Assignment of EC Numbers to Enzymatic Reactions with MOLMAP Reaction Descriptors and Random Forests

Diogo A. R. S. Latino and João Aires-de-Sousa*

CQFB, REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Received March 17, 2009

The MOLMAP descriptor relies on a Kohonen SOM that defines types of covalent bonds on the basis of their physicochemical and topological properties. The MOLMAP descriptor of a molecule represents the types of bonds available in that molecule. The MOLMAP descriptor of a reaction is defined as the difference between the MOLMAPs of the products and the reactants and numerically encodes the pattern of changes in bonds during a chemical reaction. In this study, a genome-scale data set of enzymatic reactions available in the KEGG database was encoded by the MOLMAP descriptors and was explored for the assignment of the official EC number from the reaction equation with Random Forests as the machine learning algorithm. EC numbers were correctly assigned in 95%, 90%, and 85% (for independent test sets) at the class, subclass, and subsubclass EC number level, respectively, with training sets including one reaction from each available full EC number. Increasing differences between training and test sets were explored, leading to decreased percentages of correct assignments. The classification of reactions only from the main reactants and products was obtained at the class, subclass, and subsubclass level with accuracies of 78%, 74%, and 63%, respectively.

INTRODUCTION

Large available databases of enzymatic reactions are emerging that can provide an increasingly detailed picture of the cells' biochemical machinery and its encoding on genomes.[1−6] The automatic perception of similarities between metabolic reactions, i.e. their classification, is a chemoinformatics issue with an impact in bioinformatics, biotechnology, or medicinal chemistry, as illustrated by the following examples.

Pinter et al. proposed a method for the alignment of metabolic pathways that advantageously took into account similarities between enzymatic reactions classified in terms of EC (Enzyme Commission) numbers.[7] EC numbers are the official classification of enzymatic functions and are assigned to the catalyzed chemical reactions.[8,9] Although designed for reactions, this classification system is often simultaneously employed as an identifier of enzymes and enzyme genes, linking metabolic and genomic information. The assignment or creation of EC numbers to new enzymes is performed by the Enzyme Commission and based on published experimental data that include the full characterization of the enzymes and their catalytic functions.

Yamanishi et al. have incorporated the estimation of EC numbers from substrates and products of reactions in a method to identify the genes coding for missing enzymes in the reconstruction of a metabolic pathway.[10] Studies of relationships between enzyme structure and function rely on a definition of similarities between biochemical reactions (enzyme functions) and have usually employed EC numbers for that purpose.[11−14] The definition and assignment of enzymatic reaction classes is of value also for the development of new biotechnological processes, where known biocatalytic reactions can be exploited in novel biochemical routes to specialty chemicals currently produced using organic chemistry.[15]

O'Boyle et al. defined fingerprints of enzymatic reaction features to assess the similarity between individual steps of enzymatic reaction mechanisms and to quantitatively measure the similarity of enzymatic reactions based upon their explicit mechanisms.[16] Ridder and Wagener clustered a data set of metabolic reactions using a difference fingerprint defined by the differences in occurrence of each Sybyl atom type in the reactant and product fingerprints.[17] Such a classification of metabolic reactions assisted in the establishment of reactivity rules for the prediction of metabolites in drug discovery processes.

The Kanehisa laboratory has developed Reaction Classification (RC) numbers to represent the conversion patterns of atom types in three regions of the molecular structures of reactants/products - the reaction center, the matched region, and the difference region.[18] The RC method, more specifically the identification of similarities between RC numbers of a query reaction and reactions in a database, was applied to the automatic assignment of EC numbers to enzymatic reactions in the KEGG database. Kotera et al. described a procedure for identifying possible products and/or precursors of orphan metabolites and for suggesting complete reaction equations and the corresponding EC (Enzyme Commission) number simultaneously.[19] The procedure incorporates decision trees and random forests to evaluate if an equation is possible and to assign EC numbers from general and structural features of the reactions. Faulon et al.[20] employed molecular signatures of topological atom neighborhoods to derive reaction signatures and used such descriptors with

* Corresponding author phone: (+351) 21 2948300; fax: (+351) 21 2948550; e-mail: jas@fct.unl.pt.

10.1021/ci900104b CCC: $40.75
Published on Web 07/09/2009

support vector machines to classify metabolic reactions in terms of EC numbers.

