figure 1. Reaction equations for the reactions catalyzed by the sub-subclasses of EC 3.1.c.d (hydrolases acting on ester bonds): EC 3.1.1.d (carboxylic ester hydrolases), EC 3.1.2.d (thioester hydrolases), EC 3.1.3.d (phosphoric monoester hydrolases), EC 3.1.4.d (phosphoric diester hydrolases), EC 3.1.5.d (triphosphoric monoester hydrolases), and EC 3.1.6.d (sulfuric ester hydrolases).

Data analysis methods such as Kohonen networks require the individual objects that are analyzed, in our case the reactions, to be represented by the same number of descriptors. Therefore, vectors of different length had to be filled up by standard values; here we chose values of zero. Some of the reactions in the data set of class EC 3.b.c.d are breaking two bonds. As each bond was characterized by six descriptors, all reactions in the entire data set were represented by a vector of length twelve. After calculation of the physicochemical descriptors, all values were scaled to better allow a comparison of all vectors.

EC 3.b.c.d. All reactions catalyzed by enzymes with the code EC 3.b.c.d (hydrolases) were extracted from the *BioPath* database. Hydrolases catalyze the hydrolysis of a variety of compounds such as esters, ethers, peptides, glycosides, phosphoric acid esters, acid anhydrides, or C−C bonds using water. The data set contained 135 reactions with one or two reacting bonds in the primary substrate, ignoring the bond in the water molecule involved in the reaction. The majority of reactions in the data set comprises reactions from the subclass EC 3.1.c.d, hydrolases acting on ester bonds, with 54 reactions. Subclass EC 3.2.c.d, the glycosylases, had a share of 19 reactions. There was no reaction of subclass EC 3.3.c.d (hydrolases acting on ether bonds) and the subclass EC 3.4.c.d, peptidases, comprised two reactions. A large part, consisting of 44 reactions, is constituted by subclass EC 3.5.c.d, hydrolases acting on carbon−nitrogen bonds other than in peptides. EC 3.6.c.d (hydrolases acting on acid anhydrides) is represented by twelve reactions, whereby all reactions are of subclass EC 3.6.1.d (acting on phosphorus containing anhydrides). The subclass EC 3.7.c.d (hydrolases acting on carbon−carbon bonds) contained three reactions, and there was only one reaction of subclass EC 3.8.c.d (hydrolases acting on carbon−sulfur bonds).

EC 3.1.c.d. All reactions catalyzed by enzymes falling into subclass EC 3.1.c.d (hydrolases acting on ester bonds) were selected from the data set of EC 3.b.c.d for a more detailed analysis. The original data set contained 54 reactions; all reactions have only one reacting bond on the substrate site except for the one reaction catalyzed by the enzyme with the EC number 3.1.3.36. The data set is composed of 14 reactions of sub-subclass EC 3.1.1.d (carboxylic ester hydrolases), four reactions of sub-subclass EC 3.1.2.d (thioester hydrolases), 24 reactions of sub-subclass EC 3.1.3.d (phosphoric monoester hydrolases), ten reactions of sub-subclass EC 3.1.4.d (phosphoric diester hydrolases), and one reaction each of sub-subclasses EC 3.1.5.d (triphosphoric monoester hydrolases) and EC 3.1.6.d (sulfuric ester hydrolases). An overview of the reactions catalyzed by the enzymes of these sub-subclasses is given in Figure 1.

EC 3.2.c.d. The data set of subclass EC 3.2.c.d (glycosylases) contained 19 reactions, each having one reacting bond on the substrate side. The data set consisted of two sub-subclasses: EC 3.2.1.d (glycosidases) with eleven reactions and EC 3.2.2.d (glycosylases hydrolyzing N-glycosylic compounds) with eight reactions, which in all cases contain a purine ring system. The reactions catalyzed by the enzymes of sub-subclasses 3.2.1.d and EC 3.2.2.d are shown in Figure 2.

EC 3.5.c.d. All reactions catalyzed by enzymes of subclass EC 3.5.c.d (hydrolases, acting on carbon−nitrogen bonds, other than peptide bonds) were extracted from the data set of EC 3.b.c.d. The data set contained 44 reactions with one or two reacting bonds. The data set consisted of four sub-

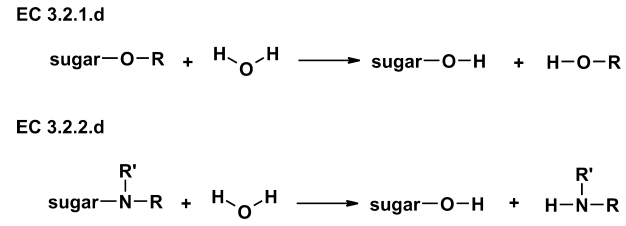

Figure 2. Reaction equations for the reactions catalyzed by the sub-subclasses of EC 3.2.c.d (glycosylases): EC 3.2.1.d (glycosidases hydrolyzing O- and S-glycosylic compounds) and EC 3.2.2.d (hydrolases hydrolyzing N-glycosylic compounds).

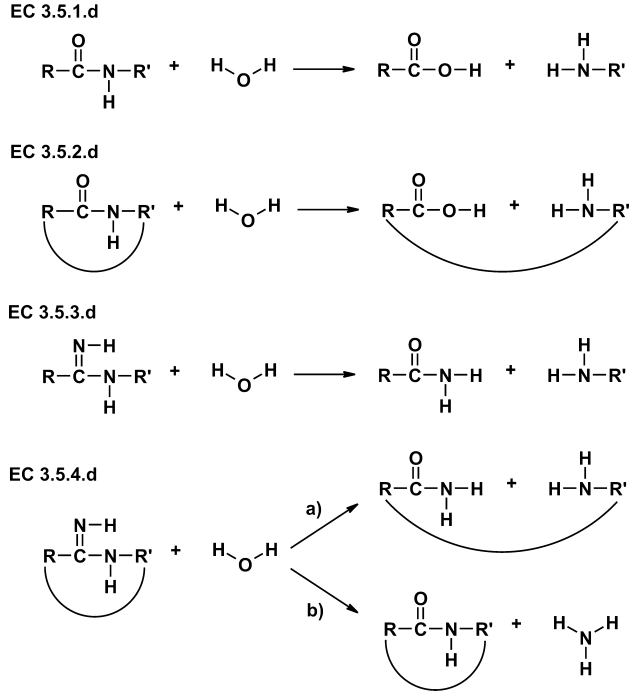

Figure 3. Reaction equations for the reactions catalyzed by the sub-subclasses of EC 3.5.c.d (hydrolases acting on carbon−nitrogen bonds, other than peptide bonds): EC 3.5.1.d (acting on open-chain amides), EC 3.5.2.d (acting on lactams), EC 3.5.3.d (acting on guanidines), EC 3.5.4.d (acting in cyclic amidines).

