Systems biology

Genome-scale classification of metabolic reactions and assignment of EC numbers with self-organizing maps

Diogo A. R. S. Latino\textsuperscript{1,2}, Qing-You Zhang\textsuperscript{3} and João Aires-de-Sousa\textsuperscript{1,*}

\textsuperscript{1}CQFB, REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, \textsuperscript{2}CCMM, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal and \textsuperscript{3}College of Chemistry and Chemical Engineering, Henan University, Kaifeng, 475001 China

Received on April 3, 2008; revised on July 25, 2008; accepted on July 29, 2008
Advance Access publication August 1, 2008
Associate Editor: Martin Bishop

ABSTRACT

Motivation: The automatic perception of chemical similarities between metabolic reactions is required for a variety of applications ranging from the computer-aided validation of classification systems, to genome-scale reconstruction (or comparison) of metabolic pathways, to the classification of enzymatic mechanisms. Comparison of metabolic reactions has been mostly based on Enzyme Commission (EC) numbers, which are extremely useful and widespread, but not always straightforward to apply, and often problematic when an enzyme catalyzes several reactions, when the same reaction is catalyzed by different enzymes, when official full EC numbers are unavailable or when reactions are not catalyzed by enzymes. Different methods should be available to compare metabolic reactions. Simultaneously, methods are required for the automatic assignment of EC numbers to reactions still not officially classified.

Results: We have proposed the MOLMAP reaction descriptors to numerically encode the structural transformations resulting from a chemical reaction. Here, such descriptors are applied to the mapping of a genome-scale database of almost 4000 metabolic reactions by Kohonen self-organizing maps (SOMs), and its screening for inconsistencies in EC numbers. This approach allowed for the SOMs to assign EC numbers at the class, subclass and sub-subclass levels for reactions of independent test sets with accuracies up to 92, 80 and 70\%, respectively. Different levels of similarity between metabolic reactions still not officially classified.

Availability: The programs to generate MOLMAP descriptors from atomic properties included in SDF files are available upon request for evaluation.

Contact: jas@fct.unl.pt

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The understanding of the multiple ways small molecules affect biological systems is demanding an integration of chemical and biological data in ‘systems biology’ approaches (Oprea et al., 2007). In this context, ‘enzymatic function’ emerges as a key gateway between the universes of biology and chemistry. Employed for the annotation of genes and proteins, or encoded as an edge in graph representations of metabolic networks, ‘enzymatic function’ has an intrinsic chemical nature—it is the catalysis of a chemical reaction. The automatic comparison and classification of enzymatic reactions, a current issue in bioinformatics, requires specific chemical knowledge and chemoinformatics methodologies. Bioinformatics applications of reaction classification and comparison include (a) computer-aided validation of classification systems, e.g. the assignment of EC (Enzyme Commission) numbers in the context of an ever increasing number of enzymatic reactions; (b) genome-scale reconstruction of metabolic pathways, where similarity searches of metabolic reactions are helpful for the proposal of enzyme sequences from their functions, and for the annotation of unknown genes (Kotera et al., 2004; Yamanishi et al., 2007); (c) classification of enzymatic mechanisms (Boyle et al., 2007); (d) visualization of reactomes; (e) enzymatic structure–function studies (Babbi, 2003; Shaknovich and Harvey, 2004; Todd et al., 2001); and (f) alignment of metabolic pathways (Pinter et al., 2005). This article reports the classification of a genome-scale data set of metabolic reactions by self-organizing maps (SOMs) using physicochemical and topological features of reactants/products to represent reactions. The data set encompasses all possible reactions in the KEGG database, which includes exhaustive lists of known metabolic reactions from different organisms.

Protein catalytic functions are officially classified by the EC numbers assigned to the catalyzed chemical reactions (Barrett et al., 1992; Tipton and Boyce, 2000). The EC number is often simultaneously employed as an identifier of reactions, enzymes and enzyme genes, linking metabolic and genomic information. The assignment of EC numbers to new enzymes is performed based on published experimental data that include the full characterization of the enzymes and their catalytic functions. Although chemically meaningful, and widespread, EC numbers have limitations (Babbi, 2003). They are often problematic in practice, for example with regard to reaction reversibility (Todd et al., 2001), or when an enzyme catalyzes more than one reaction, or when the same reaction is catalyzed by different enzymes (Green and Karp, 2005; Kotera et al., 2004). Different methods should be
available to automatically compare metabolic reactions from their reaction formulas, independently of EC numbers. Such methods are mandatory for the comparison of reactions with incomplete official EC numbers, with EC numbers still not assigned or with no EC number (if they are not catalyzed by enzymes) (Kotera et al., 2004). Advantageously, classification of reactions should take into account physicochemical features of reactants and products, as these affect reactivity and reaction mechanisms.