The Gasteiger and Funatsu groups pioneered the representation of chemical reactions by physicochemical features of the atoms and bonds at the reaction center.[21,22] The method was applied more recently to classify metabolic reactions of subclass EC 3.1 on the basis of physicochemical properties of the reactants by means of Kohonen Self-Organizing Maps (SOM).[23]

Our group has developed a method for numerically encoding the transformations resulting from a chemical reaction, the MOLMAP (MOLecular Mapping of Atom-level Properties) reaction descriptor.[24] The method was based on the mapping of the chemical bonds (described by physicochemical and/or topological properties existing in the structure of the reactants and products) on a SOM. The mapping of the chemical bonds of a molecule is converted into a numerical fixed-length fingerprint (the MOLMAP) describing the types of bonds available in each molecule. The reaction MOLMAP is obtained by the difference between the MOLMAPs of the products and the MOLMAPs of the reactants, which represents the types of bonds that disappeared from the reactants and those created in the products. This method is based on topological and physicochemical molecular features, which in principle are related to the reaction mechanism, does not require explicit assignment of the reaction center, and avoids atom-to-atom mapping.

The MOLMAP approach was applied in a preliminary study of classification of a genome-scale data set of enzymatic reactions[25] and later in the mapping of the enzymatic reactions in the KEGG database into SOMs.[26] The SOM algorithm, an unsupervised learning method, was able to show the agreement between EC numbers and MOLMAP-based classifications and revealed similarities between reactions catalyzed by enzymes of different EC classes. That study assessed the usefulness of the method for automatic comparison of enzymatic reactions and illustrated its application to the validation of reaction classification systems.

The present paper fully reports the exploration of a powerful supervised learning method - Random Forests - to assign EC numbers in the same data set of enzymatic reactions. The supervised nature of learning in RFs allowed for investigation of the possibility of representing enzymatic reactions using only the main reactants and products.

METHODOLOGY

Data Set of Chemical Reactions. Enzymatic reactions were extracted from the KEGG LIGAND database (release of November 2006)[1−4] with molecular structures in the MDL Molfile format. The data set of enzymatic reactions used is the same in ref 26. The original data set from the KEGG LIGAND database consists of 6810 reactions. Reactions that were listed with more than one assigned EC number (445 reactions), or no assigned EC number (959 reactions), or reactions with incomplete EC numbers (536 reactions) were removed. The data set with the remaining 4870 reactions was further processed as follows. In some files, general fragment symbols were replaced, such as 'X' by a chlorine atom, or 'R' by methyl, adenine, cytosine, or another fragment depending on the reaction. Structures with unclear general symbols were not changed, and the corresponding reactions were not used. Reactions involving a compound producing no output by the STANDARDIZER or CXCALC tools from the JChem package (ChemAxon, Budapest, Hungary, www.chemaxon.com) or reactions involving a compound with no MDL.mol file were removed. In these two steps 399 reactions were removed from the data set of 4870 reactions. From the remaining 4471 reactions, unbalanced reactions were removed (or manually balanced in some clear cases) - 76 reactions were removed. Detection of reaction duplicates was performed based on chemical hashed fingerprints of reactants and products generated by the GenerFP tool from the JChem package, with a length of 128 bytes, a maximum number of 10 bonds in patterns, and 3 bits switched on for each pattern in the structure. Reactions differing only in stereochemical features were also considered as duplicates and similarly were included only once. This step eliminated 551 reactions. The remaining data set of 3844 reactions was screened for null reaction MOLMAPs - 60 reactions were found and removed yielding a data set of 3784 reactions. Since most reactions are reversible, all reactions were represented in both directions. The procedure to exclude duplicated reactions was applied again after the representation in both directions, excluding 43 reactions. The whole process yielded a final data set of 3741 reactions (7482 when represented in both directions).

Table 1. Composition of the Data Set at Each Level of the EC System

	no. of classification groups	total	train	test
1 level	6 classes	3784	3156	628
2 level	49 subclasses	3764	3140	624
3 level	110 subsubclasses	3671	3066	605

Reactions represented only in one direction. The data set of 3784 reactions was partitioned into training and test sets using a 29 × 29 Kohonen SOM. The SOM was trained with all reactions, and after the training one reaction was randomly selected from each occupied neuron and moved to the test set.

The data set with the reactions represented in only one direction consisted of 3784 reactions. In the experiments to predict the subclass (or subsubclass), only subclasses (or subsubclasses) with 4 or more reactions were considered. The data set including the reactions represented in both directions (after removing the duplicates of the inverse reactions) consisted 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). The number of reactions used in the experiments to evaluate the impact of the MOLMAP descriptor size and type of bond descriptors are displayed in Table 1.