subclasses: EC 3.5.1.d (hydrolases acting on linear amides) with sixteen reactions, EC 3.5.2.d (hydrolases acting on cyclic amides (lactams)) with nine reactions, EC 3.5.3.d (hydrolases acting on linear amidines) with five reactions, and EC 3.5.4.d (hydrolases acting on cyclic amidines) with 14 reactions. An overview on the reactions considered is given in Figure 3.

2.3. Choice of Descriptors. To represent a reaction, each reacting bond on the substrate side was described by six physicochemical descriptors, suited to describe the electronic character of the bonds taking part in the reaction. The choice of descriptors was based on the work reported in ref 20 and modified for the needs of biochemical reactions. All descriptors were calculated by rapid empirical procedures implemented in the program *PETRA*.[19] The following descriptors were used for the study:

- difference in partial atomic charges (σ and π), Δq_{tot}, describes the polarity of a bond[21,22]
- difference in σ-electronegativities, $\Delta\chi_{\sigma}$, describes the ability of an atom to attract the electrons in a σ-bond[21]
- difference in π-electronegativities, $\Delta\chi_{\pi}$, describes the ability of an atom to attract the electrons in a π-bond[22]

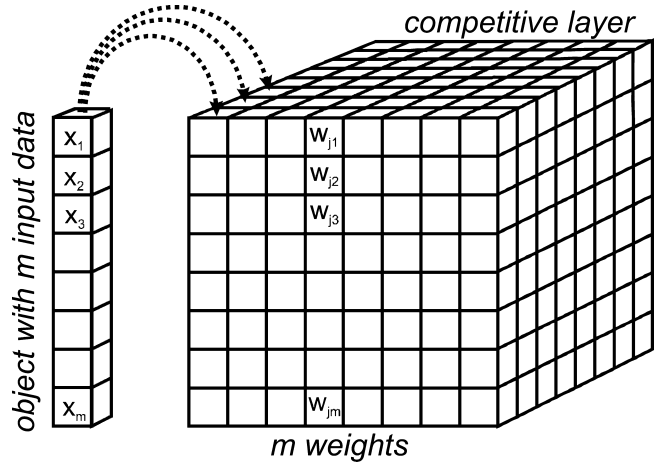

Figure 4. Schematic illustration of a Kohonen neural network. The network is composed of a two-dimensional arrangement of neurons. Each neuron has a number of weights equal to the number of descriptors representing the reacting bonds (input vector). In the training process, the input vectors are presented to the network and the neuron having weights most similar to the descriptors of reacting bonds will receive the considered reaction. A weight adjustment process is then initiated.

- effective bond polarizability, α_b, describes the tendency of the bond-electrons to be perturbated by an external electrical field[23]
- delocalization stabilization of a negative charge, D^-, describes the stabilization of a negative charge generated by heterolysis of the bond[24]
- delocalization stabilization of a positive charge, D^+, describes the stabilization of a positive charge generated by heterolysis of the bond[24]

In making these choices, we attempted to consider all major electronic effects influencing reaction mechanisms, charge distribution, inductive, resonance, and polarizability effects.

2.4. Kohonen Neural Network. By calculating the six descriptors as described above, each bond of a reaction is represented by the six descriptors. Therefore, the breaking of each bond is an event in a six- or twelve-dimensional space, spanned by the above six descriptors as coordinates for each bond. In order to determine the position of each reaction in this six- or twelve-dimensional space, each reaction is projected into a two-dimensional plane using a Kohonen self-organizing neural network. This method has already successfully been used for the classification of organic reactions.[25,26] The method projects data from multidimensional space into a two-dimensional map, preserving the topological relationships of the multidimensional space. The software used for the generation of the Kohonen maps was *SONNIA*.[27,28]

In *SONNIA*, each neuron is initialized with a random weight and has a dimension, m, identical to that of the input vectors (Figure 4). During the training process, each input vector is presented to the network and projected into the neuron with weights most similar to the input vector. The weights of the neurons are then adapted whereas this modification of weights decreases with increasing distance of the winning neuron. This process is iterated several times until convergence is obtained.

For comparison with the classical EC nomenclature, each neuron in the 2D map is colored by the reactions belonging