Chen (2003) reviewed the methods for the representation and classification of reactions that have been put forward in the last 30 years. One strategy has been to identify and encode the structural changes operated by the reaction, for example using conversion patterns of atom types, or physicochemical features of the atoms and bonds at the reaction center (Rose and Gasteiger, 1994; Satoh et al., 1998). A numerical fixed-length code representing physicochemical properties of the reaction center enabled (Chen and Gasteiger, 1996, 1997) to explore Kohonen SOMs for reaction classification. Recently, the same method was applied to classify metabolic reactions of subclass EC 3.1.x.x on the basis of physicochemical properties of the reactants (Gasteiger, 2007). A completely different approach was implemented in Daylight software, which can represent reactions by ‘difference Daylight fingerprints’—reactant fingerprints subtracted from product fingerprints (Daylight, 2008).

We have proposed the MOLMAP method for numerically encoding the structural transformations resulting from a chemical reaction (Zhang and Aires-De-Sousa, 2005). The chemical bonds existing in the structure of each reactant, and each product, are classified by a SOM on the basis of their (calculated) physicochemical and topological properties. This leads to a numerical fixed-length representation (the MOLMAP) describing the types of bonds available in each molecule. By subtracting the MOLMAPs of the products from the MOLMAPs of the reactants, a MOLMAP of the reaction is obtained, which represents the types of bonds that disappeared from the reactants and those created in the products. This method provides a fixed-length numerical representation of chemical reactions, which does not require the assignment of reaction centers, i.e. the explicit identification of the bonds that change in the reactants or are formed in the products, and avoids atom-to-atom mapping, i.e. the correspondence between atoms in the reactants and products.

Later, our lab presented preliminary results concerning the application of MOLMAPs to the classification of a genome-scale data set of enzymatic reactions (Latino and Aires-De-Sousa, 2006). [More recently, Faulon et al. (2008) reported a very similar approach but using molecular signatures of topological atom neighborhoods, and support vector machines as the learning algorithm to classify metabolic reactions.] Here, we report the training of SOM with the enzymatic reactions in the KEGG database in order to their classification, to study the agreement between EC numbers and the MOLMAP-based classification, to assign the first three digits of the EC numbers from the reaction formula and to investigate reactions revealed as similar but belonging to different EC classes.

2 METHODS

2.1 Data set of chemical reactions

Enzymatic reactions were extracted from the KEGG LIGAND database (release of November 2006) (Goto et al., 1998; Kanehisa and Goto, 2000) in MDL .mol format. Details on the pre-processing of the reaction data are available in the Supplementary Material. The final data set (‘whole data set’) includes each reaction in both directions and consists of 7482 reactions—2594 of class EC 1 (oxidoreductases), 2438 of class EC 2 (transferases), 1278 of class EC 3 (hydrolases), 666 of class EC 4 (lyases), 206 of class EC 5 (isomerases) and 300 of class EC 6 (ligases). Different criteria were used for building training and test sets, in some cases aiming at covering the maximum possible diversity of reactions, in other cases aiming at tuning the level of similarity between training and test sets on the basis of EC numbers (e.g. by not allowing the same full EC number, or the same sub-subclass, to appear both in the training set and in the test set).

2.2 Kohonen SOM

SOMs with toroidal topology were used in this study for two independent tasks, the classification of chemical bonds for the generation of a molecular descriptor, and the classification of enzymatic reactions. A short description of the learning algorithm of SOMs can be found in subsection ‘SOMs learning algorithm’ in Supplementary Material. SOMs were implemented throughout this study with an in-house developed Java application derived from the JATOO Java applets (Aires-De-Sousa, 2002).

2.3 Generation of MOLMAP reaction descriptors

The generation of MOLMAP molecular descriptors is based on a SOM that distributes chemical bonds through the grid of neurons. The chemical bonds are represented by topological and physicochemical features (Supplementary Table S1). A similar idea was proposed by Atalay and Cetin-Atalay (2005) to encode the primary structure of proteins with a SOM on the basis of aminoacid features. More details of the MOLMAP reaction methodology can be found in the subsection ‘Details of the MOLMAP methodology’ in Supplementary Material and the full description in Zhang and Aires-De-Sousa (2005). SOMs of sizes $15 \times 15$, $20 \times 20$, $25 \times 25$ and $29 \times 29$, yielding MOLMAPs of dimension 225, 400, 625 and 841, respectively, were trained with a data set of 1568 bonds extracted from compounds selected by the Ward method (1963). This is a hierarchical clustering method designed to optimize the minimum variance within clusters, and was used here with chemical hashed fingerprints as variables. The algorithm begins with one large cluster encompassing all objects to be clustered. In this case, the error sum of squares is 0. The program searches objects that can be grouped together while minimizing the increase in error sum of squares.