Generation of MOLMAP Reaction Descriptors. SOMs were used in this study for the generation of a molecular descriptor based on the classification of chemical bonds, as described elsewhere.[24,26] SOMs learn by unsupervised training, distributing objects (here bonds) through a grid of so-called neurons, on the basis of the objects properties (properties of chemical bonds). A SOM projects multidimensional objects (chemical bonds described by physicochemical and/or topological descriptors) into a 2D surface (a map), and the algorithm is designed to reveal similarities in the objects of a data set by mapping similar objects into the same or closely adjacent neurons. Each neuron of the

map contains as many elements (so-called weights) as there are input variables (objects features). Before the training starts, the weights take random values. During the training, each individual object is mapped into the neuron at the shortest Euclidean distance to its vector of features. This is the central neuron, or winning neuron, and its weights are then adjusted to make them even more similar to the properties of the presented bond. The neurons in its neighborhood also have its weights adjusted, but the extent of adjustment depends on the topological distance to the winning neuron - the closer a neuron is to the winning neuron the larger is the adjustment of its weights. The objects of the training set are iteratively fed to the map, and the weights corrected, until a predefined number of cycles is attained.

SOMs with toroidal topology were used in this study. Training was performed by using a linear decreasing triangular scaling function used with an initial learning rate of 0.1 and an initial learning span of 3 or 5. The weights were initialized with random numbers that were calculated using the mean and standard deviation of the input data set as parameters. The training was typically performed over 50, 75, or 100 cycles (the fluctuations in the results using different number of training cycles were at the same level of those resulting from starting the training of SOMs with different initial weights - 0 to 5% in the prediction accuracy), with the learning span and the learning rate linearly decreasing until zero. SOMs were implemented throughout this study with an in-house developed Java application derived from the JATOON Java applets.[27,28] The chemical bonds were represented by topological and physicochemical features (Table S1 of the Supporting Information). The SOM was trained with a diversity of bonds, taken from a diverse data set of molecules. After the training, the SOM can reveal similarities between chemical bonds and in this way can describe the diversity of bonds in a molecule. The bonds existing in a molecule can be represented as a whole by mapping all the bonds of that molecule onto the SOM. The pattern of activated neurons is interpreted as a fingerprint of the available bonds in the molecule, and it was used as a molecular descriptor (MOLMAP). A similar idea was proposed by Atalay et al. to encode the primary structure of proteins with SOM on the basis of aminoacid features.[29] For numerical processing, each neuron gets a value equal to the number of times it was activated by bonds of the molecule. The map is then transformed into a vector by concatenation of columns. In order to account for the relationship between similarity of bonds and proximity in the map a value of 0.3 was added to each neuron multiplied by the number of times a neighbor was activated by a bond.

The MOLMAP descriptor *of a reaction* is calculated as the difference between the MOLMAP of the products and the MOLMAP of the reactants. If there is more than one reactant in a reaction, the MOLMAPs of all reactants are summed, and the same for the products. The reaction MOLMAP is used in this work as a numerical representation of the structural changes operated by the reaction—a code of the reaction.

In this study, the SOM was trained with the chemical bonds from a diverse representative set of molecules involved in enzymatic reactions. The molecules were selected from the database of reactions by the Ward's minimum variance method[30−32] implemented in the WARD tool of the JChem package, based on molecular chemical hashed fingerprints with a length of 64 bytes, a maximum number of 5 bonds in patterns, and 2 bits switched on for each pattern in the structure. The Kelley method[33] was used to decide the number of clusters - 45. The central compounds of the clusters were chosen, and all their bonds were extracted to the training set of the SOM.

The original molfiles of the compounds were treated with the JChem STANDARDIZER tool to add hydrogens, clean stereochemistry, and aromatize. The physicochemical and topological descriptors listed in Table S1 of the Supporting Information were computed with in-house developed software from properties calculated with the JChem CXCALC tool. As the descriptors for a bond depend on the orientation of the bond, each bond was oriented from the atom with higher charge to the atom with lower charge. To make all the descriptors equally relevant in the training of the SOM, z-normalization was applied to each of the descriptors 42−68 (the set of 27 physicochemical descriptors) based on the whole data set of chemical bonds. Descriptors 1−14 and 35−41 were multiplied by 3. This normalization scheme was manually chosen in a different study, with a different set of compounds, to produce as much as possible mappings of bonds in accordance with chemical intuition (concerning similarity of bonds).