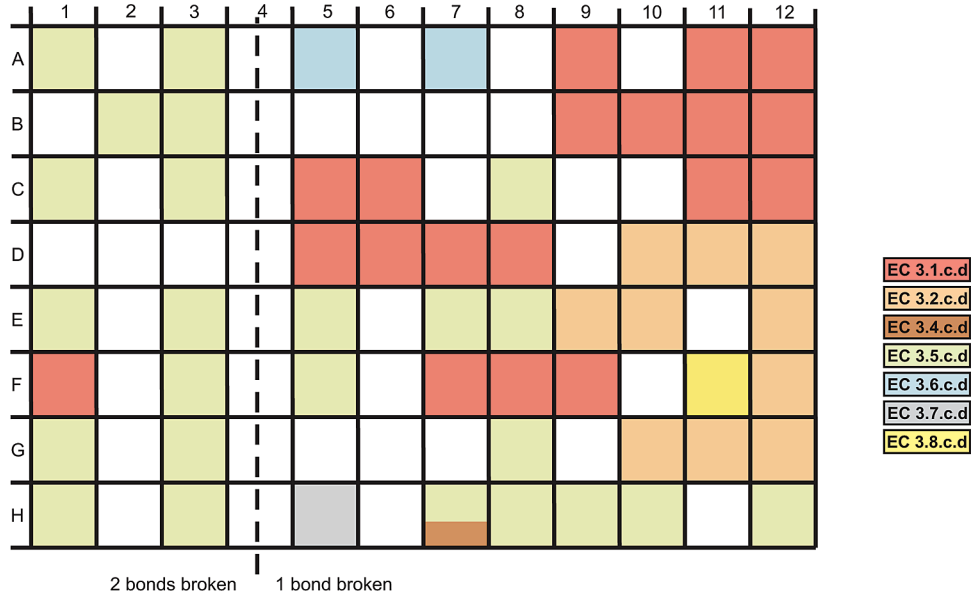

Figure 5. Projection of the reactions of class EC 3.b.c.d (hydrolases; 135 reactions) into a two-dimensional Kohonen neural network. The coloring of the neurons is assigned by the subclass b. Neurons, in which reactions of a different EC 3 subclass b are mapped into are called 'conflicting' neurons and are indicated by multiple colors. In this case, the filled area corresponds to the amount of reactions of a subclass. The dotted line marks the border between the area with reactions where one bond is broken and the area where two bonds are broken during the reaction process.

to a specific EC subclass. Neurons where reactions of different EC subclasses come together are called 'conflicting neurons'. In this case, the neuron is filled with different colors, whereas the colored area corresponds to the amount of reactions belonging to a specific EC subclass. In conflicting neurons, the method described here considers reactions similar that are put into different sub- or sub-subclasses by the EC system. In the following figures the Kohonen network is inspected from the top of Figure 4 in order to show how the various reactions are distributed into the individual neurons.

To objectively interpret such Kohonen maps the final weight values of the neurons of the trained network are used. The relationships and clusters of the enzyme-catalyzed reactions are derived by calculating the Euclidean distances of the weight values of each adjacent neuron pair. Neighboring neurons that show small weight distance are quite similar and may belong to the same cluster, large weight distances separate different clusters. Finally, using a 2D map for similarity perception has several advantages: (i) different directions can represent different kinds of similarities and (ii) distances can indicate the degree of similarity.

3. RESULTS AND DISCUSSION

First, the overall classification of the entire data set of reactions catalyzed by hydrolases (EC 3.b.c.d) is investigated. Then, the individual subclasses for which enough examples were available, that is, the data sets of EC 3.1.c.d, EC 3.2.c.d, and EC 3.5.c.d were subjected to a more detailed analysis. We refrain from a too detailed discussion of the mapping of the reactions by the self-organizing neural network as these can be discerned from the individual figures. Only the major features of the mapping will be discussed.

3.1. EC 3.b.c.d. In this study, all reactions catalyzed by EC class 3, hydrolases, were examined. For this data set containing 135 reactions, the six descriptors described in section 2.3 were calculated for each reacting bond and the resulting data were then mapped into a planar 12 × 8 Kohonen map. The resulting map is shown in Figure 5.

To allow a comparison with the EC system, the neurons in Figure 5 were colored by their affiliation with the different subclasses b (EC 3.b.c.d). A neuron into which reactions of different EC 3 subclasses b are mapped is called a "conflicting" neuron and is indicated by multiple colors.

Inspection of Figure 5 shows a clear separation of the reactions into two groups, indicated by a dashed line. On the left-hand side, there is a widespread cluster of reactions belonging to subclasses EC 3.5.c.d and EC 3.1.c.d. They all have in common that two bonds of the substrate are broken in the reaction. This cluster is separated by empty neurons in column 4 from the reactions on the right-hand side, where only one bond takes part in the reaction. Clearly, the number of descriptors, caused by the different number of reacting bonds, must result in a strong separation of the two groups of reactions. Observe also that changes in bond order, converting a double bond into a single bond are indicated. It should be emphasized that two bonds are not necessarily broken simultaneously in these reactions catalyzed by a single enzyme. Either in the breaking of a bond another bond changes its bond order or, after the first bond is broken, a second consecutive step follows (which might be spontaneous). However, overall, eventually two bonds are broken or change bond order.

For the further interpretation of the Kohonen map the weight distance information of adjacent neurons was calculated and is shown in Figure 6.

As shown in Figure 6, the already discussed strong separation of the two groups of reactions can also be seen in large weight distances in neurons A4−H4.

By and large, the similarity analysis of reactions based on physicochemical descriptors reproduces the classification of enzymes by the EC number. This is true for the situation given here where the identity of the reacting bond is considered in the EC classification.