To focus on substructures around functional groups, only bonds were considered that include (or are at a one bond distance from) a heteroatom or an atom belonging to a pi system. Experiments were performed using only topological descriptors of bonds, only physicochemical descriptors, the whole set of descriptors, or the subset of descriptors 1–43, 45, 46, 48, 49, 55, 56, 58, 59, 65, 66, 67, 68 from Supplementary Table S1.

2.4 Classification of enzymatic reactions

MOLMAP reaction descriptors were calculated for all the reactions in the data set. Then, classification of the reactions was performed by new SOMs, unrelated and independent from those first trained with chemical bonds to generate molecular descriptors. The new SOMs were trained with enzymatic reactions to predict EC numbers. The data set of reactions was partitioned into a training and a test set, using different criteria depending on the experiment. After the training, the whole training set was mapped on the surface, and each neuron was classified according to the majority of reactions that activated the neuron (or its neighbors, if the neuron was empty). When a majority could not be obtained, the neuron was classified as undecided. The test set was then submitted to the SOM, and each reaction was classified based on the classification of the neuron it activated.

In the experiments for predicting EC subclass, or sub-subclass, reactions belonging to each EC class were treated separately, i.e. six experiments were performed, one for each EC class. Due to the large number of sub-subclasses
of classes EC1 and EC2, counterpropagation neural networks (CPGNNs) were used in those cases, instead of SOMs. Some details about CPGNNs learning can be found under 'CPGNNs learning algorithm details' in Supplementary Material.

To overcome fluctuations induced by the training random factors, five or ten independent SOMs were trained, generating an ensemble. Ensemble predictions were obtained by majority vote of the individual maps. The relationship between the number of votes for the winning class and the reliability of the prediction was investigated. Another measure of reliability was also investigated, the Euclidean distance between a reaction MOLMAP (a vector with the reaction descriptors) and the weights of the corresponding winning neuron.

### 2.5 Classification of racemase and epimerase reactions

An experiment was performed with 18 reactions of subclass 5.1 extracted from the BioPath database version Nov. 2005 (Molecular Networks GmbH, Erlangen, Germany) with stereochemistry explicitly assigned in molecules. Chirality codes (Aires-De-Sousa and Gasteiger, 2001) were used as molecular descriptors instead of MOLMAPs. Chirality codes are descriptors of molecular chirality that can distinguish stereoisomers, namely enantiomers. Here, conformation-independent chirality codes (CICC) with a dimension of 50 were calculated with the following parameters: (a) all the atoms in each molecule were considered, including hydrogen atoms, (b) partial atomic charge (calculated by PETRA 3.2) was used as the atomic property, (c) the chirality code function was sampled in the interval \([-r, +r]\) with \(r = 0.070 \, \text{Å}^{-1}\), (d) only combinations of four atoms with maximum interatomic path distances of six bonds were considered, (e) the smoothing parameter was set to \((\text{code length/range of } u)^2\). The 50 dimensional vectors were normalized by their vector sum.

### 3 RESULTS AND DISCUSSION

#### 3.1 Mapping the whole data set of enzymatic reactions on a SOM

The MOLMAP representation yielded numerical descriptors of enzymatic reactions that could be further processed by a SOM. SOMs were here applied to evaluate to which extent similarities between reaction MOLMAPs correspond to similarities in EC classification, and whether MOLMAPs could be used for the classification of enzymatic reactions. Note that MOLMAPs represent overall reactions, and do not explicitly consider the mechanisms (although the physicochemical and topological bond properties used for the calculation of MOLMAPs are in principle related to mechanisms). EC numbers are also officially assigned based on overall reactions. A 49×49 Kohonen SOM was trained with the whole data set of 7482 enzymatic reactions encoded by MOLMAPs of size 625 using topological and physicochemical descriptors. During this stage, the SOM made no use of the information related to EC numbers. After the training, each neuron of the surface was assigned to a class (first digit of the EC number), according to the majority class of the reactions activating that neuron. The resulting map is displayed in Figure 1. It clearly shows a trend for reactions to cluster according to the EC class, particularly those catalyzed by oxidoreductases (EC 1), transferases (EC 2), hydrolases (EC 3) and ligases (EC 6). Consistent mapping was observed for 91.4% of the data set—the percentage of correctly classified reactions in terms of EC class when the whole data set is submitted again to the SOM trained with the whole data set. (In five experiments with randomized class labels of the reactions, consistency was only 41–42%.) The more scattered lyase class has a not so typical pattern of overall changes in chemical bonds, which suggests similarities with reactions of other classes. Isomerase reactions (EC 5) are particularly difficult to classify by MOLMAP descriptors, as they often consist of subtle changes in the molecular structures of reactants and products. These problems will be discussed later with a more detailed analysis of the map and the assessment of internal consistency of the EC system.