SOMs of sizes 7 × 7, 10 × 10, 15 × 15, 20 × 20, 25 × 25, and 29 × 29, yielding MOLMAPs of dimension 49, 100, 225, 400, 625, and 841, respectively, were trained with a data set of 1568 bonds extracted from the selected compounds using the Ward method. Experiments were performed using exclusively 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, and 68 from Table S1 of the Supporting Information. These MOLMAPs with different sizes and using different bond descriptors were later tried for the classification of enzymatic reactions at the first three levels of the EC system.

Random Forests Assignment of EC Numbers. With MOLMAP reaction descriptors calculated for the different data sets, automatic assignment of EC numbers was performed by Random Forests (RF) for the three first levels of the EC classification system. A Random Forest[34,35] is an ensemble of unpruned classification trees created by using bootstrap samples of the training data and random subsets of variables to define the best split at each node. It is a high-dimensional nonparametric method that works well on large numbers of variables. The predictions are made by majority voting of the individual trees. It has been shown that this method is very accurate in a variety of applications.[35] Additionally, the performance is internally assessed with the prediction error for the objects left out in the bootstrap procedure (out-of-bag estimation, OOB). In this work, RFs were grown with the R program,[36] using the randomForest library.[37] The number of trees in the forest was set to 1000, and the number of variables tested for each split was set to default (square root of the number of variables). The data set of reactions was partitioned into a training and a test set, using different criteria depending on the experiments (see section Results and Discussion). The voting system of a RF allows the association of a probability to each prediction that

Table 2. Impact of MOLMAP Parameters on the Accuracy of EC Number Assignment by RF

		% incorrect predictions[c]								
		1st level			2nd level			3rd level		
MOLMAP size[a]	set of bond descriptors[b]	all	tr	te	all	tr	te	all	tr	te
49	T	19.32	19.42	22.61	33.40	33.89	35.42	37.10	37.77	41.98
	PC1	11.44	10.96	17.20	19.95	19.78	25.32	23.26	23.91	30.41
	PC1T	18.23	18.00	21.82	28.21	28.63	30.77	33.97	33.20	38.02
	PC2T	20.35	20.09	23.73	33.79	34.68	37.34	37.95	38.01	43.14
100	T	10.17	9.66	14.65	18.25	18.41	23.72	23.78	24.08	30.41
	PC1	9.12	8.75	15.45	16.50	17.52	20.99	19.56	20.91	25.12
	PC1T	7.53	7.38	12.10	18.20	19.08	20.51	24.30	24.72	30.41
	PC2T	10.60	10.17	15.29	20.43	20.67	24.68	25.01	25.24	31.40
225	T	8.27	8.24	12.42	15.14	15.48	19.23	18.71	19.50	23.64
	PC1	7.43	7.16	12.58	14.35	15.16	18.43	15.83	17.78	21.65
	PC1T	6.50	6.27	10.83	13.26	13.54	17.63	16.04	16.99	21.49
	PC2T	6.45	6.18	10.67	14.16	14.94	18.91	17.62	18.59	23.97
400	T	6.37	6.46	10.03	12.59	13.82	16.51	16.32	16.77	20.83
	PC1	6.92	6.81	11.15	13.71	14.81	18.11	15.77	16.99	21.98
	PC1T	6.21	6.31	9.55	12.11	12.80	16.19	14.41	15.92	20.33
	PC2T	6.53	6.43	9.87	12.54	13.22	17.63	15.45	15.99	20.17
625	T	6.34	6.34	9.55	12.49	13.25	16.67	16.07	16.84	20.99
	PC1	7.16	7.54	11.46	13.42	14.87	18.39	15.36	16.37	21.82
	PC1T	6.05	6.21	9.55	12.33	13.06	17.31	14.14	14.87	20.50
	PC2T	6.32	6.34	10.99	12.01	13.76	16.83	13.70	14.68	20.99
841	T	6.63	6.40	10.67	13.04	13.57	17.31	16.18	17.09	21.65
	PC1	6.42	6.65	10.99	13.15	14.43	18.11	14.85	16.11	21.16
	PC1T	6.40	6.18	10.19	12.06	12.71	16.99	13.67	14.51	20.00
	PC2T	6.21	6.12	10.51	11.90	12.74	17.15	13.81	14.88	19.67