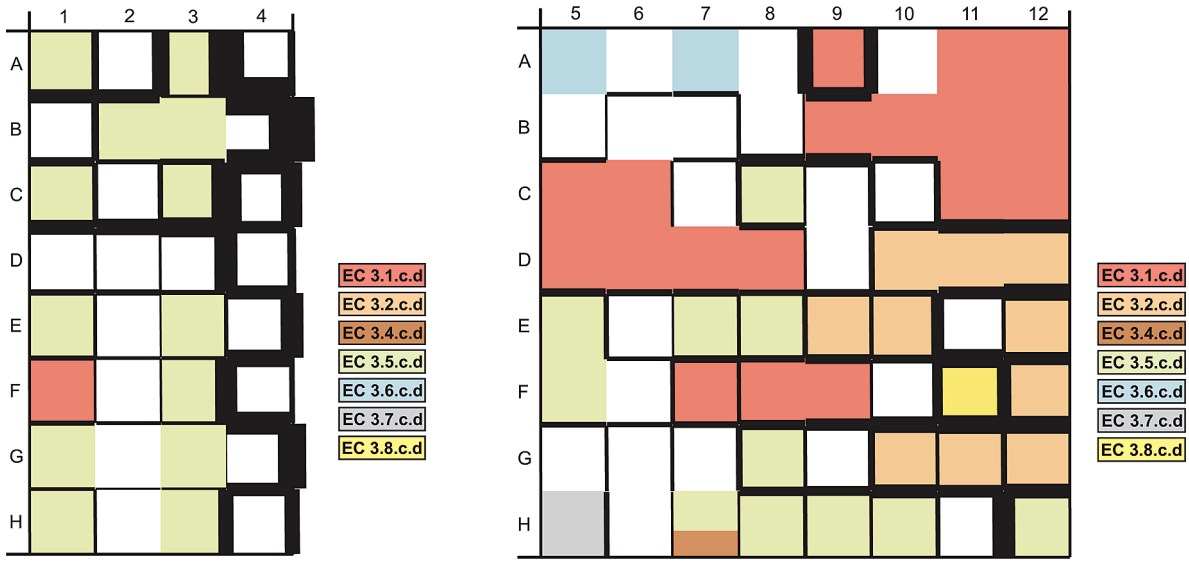

Figure 6. Same Kohonen map as in Figure 5 showing weight distance information between adjacent neurons by the thickness of lines separating them. The map was split into two parts as the scaling of the weight distance values is different: On the left-hand side these values were divided by factor 5 and then visualized as thickness of lines in millimeter; on the right-hand side these values were not scaled at all. On the left-hand side a line is only shown if the unscaled value is larger than 1.0, on the right-hand side the value must be larger than 0.5.

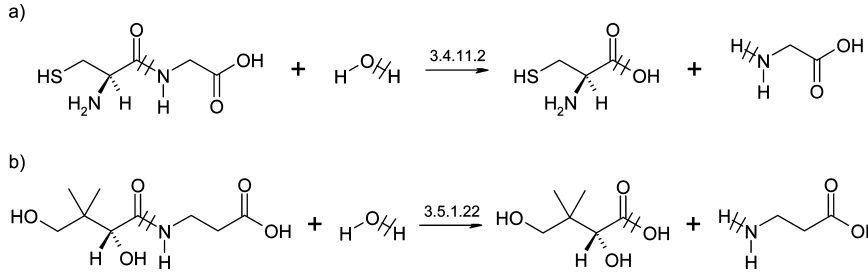

Figure 7. Comparison of two reactions one being catalyzed by a peptidase (EC 3.4.c.d) the other one by an enzyme acting on carbon−nitrogen bonds, other than peptide bonds (EC 3.5.c.d). The physicochemical effects on the reaction center when hydrolyzing the carbon−nitrogen bond are very similar for a) the dipeptide molecule (catalyzed by EC 3.4.11.2) and b) the substrate (R)-pantothenate (catalyzed by 3.5.1.22). Both reactions are projected into the conflict neuron H7 of Figure 5. Double crossed bonds indicate a bond broken in the reaction.

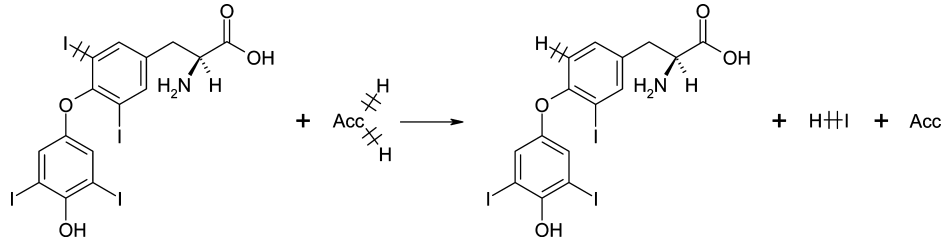

Figure 8. Reaction catalyzed by the enzyme thyroxine deiodinase. Until 2003, this enzyme had the EC number EC 3.8.1.4, but since 2003 it is assigned as oxidoreductase of subclass EC 1.97.c.d. The abbreviation Acc stands for an acceptor molecule.

We will defer a more detailed discussion of this mapping by looking at the mappings of the EC sub-subclasses in the following section. Here we only want to comment two of the apparently more unusual mappings.

The reactions of subclass EC 3.4.c.d, peptidases, are projected into the conflict neuron H7, where they fall together with reactions of subclass EC 3.5.c.d, hydrolases acting on carbon−nitrogen bonds other than peptide bonds. To put these reactions together with subclass EC 3.5.c.d does in our opinion make sense on the basis of mechanistic considerations. To illustrate this similarity two reactions from this neuron are shown in Figure 7.

The reaction of the dipeptide catalyzed by the enzyme with EC 3.4.11.2 (Figure 7a) is very similar to the substrate (R)-pantothenate of the reaction catalyzed by EC 3.5.1.22 (Figure 7b). Accordingly, the physicochemical effects of the carbon−nitrogen bond broken in both reactions are quite similar. The separation of the hydrolysis of peptides from those of amides is only of a phenomenological manner not warranted by the reaction mechanism which is of the same nature, and thus they are not separated by our method.

From subclass EC 3.8.c.d only one reaction is contained in the data set, the one catalyzed by thyroxine deiodinase (EC 3.8.1.4), where a carbon−iodine bond is broken (see Figure 8). The reaction is located in neuron F11 within the area of glycosylases, but separated by quite large weight distances (see Figure 6). Interestingly, the enzyme code of this enzyme that had been assigned as EC 3.8.1.4 in 1984 by the Nomenclature Committee of the IUBMB has been reassigned to EC number 1.97.1.10 in 2003. Thus, the