The influence of MOLMAP size, and type of MOLMAP bond descriptors on the consistency of EC class mapping was verified with a smaller version of the data set (encompassing reactions represented only in the direction of the KEGG reaction file—3784 reactions) on a 29×29 SOM (Supplementary Table S2). Variation of MOLMAP size between 225 (15×15) and 841 (29×29) affected the consistency of mapping in only 0.3–4.1%, depending on the type of bond descriptors used for the MOLMAPs. Topological and physicochemical bond descriptors combined yielded the most robust MOLMAPs in such experiments. MOLMAPs were generated with four sets of bond descriptors (see Section 2): topological, physicochemical, topological + physicochemical and topological + subset of physicochemical descriptors. For the same size of MOLMAPs, the type of descriptors never affected the consistency of mapping in >3%. Among the 16 tried combinations of MOLMAP size and type of bond descriptors, the highest and lowest consistency of mapping differed in only 4.1%. Experiments with bond features derived from PETRA software (Molecular Networks GmbH) did not yield superior results.

A map such as the one shown in Figure 1 has a number of possible applications. The trend for reactions to cluster according to EC numbers allows to explore the SOM for the automatic assignment of EC numbers from the reaction equation (see below). Simultaneously, the inspection of reactions belonging to different EC classes but activating the same neuron may indicate possible inconsistencies in EC numbers, problematic classification of enzymes, similarity between reactions hidden by specific EC rules, mistakes in database entries, as well as limitations of the MOLMAP approach. A systematic analysis of such cases uncovered reaction similarities hidden by differences in EC numbers at the class level. Some examples are displayed in Table 1. While some cases may deserve a revision of EC numbers, others illustrate problematic aspects of the application of EC rules related to reversibility of reactions, or enzymes catalyzing more than one type of reactions.

The first entry of Table 1 corresponds to reactions R00603 and R03523 of the KEGG database, the first is listed as a lyase (dichloromethane dehalogenase, EC 4.5.1.3) and the second is listed as a hydrolase (alkylhalidase, EC 3.8.1.1). Surprisingly, the only difference between them is that one eliminates two chloride ions and the other eliminates one chloride and one bromide ion. In entry 2, the two reactions are the same, only the substrates are different. They get different classes probably because the enzyme responsible for one of them (R05086, EC 4.2.3.1) also catalyzes a different type of reaction. The two overall reactions of entry 3 are essentially the same, although they are catalyzed by different enzymes, and listed as belonging to different EC classes. The first is catalyzed by the broad group of glutathione S-transferases (EC 2.5.1.18), and the second is catalyzed by leukotriene-C4 synthase (EC 4.4.1.20). Significantly, the latter was named as 2.5.1.37 until 2004. Globally, the first reaction of entry 4 is an intramolecular version of the...
second reaction, but the first is catalyzed by an enzyme officially classified as a lyase (ornithine cyclodeaminase, EC 4.3.1.12), while the second is catalyzed by a transferase (methylamine-glutamate N-methyltransferase, EC 2.1.1.21). The two reactions of entry 5 are currently listed in the KEGG database as irreversible, and are displayed in the table in the same directions as in the metabolic pathways to which they belong. However, if the first reaction is considered in the opposite direction, then both reactions are hydrolysis of amides. The second reaction has in fact an EC number corresponding to a hydrolase (EC 3.5.1.18), but the first is associated with a transferase (EC 2.3.2.2). The two reactions of entry 7 are hydrolysis of epoxides. While the first is catalyzed by hepxoxilin-oxepoxide hydrolase (EC 3.3.2.7), the second is catalyzed by a P-450 unspecific monoxygenase and the EC number 1.14.14.1 is more an identifier of the enzyme than of the catalyzed reaction. The main difference between the two overall reactions of entry 8 is that one involves a ketone and an aldehyde, while the other involves two aldehydes. Apart from that, they are formally the same. However, the first is associated with a transferase (2-hydroxy-3-oxoadipate synthase, EC 2.2.1.5) and the second with a lyase (tartronate-semialdehyde synthase, EC 4.1.1.47). Although the two reactions of entry 9 are officially classified into different EC classes, both are formally hydrolysis of amides (the second occurring intramolecularly, and represented in the opposite direction). Such a similarity is completely hidden by their EC numbers. The reactions of entry 10 are represented in the directions corresponding to their metabolic pathways (both are currently listed in KEGG as irreversible). The first reaction is a hydrolysis (and its EC number corresponds to a hydrolase, EC 3.3.1.1), and the second reaction involves the cleavage of a C–O bond with an EC number corresponding to a carbon–oxygen lyase (EC 4.2.1.22). However, if we consider both reactions in both directions, they are the same overall reaction.