[a] 49, 100, 225, 625, and 841 are the MOLMAP size. [b] T - topological bond descriptors (41 descriptors); PC1 - physicochemical bond descriptors (27 descriptors); PC1T - physicochemical bond descriptors and topological bond descriptors (27 + 41 descriptors); PC2T - subset of physicochemical bond descriptors and topological bond descriptors (14 + 41 descriptors). [c] all - out of bag predictions from a model trained with all data; tr - out of bag predictions from a model trained with a training set; te - predictions for an independent test set.

reflects the percentage of votes obtained by the winning class. This probability was investigated as a measure of reliability.

To evaluate the influence of MOLMAP size and type of bond descriptors on the accuracy of the EC assignments at the first three levels of the EC system, experiments were performed with MOLMAPs of different sizes and using the different sets of descriptors.

RESULTS AND DISCUSSION

Evaluation of the Impact of MOLMAP Parameters (Size and Bond Descriptors). The influence of MOLMAP size and type of MOLMAP bond descriptors on the accuracy of the predictions for the first three levels of EC numbers (class, subclass and subsubclass) was evaluated with the data set of enzymatic reactions represented only in the direction of the KEGG reaction file - 3784 reactions. Random Forests were trained with MOLMAPs of dimension 49, 100, 225, 400, 625, and 841 each one generated using the four sets of bond descriptors (topological, physicochemical, topological + physicochemical, topological + subset of physicochemical). Experiments were performed for the assignment of the first three levels of the EC number, and the results were analyzed using the internal cross-validation of the RF obtained by out-of-bag (OOB) estimation for the training set, or the predictions for a test set. As mentioned in the Methodology section (Table 1), the data set of 3784 reactions was partitioned into training and test sets using a Kohonen SOM. The results are displayed in Table 2. In general the results for the test set were ~5% worse than the OOB results for the total data sets or for the training set.

The following discussion is based on the results of the internal cross-validation (OOB estimation) for the training sets. MOLMAPs of size 49 (7 × 7) were clearly inferior to larger sizes, suggesting an insufficient resolution to distinguish between the different types of chemical bonds, and so to encode the different types of chemical changes operated in the reactants by a reaction. The variation of the MOLMAP size between 100 (10 × 10) and 841 (29 × 29) affected the accuracy of the predictions (the best result compared to the worst) in 4.1, 8.0, and 10.7% for the first three levels of the EC hierarchy depending on the type of bond descriptors. In experiments performed with the same MOLMAP size (100), the type of descriptors affected the accuracy at most in 2.8, 3.2, and 4.3% for the first three levels of the EC number, respectively. With MOLMAPs of size 400, the accuracy of the predictions was only affected in 0.5, 2.0 and 1.1%. With MOLMAPs of size larger than 100, topological descriptors generally performed poorer than the others. The best results were obtained using the combination of topological descriptors and the subset of physicochemical descriptors for MOLMAPs of sizes 625 and 841. No significant benefit was observed from increasing the MOLMAP size from 625 to 841. MOLMAPs of size 625 using topological descriptors and the subset of physicochemical bond descriptors were thus chosen for the subsequent experiments. This combination yielded an error of 6.3, 13.8, and 14.7% for the assignment of the class, subclass and subsubclass, respectively.

RFs Assignment of EC Numbers from the Reaction Equation. MOLMAPs of size 625 (25 × 25) generated using topological and the subset of physicochemical bond descriptors were used for the automatic assignment of the EC

Table 3. Classification of Enzymatic Reactions by RFs

		% incorrect predictions (number of reactions)			
		1st level		2nd level	3rd level
data sets		RFs	SOMs[a]	RFs	RFs
all	data	6.19 (7482)	–	11.07 (7442)	13.25 (7258)
partition	training 1	5.84 (5855)	–	11.62 (5794)	14.03 (5659)
1[b]	test 1	9.17 (1646)	15.7	15.95 (1624)	18.60 (1575)
partition	training	7.68 (5246)	–	15.35 (5206)	17.34 (5046)
2[c]	test	5.05 (2236)	8.3	10.24 (2236)	15.05 (2212)
partition	training	28.57 (350)	–	– (320)	–
3	test 1	20.08 (4896)	25.6	49.51 (4886)	–
(subsub)[d]	test 2	18.28 (7132)	24.3	48.16 (7122)	–
partition	training	29.68 (310)	–	–	–
4[e]	test	27.50(40)	32.5	–	–
partition	training	8.7 (4576)	–	17.4 (4534)	19.3 (4366)
5[f]	test	8.1 (1482)	12.1	13.7 (1476)	19.3 (1460)