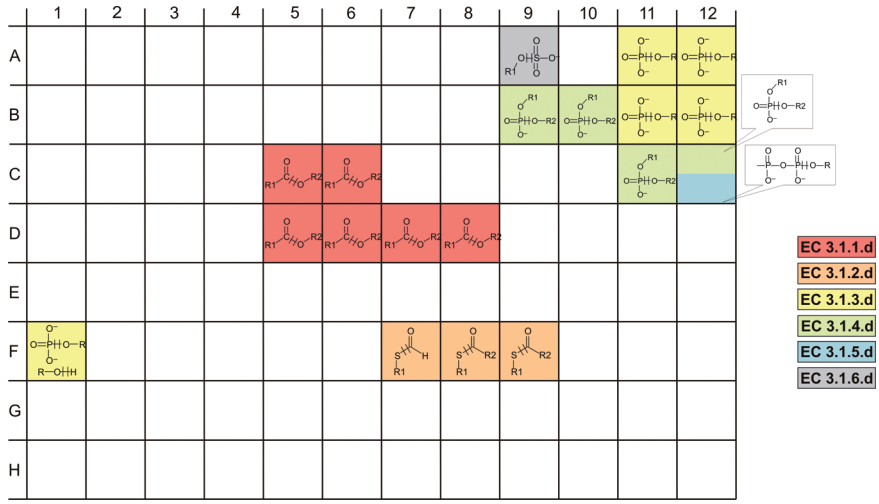

Figure 9. Highlighting of reactions of subclass EC 3.1.c.d (hydrolases acting on ester bonds; 54 reactions) into the trained Kohonen neural network of Figure 5. The coloring of the neurons is determined by the sub-subclass c. Neuron C12, where reactions of different sub-subclasses fall together is marked by two colors. Here, the ratio of the colored area corresponds to the amount of reactions of the two sub-subclasses.

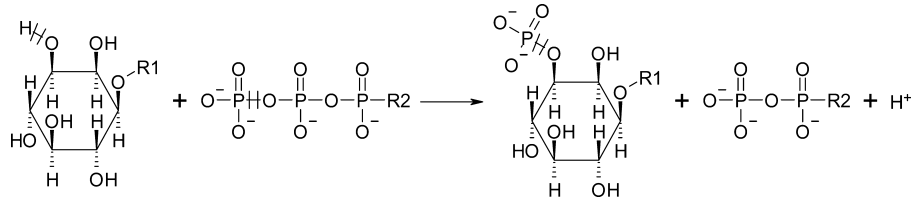

Figure 10. Reaction catalyzed by enzyme phosphatidylinositol 3-kinase having the EC number 2.7.1.137. R1 represents the phosphatidyl residue, R2 stands for adenosine. Because of a coding error in the BioPath database, this reaction was assigned to enzyme phosphoinositide 5-phosphatase with the EC number 3.1.3.36. This error is obvious as the reaction is classified as an outlier by its location in an area where only reactions of EC subclass 3.5.c.d are projected.

enzyme is no longer classified as a hydrolysis but as an oxidoreductase of subclass EC 1.97.c.d "other oxidoreductases". In the meantime, the EC number of this reaction was also updated in the BioPath database.

In the following, we investigate now the individual subclasses EC 3.1.c.d, EC 3.2.c.d, and EC 3.5.c.d in more detail.

3.2. EC 3.1.c.d. For this study, all reactions of EC subclass 3.1.c.d were explored. The trained Kohonen map of Figure 5 was taken, and only those neurons were marked in which at least one of the 54 reactions of subclass 3.1.c.d were projected. The coloring of the neurons is based on the sub-subclass c (EC 3.1.c.d) and the resulting map with the reaction centers indicated is shown in Figure 9.

The data set is mainly split into four separated areas. The first area consists of the reactions from the sub-subclasses EC 3.1.1.d (carboxylic ester hydrolases) (red area), the second from reactions catalyzed by enzymes from EC 3.1.2.d (thioester hydrolases) (orange area). The third area consists of EC 3.1.3.d (phosphoric ester hydrolases and sulfuric ester hydrolases) (neurons marked in yellow), EC 3.1.4.d (phosphoric diester hydrolases) (neurons marked in green), EC 3.1.5.d (triphosphoric monoester hydrolases) (neurons marked in blue), and EC 3.1.6.d (sulfuric ester hydrolases) (gray area). Finally, the fourth area is located on neuron F1 that contains only one reaction.

The classification of enzyme catalyzed reactions on the basis of the physicochemical effects of the reacting bond works quite well, even on the level of EC sub-subclasses: all reactions catalyzed by the enzymes of these six sub-subclasses were separated quite well from each other and overlap only in a single neuron C12. In neuron C12 a reaction that is catalyzed by EC 3.1.4.d (phosphoric diester hydrolases) is projected, together with a reaction catalyzed by EC 3.1.5.d (triphosphoric monoester hydrolases): phospholipase C (EC 3.1.4.1) that cleaves O-phosphocholine from lecithin and dGTPase (EC 3.1.5.1), hydrolyzing a triphosphate group from dGTP (see structures in Figure 9). The classification of the reaction of dGTPase into an area of neurons carrying reactions breaking bonds of phosphodiesters is clearly warranted, as here also a nonterminal phosphoester bond is broken.

On the basis of our physicochemical effects the hydrolysis of a terminal phosphoester bond can clearly be separated from the hydrolysis of a central phosphoester bond. However, our selection of physicochemical effects can hardly distinguish between cleaving a diphosphate or triphosphate group.

The reaction that hydrolyzes a sulfate group from a steryl-substrate is the only representative of a sulfuric ester hydrolysis. The reaction, catalyzed by sterylsulfatase (EC 3.1.6.2), was projected into neuron A9. The reaction of sterylsulfatase is apparently governed by physicochemical effects similar to those operating in phosphodiesters. This has the result that this reaction is projected into a neuron of its own but which is adjacent to reactions in which phosphodiesters are hydrolyzed.

One reaction of EC 3.1.c.d (EC number 3.1.3.36, see Figure 10) has two reacting bonds on the substrate side and was projected into neuron F1 that is far removed from the other clusters of EC 3.1.c.d.

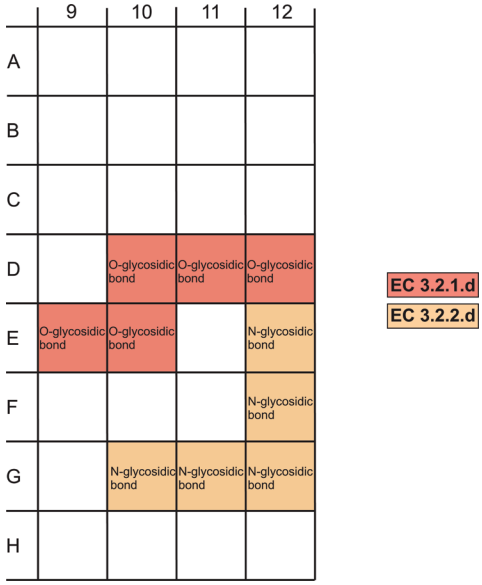

Figure 11. Projection of reactions of subclass EC 3.2.c.d (glycosylases; 19 reactions) into the trained Kohonen neural network of Figure 5. Only neurons containing reactions of this subclass are shown. The coloring of the neurons is determined by the sub-subclass c. The label indicates the type of bond broken by the reactions in the neurons.