Other neurons activated by reactions of conflicting EC classifications highlighted limitations of the MOLMAP/SOM approach that perceived truly non-similar reactions as similar.
Table 1. Examples of reactions activating the same neuron, but labeled with different EC classes

<table>
<thead>
<tr>
<th>Entry</th>
<th>Reactions diagrams</th>
<th>EC numbers</th>
<th>KEGG identifier</th>
<th>Neuron</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>![Diagram 1]</td>
<td>4.5.1.3 and 3.8.1.1</td>
<td>R00603 and R03523</td>
<td>W:37</td>
</tr>
<tr>
<td>2</td>
<td>![Diagram 2]</td>
<td>3.1.3.3 and 4.2.3.1</td>
<td>R00582 and R05086</td>
<td>A:3</td>
</tr>
<tr>
<td>3</td>
<td>![Diagram 3]</td>
<td>2.5.1.18 and 4.4.1.20</td>
<td>R07069 and R03059</td>
<td>AR:32</td>
</tr>
<tr>
<td>4</td>
<td>![Diagram 4]</td>
<td>4.3.1.12 and 2.1.1.21</td>
<td>R00671 and R01586</td>
<td>A V:33</td>
</tr>
<tr>
<td>5</td>
<td>![Diagram 5]</td>
<td>2.3.2.2 and 3.5.1.18</td>
<td>R03970 and R02734</td>
<td>AK:16</td>
</tr>
<tr>
<td>6</td>
<td>![Diagram 6]</td>
<td>4.1.1.71 and 1.1.1.87</td>
<td>R00272 and R01936</td>
<td>AF:10</td>
</tr>
</tbody>
</table>

(Continued)
(Supplementary Table S3). In many of those cases, the two reactions result in the formation of similar bonds, although the bonds broken in the reactants are different. Globally, the reaction MOLMAPs bear some similarities and may activate the same neuron even if the reactants and the reactions are not similar. There are also cases in which reactants very similar to the products result in MOLMAPs with only few non-null components—two such reactions may yield globally similar (almost null) MOLMAPs, although the non-null values are different because they correspond to different types of bonds being broken, changed or formed (different types of reactions). In other situations, the overall reaction involves more than one transformation making the MOLMAP somewhat similar to that of another reaction where only one of the transformations occurs. Still in other cases, two reactions activate the same neuron but one or both are at a high Euclidean distance to the neuron (probably due to a lack of similar reactions in the database) — the fact that the winning neuron is the most similar neuron to the MOLMAP does not imply they are very similar. Wrongly perceived similarities between reactions may also derive from wrongly perceived similarities and differences between bonds in the SOM employed for the mapping of bonds.

### 3.2 SOM-based assignment of EC first digit from the reaction equation

The MOLMAP/SOM approach was explored to automatically assign EC numbers of reactions from the structures of reactants and products. MOLMAPs of size 625 generated with topological and physicochemical bond descriptors were employed. In a first experiment (Partition 1), training and test sets were selected with a $49 \times 49$ Kohonen SOM. The SOM was trained with all reactions, then one reaction was randomly taken from each occupied neuron and moved to the test set, resulting in a training set with 5855 and a test set with 1646 reactions. Correct predictions of the EC class were achieved for up to 79.6% of the test set (Table 2). In order to overcome random fluctuations, and to improve predictive ability,
Table 2. SOM assignment of the first digit of EC numbers

<table>
<thead>
<tr>
<th>Datasets</th>
<th>No. of reactions</th>
<th>Correct predictions (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>One SOM</td>
</tr>
<tr>
<td>Partition 1</td>
<td>Training 1</td>
<td>5855</td>
</tr>
<tr>
<td></td>
<td>Test 1</td>
<td>1646</td>
</tr>
<tr>
<td></td>
<td>Training 2</td>
<td>5855</td>
</tr>
<tr>
<td></td>
<td>Test 2</td>
<td>1646</td>
</tr>
<tr>
<td>Partition 2</td>
<td>Training</td>
<td>5246</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>2236</td>
</tr>
<tr>
<td>Partition 3</td>
<td>Training</td>
<td>350</td>
</tr>
<tr>
<td></td>
<td>Test 1</td>
<td>4896</td>
</tr>
<tr>
<td></td>
<td>Test 2</td>
<td>7132</td>
</tr>
<tr>
<td>Partition 4</td>
<td>Training</td>
<td>310</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>40</td>
</tr>
</tbody>
</table>