[a] SOMs - results obtained from an ensemble of 10 SOMs (reported in ref 26 except for partition 5). [b] Partition 1 - The data set with all reactions was partitioned into training and test sets using a 49 × 49 Kohonen SOM. The SOM was trained with all reactions, and after the training one reaction was randomly selected from each occupied neuron and moved to the test set. [c] Partition 2 - Training set with one reaction of each EC number of the data set and test set with the remaining reactions. [d] Partition 3 - Training set with one reaction of each subsubclass, 350 reactions. Two test sets, the first with one reaction of each EC number of the remaining reactions, 4896 reactions, and the second set with all reactions excluding the 350 used in training, 7132 reactions. [e] Partition 4 - Training set with one reaction of each subsubclass excluding 40 reactions for the test set. Training set with 310 reactions and test set with 40. [f] Partition 5 - Data set of 6058 reactions without duplicated reaction MOLMAPs: training set with one reaction of each EC number and test set with the remaining reactions.

numbers, at the first three levels of the EC hierarchy, from the structures of all reactants and products (as described in the KEGG reaction file) using Random Forests. Several partitions in training and test sets using different criteria were explored, to assess the robustness of the method on exercises of increasing difficulty.

The results obtained with the different partitions are shown in Table 3 and compared, when possible, with those obtained in a previous study with unsupervised Self-Organizing Maps (SOM).[26] In all experiments, the reactions were represented in both directions. The results presented for the experiments with all data and for training sets were from the internal cross-validation of RFs obtained by out-of-bag (OOB) estimation.

The first experiment (all data) was carried out with all the available data at each level of classification. Errors of 6.19, 11.07, and 13.25% were obtained for the EC class, subclass, and subsubclass.

In partition 1 the training and test sets were selected using a 49 × 49 SOM. The SOM was trained with all reactions, and then the test set was chosen by random selection of one reaction from each occupied neuron, resulting in a training and a test set with 5855 and 1646 reactions, respectively. Wrong predictions were obtained for 9.17, 15.95, and 18.60% of the test set for the first, second, and third digit of the EC system, respectively. Partition 1 is the only one where it is not guaranteed that the two entries for each reaction (corresponding to both directions) were included in the same set (training or test).

Partition 2 relied on a different criterion to cover the reaction space as much as possible. In this case one reaction of each full EC number was randomly selected to the training set, and the remaining reactions labeled with the same full EC number were moved to the test set. With this partition every full EC number represented in the test set was also represented in the training set. There are some EC numbers represented in the training set without reactions in the test set. The high similarity between training and test sets was reflected in the results. The results for the test sets were better than the OOB estimation for training sets. As each full EC number is unique in the training set, the results for the OOB estimations illustrate the ability of the model to classify reactions belonging to full EC numbers not available in the training.

The next partitions were designed to reduce the similarity between the training and test sets and therefore to test the ability of the MOLMAP approach to assign EC numbers to reactions increasingly different from those provided during the training. In partition 3 the model was trained only with one reaction of each subsubclass (350 reactions), which means that all reactions in the training set were different at the subsubclass level, and all subsubclasses were represented in the training set. Two test sets were used: the first including one reaction from each EC number not included in the training set (4896 reactions) and the second including all the reactions not included in the training set (7132 reactions). This means the first test set does not include any reaction with a full EC number available in the training set. For the class level, 80% and 82% of the test sets were correctly assigned despite the rather small size of the training set, and the fact that in the training set there is only one reaction of the same subsubclass of each reaction in the test sets. The high diversity within the training set explains the higher percentage of wrong predictions in the OOB estimation (for the training set) than for the test set. The low level of similarity between training and test set is reflected in the results for the subclass level. The percentage of correct assignments decreases to 50% and 52% for the two test sets.

With partition 4 all reactions of the test set are from different subsubclasses of the training set. This set was built by random selection of 40 reactions from the data set of 350 reactions used in partition 3 (which contained only one reaction of each subsubclass). The test and training set were predicted with an error of ~30% for the class level and with approximately the same accuracy as the OOB estimation for the training set of partition 3 (which has the same meaning, i.e. predicting the class with no cases of the same subsubclass in the training). The results with partitions 3 and 4 are also indicative of the heterogeneity of reactions within the same class and subclass.