This reaction had been assigned in BioPath to be catalyzed by the enzyme with EC number 3.1.3.36 (phosphoinositide 5-phosphatase). However, this reaction is not a hydrolysis as the hydroxyl group of the 1D-*myo*-inositol reacts with ATP by transferring a phosphate group. The reaction is actually catalyzed by phosphatidylinositol 3-kinase (EC 2.7.1.137). Thus, the projection of this reaction into a neuron far removed from the other neurons containing reactions of the subclass EC 3.1.c.d led to the detection of a coding error. In the meantime, this coding error has been corrected in the BioPath database.

In summary, the differences in reactivity of the several reaction types grouped within this EC subclass are quite distinct and this is reflected by their clear separation in the map of Figure 9. We believe that these distinct differences should be considered on a higher level of the classification scheme.

3.3. EC 3.2.c.d. Nineteen reactions of subclass EC 3.2.c.d, glycosylases were studied in more detail. Therefore, the trained Kohonen map of Figure 5 was used after coloring the neurons according to their sub-subclass (Figure 11).

The data set contained reactions of two sub-subclasses, EC 3.2.1.d (glycosylases hydrolyzing O- and S-glycosyl compounds) and EC 3.2.2.d (glycosylases hydrolyzing N-glycosyl compounds). Both subgroups are completely separated by the method based on physicochemical effects without any conflicts. This study shows that the original, unrevised classification of EC 3.b.c.d is suitable to even get a clear separation on the level of sub-subclasses.

3.4. EC 3.5.c.d. For this investigation, the data set, containing 44 reactions of subclass EC 3.5.c.d (hydrolases, acting on carbon−nitrogen bonds, other than peptide bonds), was projected into the trained Kohonen map of Figure 5. The coloring of the neurons is based on the sub-subclass c (EC 3.5.c.d).

The data set contains reactions in which one or two bonds are broken. A clear separation can be detected between reactions where two bonds are broken versus those where one bond is broken, each having 22 reactions.

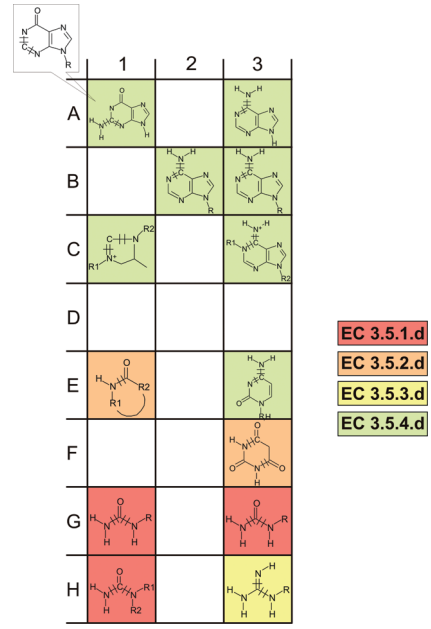

Figure 12. Highlighting the reactions of subclass EC 3.5.c.d (hydrolases acting on carbon−nitrogen bonds, other than peptide bonds). The coloring of the neurons is assigned by the sub-subclass c. In this figure, the contents of the neurons containing reactions breaking two bonds simultaneously are labeled by the reaction centers of the mapped reactions. For each neuron, the reaction center and its neighboring atoms and bonds are shown. Crossed bonds indicate a change in bond order; bonds doubly crossed indicate a bond broken. The color of the neurons is determined by the sub-subclass to which the reactions belong.

In the following analysis, we will inspect the reactions breaking two bonds first and then concentrate on the reactions breaking one bond. The reaction centers for the reactions with two bonds broken are indicated for the various neurons in Figure 12, showing how the reactions are projected into the network.

The different sub-subclasses EC 3.5.1.d to EC 3.5.4.d are well separated (see also Figure 6). Within sub-subclasses EC 3.5.4.d reactions of purines are well distinguished from reactions of pyrimidines. Furthermore, within the reactions of purine derivatives, the reactions at different positions in the ring are well separated.

Now let us turn toward the reactions of subclass EC 3.5.c.d breaking one bond. These reactions are more interwoven than the reactions breaking two bonds, projected into several neurons. The reaction centers for these reactions are indicated in Figure 13.

The twelve reactions of sub-subclass EC 3.5.1.d, hydrolases breaking open-chain amides, are distributed into neurons E5, F5, H7, H8, H10, and G8 (see Figure 13). The distribution of these reactions into the neurons is strongly correlated with the substructure: While in neurons E5 and F5 always a primary amide is broken with release of ammonia, in neurons H7 and H8 always a secondary amide is hydrolyzed. For the reactions in neurons G8 and H10, a formyl residue is hydrolyzed from the substrate distinguishing between a secondary and tertiary amide. This clearly reflects the influence of the environment of the reaction center on the reactivity of a bond.