Table 3. Confusion matrix for the classification of enzymatic reactions according to the first digit of the EC number (test set of partition 1; ensemble of ten 49×49 Kohonen SOMs)

<table>
<thead>
<tr>
<th>EC1</th>
<th>EC2</th>
<th>EC3</th>
<th>EC4</th>
<th>EC5</th>
<th>EC6</th>
<th>X</th>
<th>Correct predictions (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC1</td>
<td>573</td>
<td>31</td>
<td>11</td>
<td>21</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>EC2</td>
<td>13</td>
<td>414</td>
<td>13</td>
<td>9</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>EC3</td>
<td>1</td>
<td>10</td>
<td>214</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>EC4</td>
<td>11</td>
<td>20</td>
<td>19</td>
<td>115</td>
<td>2</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>EC5</td>
<td>5</td>
<td>13</td>
<td>2</td>
<td>11</td>
<td>19</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>EC6</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>52</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

Consensus predictions were obtained with ensembles of 5 or 10 independent SOMs trained with the same data set. An increased accuracy of predictions was in fact observed for the training set, and then also for the test set (84.3%), particularly for the lyase class. The number of wrong classifications obtained with an individual SOM and an ensemble of ten SOMs is similar—the improvement in the number of correct predictions using an ensemble mainly derives from reactions that were classified as undecided with a single SOM, and became correctly classified by the ensemble. The confusion matrix for the test set (Table 3) shows a higher prediction accuracy for classes EC2 and EC3, and the worst predictions are for classes EC4 and EC5. Reactions catalyzed by isomerases (EC5) are generally more problematic as very often they involve no substantial structural changes, and they are also less represented in the data set. The results (confirmed by the map of Fig. 1) reveal that MOLMAP patterns for reactions catalyzed by lyases (class EC4) are not so well defined, and most frequently confused with those corresponding to hydrolases and transferases.

Partition 2 also involved a strategy for the training set to cover as much the reaction space as possible—one reaction of each available EC number (full EC number) was selected into the training set (5246 reactions), and the remaining reactions were moved to the test set (2236 reactions). This partition guaranteed that all the full EC numbers in the test set were represented in the training set, which generally means a high similarity between reactions in both sets. In fact, assignment of the EC class was obtained with an accuracy of 91.7% for the test set. For this and the next partitions, the two entries of each reaction (in the two opposite directions) were included in the same set (training or test).

Then experiments were performed with lower similarities between training and test sets. With Partition 3, the model was trained only with one reaction from each sub-subclass (first three digits of the EC number, 350 reactions), and tested with one reaction from each of the remaining full EC numbers (4986 reactions). In this way, all reactions in the training set belonged to a different sub-subclass, and all sub-subclasses were represented. The test set includes only reactions with full EC numbers non-available in the training set. Despite the test set being 14 times larger than the training set, and the exclusion of similarities at the level of the full EC number, 74.4% of the test set could still be correctly predicted in terms of EC class by a 25×25 SOM. An even more stringent test was performed with Partition 4, where all the reactions of the test set belonged to different sub-sub-classes of those in the training set. The map for the training set is available as Supplementary Figure S2. From the set of 350 reactions with no duplicated sub-subclass, a test set of 40 reactions was randomly selected that could be predicted with 67.5% accuracy by a 20×20 SOM. In five experiments with randomization of the reaction labels in the training set, only 20–30% of correct predictions were observed for the test set.

Another independent test was performed using 930 reactions in the KEGG database with incomplete EC numbers. These are often problematic cases, and are therefore expected to present an increased level of difficulty. The ensemble of 10 SOMs trained with all the 7482 reactions with full EC numbers was able to correctly predict the EC class for 73.8% of such cases.

We explored the possibility of obtaining some measure of reliability associated to each prediction. The Euclidean distance to the winning neuron, and the number of votes in the consensus prediction by the ensemble of SOMs were evaluated for the test set of Partition 1 (details of their performance are in Supplementary Tables S4 and S5). With the ensemble of 10 SOMs, 97% of the predictions obtained with 10 votes (for the test set) are correct, and the percentage decreases to 82.3, 75.7, 68.8, 64.6, and 47.1% for predictions obtained with 9, 8, 7, 6 and 5 votes, respectively. The Euclidean distance between the winning neuron and the MOLMAP of the query reaction also performed well. The percentage of correct predictions gradually decreased from 100% for reactions at an Euclidean distance ≤ 101, to 59.3% for reactions at a distance ≥ 5001.

Although SOMs are trained in an unsupervised manner, it is possible to render the training supervised by adding new descriptors encoding the classes of the objects. In our case, six new descriptors were appended to the MOLMAP descriptors, each one corresponding to an EC class. For one reaction, five of the descriptors were zero, and the descriptor corresponding to the EC class took a value of 80. This increases the overall similarity between reaction descriptors of the same class, forcing the SOM to cluster according to class (Supplementary Figure S3). After the training, for the SOM to map a new reaction (possibly of unknown class), the six layers of the network corresponding to the class codes are not used to determine the winning neuron. For Partition 1, such a technique allowed a single SOM to improve the accuracy of predictions for the test set up to 85.4% (similar to the results obtained with the ensemble of ten unsupervised SOMs), but ensembles of supervised SOMs could only marginally improve this number.
Table 4. SOM assignment of the second and third digit of EC numbers

<table>
<thead>
<tr>
<th>Subclass</th>
<th>Partition 1 Training/Test</th>
<th>Partition 2 Training/Test</th>
<th>All reactions</th>
<th>All reactions</th>
<th>Partition 1 Training/Test</th>
<th>Partition 2 Training/Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC1</td>
<td>90.7 (2584)</td>
<td>91.6/41.7 (2548/36)</td>
<td>93.5/74.3 (1805/779)</td>
<td>75.1 (2474)</td>
<td>76.7/30.0 (2394/80)</td>
<td>77.7/65.4 (1864/610)</td>
</tr>
<tr>
<td>EC2</td>
<td>95.5 (2436)</td>
<td>94.9/87.5 (2420/16)</td>
<td>96.1/86.1 (1759/677)</td>
<td>84.8 (2412)</td>
<td>84.6/72.7 (2368/44)</td>
<td>85.0/73.9 (1960/452)</td>
</tr>
<tr>
<td>EC3</td>
<td>96.7 (1262)</td>
<td>96.0/62.5 (1246/16)</td>
<td>96.4/87.4 (912/350)</td>
<td>92.7 (1238)</td>
<td>91.3/73.7 (1200/38)</td>
<td>93.7/57.7 (859/379)</td>
</tr>
<tr>
<td>EC4</td>
<td>89.3 (666)</td>
<td>94.8/57.1 (652/14)</td>
<td>90.1/69.5 (453/213)</td>
<td>86.9 (658)</td>
<td>87.2/26.9 (632/26)</td>
<td>90.4/65.3 (439/219)</td>
</tr>
<tr>
<td>EC5</td>
<td>93.9 (198)</td>
<td>91.2/100 (192/6)</td>
<td>94.6/70.6 (130/68)</td>
<td>88.3 (180)</td>
<td>87.4/85.7 (166/14)</td>
<td>91.4/57.8 (116/64)</td>
</tr>
<tr>
<td>EC6</td>
<td>97.6 (296)</td>
<td>97.9/80.0 (286/10)</td>
<td>99.5/92.1 (208/88)</td>
<td>93.6 (296)</td>
<td>92.8/66.7 (278/18)</td>
<td>95.6/75.8 (205/91)</td>
</tr>
<tr>
<td>Total</td>
<td>93.5 (7442)</td>
<td>94.0/62.2 (7344/98)</td>
<td>94.8/80.2 (5267/2175)</td>
<td>83.5 (7258)</td>
<td>83.7/52.3 (7038/220)</td>
<td>84.8/69.9 (5443/1815)</td>
</tr>
</tbody>
</table>

3.3 SOM-based assignment of EC second and third digits from the reaction equation

SOMs were also used to classify reactions according to the second and third digits of the EC number (subclass and sub-subclass). Independent networks were trained for different classes. For the subclass experiments concerning classes EC1 and EC2, CPGNNs were used instead of SOMs due to the larger number of subclass to classify. The data sets were partitioned into training and test sets in two different ways. In Partition 1, the test set included one (random) reaction of each subclass (or sub-subclass), and the training set the remaining reactions. Accurate predictions of subclass were achieved for 62% of the test set (42% for EC1 reactions, 88% for EC2, 63% for EC3, 57% for EC4, 100% for EC5, and 80% for EC6). Correct prediction of the sub-subclass was obtained for 52% of the test set. In Partition 2, the test set was selected with a SOM as for the experiments at the class level. While the second test set has a similar distribution of subclasses (or sub-subclasses) to the whole data set, in the test set of Partition 1 the subclasses (or sub-subclasses) with smaller number of reactions have proportionally more impact on the calculated prediction accuracy. Accurate predictions of subclass were achieved for 80% of the test set of Partition 2 (74% for EC1 reactions, 86% for EC2, 87% for EC3, 70% for EC4, 71% for EC5 and 92% for EC6). Correct prediction of the sub-subclass was obtained for 70% of the test set. Table 4 shows the number of reactions in each training and test sets, and details the predictions for the second and third digits of the EC number. The size of the maps was chosen such that the number of reactions was approximately twice the number of neurons. Maps of size $35 \times 35$, $35 \times 35$, $25 \times 25$, $18 \times 18$, $10 \times 10$ and $12 \times 12$ were used for classes EC1, EC2, EC3, EC4, EC5, and EC6 respectively. CPGNNs of size $35 \times 35$ were used to assign sub-subclasses of the EC1 and EC2 classes.