Six reactions of sub-subclass EC 3.5.2.d, hydrolases breaking bonds in lactams, fall into neurons H9 and E8 (see structures in Figure 13). In the Kohonen map, especially,

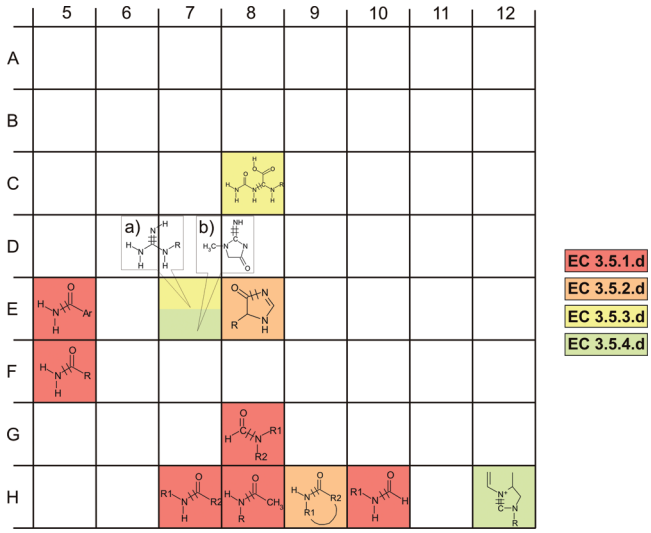

Figure 13. Highlighting the reactions of subclass EC 3.5.c.d (hydrolases acting on carbon−nitrogen bonds, other than peptide bonds). The coloring of the neurons is assigned by the sub-subclass c. In this figure, the contents of the neurons containing reactions breaking one bond in the reaction process are labeled by the reaction centers of the mapped reactions. For each neuron, the reaction center and its neighboring atoms and bonds are shown. Crossed bonds indicate a change in bond order; bonds doubly crossed indicate a bond broken. The color of the neurons is determined by the sub-subclass to which the reactions belong; conflicting neurons are indicated by multiple colors. The letter captions are explained in the text.

the reactions of neuron H9 are surrounded by reactions of sub-subclass EC 3.5.1.d. This indicates, that the distinction between open-chain and cyclic amides and amidines is not retraced by the method presented here, as no strong impact on the reactivity of the amide bond is to be expected. The distinction between hydrolysis of amides and of lactams is more of a phenomenological manner.

The two reactions of sub-subclass EC 3.5.3.d (hydrolyses of carbon−nitrogen bonds in linear amidines) fall into neuron E7, arginine deiminase, and C8, allantoicase. For arginine deiminase (EC 3.5.3.6), the carbon−nitrogen double bond of the guanidinium group is broken. Allantoicase (allantoate amidinohydrolase, EC 3.5.3.4) hydrolyzes the ureide allantoate to ureidoglycolate and urea. In the tautomeric form shown in neuron C8 of Figure 13 a carbon−nitrogen bond is broken. Based on another tautomeric form (CH acidity of central fragment), this bond can also be drawn as amidine. However, as stated by Leulliot et al., the exact mechanism how this carbon−nitrogen bond is hydrolyzed is not yet clarified: "The carbon−nitrogen bond hydrolyzed by allantoicase is not very reactive, and with the present structure in hand, how this may be achieved remains an intriguing question".[29] In our opinion, the classification of this reaction by the EC system is deficient; perhaps allantoicase should be accommodated in a sub-subclass of its own. However, the classification method we present here estimates a higher similarity to the reactions of subclass EC 3.1.c.d surrounding neuron C8, hydrolyzing esters, among others.

The two reactions of sub-subclass EC 3.5.3.d are very different, indeed, and a classification into the same sub-subclass is not warranted in our opinion.

Finally, the two reactions of sub-subclass EC 3.5.4.d fall into neuron H12, methenyl-THF-cyclohydrolase (EC 3.5.4.9) and E7, creatinine deiminase (EC 3.5.4.21). While for the reaction in E7, creatinine deiminase, a carbon−nitrogen bond attached to a five membered ring is broken (see structure a in Figure 13), in the reaction of methenyl−THF−cyclohydrolase, a carbon−nitrogen bond that is part of a five-membered ring and where the nitrogen atom carries a positive charge is broken (see structure in neuron H12 of Figure 13). Clearly, the physicochemical effects on the reactions are different, leading to this clear separation. Also here, the classification into the same sub-subclass is not warranted when considering the reactivity of the reacting bond. The reaction catalyzed by creatinine deiminase (EC 3.5.4.21) is more similar to the reaction catalyzed by arginine deiminase (EC 3.5.3.6) as discussed above (see structure b in Figure 13).

Looking into the reactions not from the standpoint of EC sub-subclasses, but from their affiliation to neurons, the following observations can be made: In neuron E7, all reactions break the terminal carbon−nitrogen bond of a guanidinium group releasing ammonia. In the neighborhood of this neuron the reaction catalyzed by imidazolonepropionase (EC 3.5.2.7) has been projected. Here, a cyclic amide bond is broken within an aminomethylidene acetamide fragment. For the reactions located in neurons E5 and F5, in all cases an amide bond of a secondary amide is broken. In all five reaction instances projected into neuron H9 the amide bond of a lactam is broken. In neurons G8 and H10, two reactions break an amide bond under release of a formyl residue. Finally, neuron H12 contains one reaction catalyzed by methenyl-THF-cyclohydrolase (EC 3.5.4.9), breaking a cyclic amidine under opening of a ring structure and the reaction of allantoicase (EC 3.5.3.4) in neuron E7 where the broken bond lies outside of the amide bond as stated before.

In conclusion, this study gives a nice example how the character of a bond is strongly influenced by the atoms and bonds of its environment. Analyzing the reactions catalyzed by enzymes of sub-subclass 3.5.4.d shows a strong separation of these reactions, even if the same type of bond is broken. Here, a strong separation is obtained by the type of ring system, purine or pyrimidine, or the location of the bond in the purine ring system. Such evident differences in bond reactivity are only hard to recognize by visual inspection of the substructure but become obvious when analyzing the finer details on the level of sub-subclasses. Apart from the separation caused by the different number of broken bonds, the reactions of the sub-subclasses are indeed mapped into the same region of the map. This means that, even through the differences in bonds broken, the similarity of the reaction center is recognized by the network. This highlights the advantage of similarity perception by a self-organizing neural network. Neurons of a certain part of the maps reflect a certain *type* of similarity. On top of that, distances between the neurons in such an area reflect the *degree* of similarity.