The general trend for Partitions 2 to yield better predictions than Partitions 1 suggests that predictions are generally easier for subclasses (and sub-subclasses) with more examples available (in the test sets of Partitions 2 these are proportionally more abundant). It is thus expected that EC numbers can be more easily predicted when similar reactions are known catalyzed by the same enzyme with other substrates. Even though, 62% and 52% of the reactions in the test set (respectively at the subclass and sub-subclass levels) could be correctly predicted in the more severe situations of Partitions 1. Inspection of the results for individual subclasses (or sub-subclasses) reveals in general a lower percentage of correct predictions for subclasses (or sub-subclasses) with a smaller number of reactions in the data set. While the experiments at the class level with Partition 4 (Section 3.2 above) assessed the ability of the MOLMAPs to identify similarities between subclasses of the same class, these experiments with classifications at the sub-subclass level assess the ability to discriminate between different sub-subclasses. The surface of the best SOM obtained for the classification into subclasses of oxidoreductase and transferase reactions were illustrated in the Supplementary Material (Figures S4 and S5), as well as a map of the reactions of the hydrolase class, colored according to sub-subclasses.

3.4 Classification of racemases and epimerases reactions

The connectivity of reactants is not changed in isomerase reactions of subclass 5.1 (racemases and epimerases)—only stereochemical changes occur. MOLMAPs cannot perceive stereochemical features, and therefore cannot represent such reactions. We retrieved 18 reactions of subclass 5.1 from the BioPath database, where stereochemistry was assigned to chiral structures, and tried to represent reactions using chiral descriptors. Chirality codes were used as molecular descriptors instead of MOLMAPs. Reactions were represented by the difference of the chirality code of the product and the chirality code of the reactant. Then a SOM was trained to assess the ability to distinguish between sub-subclasses of racemases and isomerases (Supplementary Figure S6). Each reaction was included twice, corresponding to the two directions. Although the data set is rather small, the map shows some separation between the two main sub-subclasses (EC 5.1.1 acting on amino acids and derivatives, and EC 5.1.3 acting on carbohydrates and derivatives). Within each of these sub-subclasses, two regions are differentiated corresponding to the two directions of the reactions.

4. CONCLUSIONS

The results of unsupervised mapping of metabolic reactions show a general agreement with the EC classification. The general reasonable clustering of reactions according to the EC classification allowed...
for the SOMs to assign EC numbers at the class, subclass and sub-subclass levels for reactions of independent test sets with accuracies up to 92, 80, and 70%, respectively. These numbers reflect the similarity between reactions within the first three levels of the EC hierarchy. Accuracy of predictions was correlated with the number of votes in consensus predictions, and with the Euclidean distance to the winning neuron of SOMs.

Similarity of reactions across sub-subclasses (within the same class) was assessed by test sets only including reactions of sub-subclasses not available in the training set—68% of correct predictions were obtained for the class level. At the same time, experiments to predict the third digit of the EC number demonstrated the ability of MOLMAP descriptors to discriminate sub-subclasses.

The correspondence between chemical similarity of metabolic reactions and similarities in their MOLMAP descriptors was confirmed with a number of reactions detected in the same neuron but labeled with different EC classes. Such an exercise also demonstrated the possible application of the MOLMAP/SOM approach to the verification of internal consistency of classifications in databases of metabolic reactions. Conveniently, the MOLMAP method avoids the assignment of reaction centers and atom-to-atom mapping previous to the classification of reactions.

ACKNOWLEDGEMENTS

The authors thank ChemAxon Ltd (Budapest, Hungary) for access to JChem and Marvin software, Kyoto University Bioinformatics Center (Kyoto, Japan) for access to the KEGG database and Molecular Networks GmbH (Erlangen, Germany) for access to the BioPath Database and PETRA software.

Funding: Fundação para a Ciência e Tecnologia (Lisbon, Portugal) Ph.D. grant (SFRH/BD/18347 to D.A.R.S.L.)

Conflict of Interest: none declared.

REFERENCES