4. CONCLUSIONS

These studies have shown that for the enzyme class EC 3.b.c.d, hydrolases, the reaction center bond properties show similarity in these reactions that overall compares well with the EC system. For this type of reaction, this result meets the expectations, as here the EC classification system also is to a large extent based on the nature of the reacting bond. However, the method presented here perceives finer details

showing how the reacting bonds are influenced by the atoms and bonds in their neighborhood. 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 as shown in the detailed discussions of this publication. Several points have been indicated where even for hydrolases the EC classification could be improved. For other classes of the EC system, the classification of reactions by physicochemical properties of the reaction center will provide a better basis for defining the similarity of reactions and of enzymes.

The work presented here has shown that the physicochemical descriptors form an excellent basis for assessing the similarity of enzyme catalyzed reactions. Thus, this method can also be used for assigning EC codes for known or novel enzyme catalyzed reactions. First, the physicochemical descriptors at the reaction centers have to be calculated and then used for projecting this reaction into a trained Kohonen network. The neuron where this reaction is projected into allows one to perceive the EC subclass or sub-subclass.

It should be emphasized that the similarity of reactions is encoded by the physicochemical effects. Data analysis by a Kohonen neural network was performed for several reasons: (i) the unsupervised learning method is free of bias in looking at a descriptor space and does not impose any prior classification scheme, (ii) a two-dimensional map allows one to perceive similarities in an intuitive manner reflecting both different types and different degrees of similarities. Clearly, a variety of other data analysis methods could be used for perceiving such similarities. In future work we will explore hierarchical classification methods.

ACKNOWLEDGMENT

We thank the Bundesministerium fuer Bildung and Forschung (BMBF) for funding our research within the project Bioinformatics for the Functional Analysis of Mammalian Genomes (BFAM), part of the Nationales Genomforschungsnetz Deutschland (NGFN), projects 031U112D, 031U212D, 031U112A, and 031U212A. We also thank Dr. Thomas Kleinoeder and Dr. Lothar Terfloth, both of Molecular Networks GmbH, for their support in the use of the *PETRA* and *SONNIA* software.

Note Added after ASAP Publication. This paper was released ASAP on May 15, 2009, with errors in Figures 9, 12, and 13. The correct version was posted on May 19, 2009.

REFERENCES AND NOTES

(1) Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). 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 catalyse. http://www.chem.qmul.ac.uk/iubmb/enzyme/ (accessed March 16, 2009).

(2) Arita, M. The metabolic world of *Escherichia coli* is not small. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 1543−1547.

(3) Izrailev, S.; Farnum, M. A. Enzyme classification by ligand binding. *Proteins* **2004**, *57*, 711–724.

(4) Dobson, P. D.; Doig, A. J. Predicting enzyme class from protein structure without alignments. *J. Mol. Biol.* **2005**, *345*, 187–199.

(5) 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.

(6) 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.

(7) 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.

(8) Ruepp, A.; Zollner, A.; Maier, D.; Albermann, K.; Hani, J.; Mokrejs, M.; Tetko, I.; Gueldener, U.; Mannhaupt, G.; Muensterkoetter, M.; Mewes, H.-W. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.* **2004**, *32*, 5539–5545.

(9) 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.

(10) 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.

(11) 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.

(12) Nagano, N. EzCatDB: The enzyme catalytic-mechanism database. *Nucleic Acids Res.* **2005**, *33*, D407−D412.

(13) Gariev, I. A.; Varfolomeev, S. D. Hierarchical classification of hydrolases catalytic sites. *Bioinformatics* **2006**, *22*, 2574–2576.

(14) 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.

(15) Reitz, M.; Sacher, O.; Tarkhov, A.; Truembach, D.; Gasteiger, J. Enabling the exploration of biochemical pathways. *Org. Biomol. Chem.* **2004**, *2*, 3226–3237.

(16) *Biochemical Pathways Wall Chart*; Michal, G., Ed.; Boehringer Mannheim (now Roche): Mannheim, Germany, 1993. It can also be accessed on the internet at http://www.expasy.org/tools/pathways/ (accessed March 16, 2009).

(17) Michal, G. *Biochemical Pathways—An Atlas of Biochemistry and Molecular Biology*; Spektrum Akademischer Verlag: Heidelberg, Germamy, 1999.

(18) *BioPath.Explore*, version 1.0; Molecular Networks GmbH: Erlangen, Germany, http://www.molecular-networks.com, July 2006. (accessed March 16, 2009).

(19) 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.

(20) Sacher O. PhD Dissertation. University of Erlangen-Nuernberg, Erlangen, Germany, 2001.

(21) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3228.

(22) Gasteiger, J.; Saller, H. Calculation of the charge distribution in conjugated systems by a quantification of the resonance concept. *Angew. Chem., Int. Ed. Engl.* **1985**, *24*, 687–689.

(23) Gasteiger, J.; Hutchings, M. G. Quantitative models of gas-phase proton transfer reactions involving alcohols, ethers, and their thio analogs. Correlation analyses based on residual electronegativity and effective polarizability. *J. Am. Chem. Soc.* **1984**, *106*, 6489–6495.

(24) Froehlich A. PhD Dissertation, Technical University Muenchen, Muenchen, Germany, 1993.

(25) 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.

(26) (a) Chen, L.; Gasteiger, J. Reactions classified by neural networks: Michael additions, Friedel−Crafts alkylations by alkenes, and related reactions. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 763–765. (b) Chen, L.; Gasteiger, J. Organische Reaktionen mit Hilfe neuronaler Netze klassifiziert: Michael-Additionen, Friedel-Crafts-Alkylierungen durch Alkene und verwandte Reaktionen. *Angew. Chem.* **1996**, *108*, 844–846.

(27) *SONNIA*, version 4.2; Molecular Networks GmbH: Erlangen, Germany, http://www.molecular-networks.com, Dec 2006. (accessed March 16, 2009).

(28) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, Germany, 1999.

(29) Leulliot, N.; Quevillon-Cheruel, S.; Sorel, I.; Graille, M.; Meyer, Ph.; Liger, D.; Blondeau, K.; Janin, J. I.; van Tilbeurgh, H. Crystal structure of yeast allantoicase reveals a repeated jelly roll motif. *J. Biol. Chem.* **2004**, *279*, 23447−23452.

CI800277F