Before generating partition 5, duplicated reaction MOLMAPs were removed from the data set of 7482 reactions - these were generally pairs of very similar reactions, although involving different compounds. The remaining 6058 reactions were then divided in training and test sets in the same way as partition 2 (random selection of one reaction of each available full EC number for the training set). The results for this partition were only slightly inferior to those obtained with partition 2 (3−4% decreased accuracy for the three first digits of the EC number).

The comparison of results obtained with RFs with those previously obtained[26] with ensembles of 10 SOMs reveals 3−6% higher percentages of correct predictions by the RFs for the EC number class level in all experiments with a test

Table 4. Confusion Matrix for the Classification of Enzymatic Reactions According to the First Digit of the EC Number (Test Set of Partition 2)

	EC1	EC2	EC3	EC4	EC5	EC6	% incorrect predictions
EC1	950	8	6	4	−	4	2.26
EC2	10	612	22	8	−	−	5.85
EC3	−	2	372	2	−	2	1.59
EC4	2	3	18	103	−	−	18.25
EC5	−	10	−	8	16	−	52.94
EC6	−	4	−	−	−	70	5.41

set. The experiments with RFs and SOMs were performed using exactly the same MOLMAPs reaction descriptors and training and test sets. For the assignment of the second and third digits, rigorous comparisons are not possible since training and test set partitions were not exactly the same at those levels for the SOM experiments. For the prediction of subclass and subsubclass, each SOM only processed reactions from a single class, while a RF was trained with reactions of all classes simultaneously, even though the 15.95% and 18.6% incorrect predictions obtained by the RFs for the second and third digits (test sets of partition 1) compare favorably with the 19.8% and 30.1% incorrect predictions obtained by the SOMs for test sets of comparable sizes obtained using a similar partition procedure.

One additional experiment, with an independent test set, was performed with a data set of 930 enzymatic reactions from the KEGG database with an incomplete EC number (both directions included). The RFs model trained with all reactions (7482 reactions) with a full EC number were applied to assign the first EC digit to this data set of reactions. The first digit was correctly assigned in 82.7% of such cases. In the same experiment using an ensemble of 10 SOMs[26] only 73.8% of the reactions had the first digit correctly assigned.

Table 4 shows the confusion matrix obtained for the test set of partition 2 at the class level (2236 reactions).

The confusion matrix shows a higher prediction accuracy for the most represented classes (oxidoreductases, transferases, and hydrolases) and for ligases, and the worst was for lyases and isomerases with 18% and 53% of uncorrect predictions. Reactions catalyzed by isomerases, EC5, are more difficult to classify as they usually involve no substantial structural changes, and also they are the least represented in the data set. Reactions catalyzed by lyases, EC4, gave 18.25% of the errors (23 reactions out of 126) with 18 reactions classified as hydrolases. This result suggests that the patterns of reactions catalyzed by lyases are similar in many cases to hydrolases. Hydrolases present the higher prediction accuracy, with only 6 wrongly classified reactions. At the same time it is the class of reactions with the larger number of false positives (18 lyases and 22 transferases wrongly classified as hydrolases). This fact may indicate possible inconsistencies in EC numbers for the classification of these reactions. These results are in general agreement with those obtained by SOMs.[26]

The usefulness of the probability associated by a RF to each prediction as a measure of the reliability of the prediction was assessed by comparing prediction accuracies attained above several thresholds of probability. Table 5 shows the results for the prediction of the EC class with the test set of partition 2. Prediction with a probability higher than 0.5 was observed for 2098 reactions out of 2236 (93.8%), and 2035 of these were correct at the class level. The number of reactions decreases to 1434 (64%) if we consider only reactions predicted with a probability higher than 0.9, but in this case almost all reactions were correctly classified (1424 out of 1434).

The results are similar for the same experiment performed at the subclass level. The percentages of correct predictions for each level of probability are almost the same. For probabilities higher than 0.5, 0.7, and 0.9 the reactions are correctly assigned in 97.2%, 98.5%, and 99.2% of the cases at the subclass level, respectively. The main difference is in the number of reactions predicted with these levels of probability. In the first case 93.8%, 84.0%, and 64.1% of the entire data set are predicted with probabilities higher than 0.5, 0.7, and 0.9 respectively, while in the second case the percentages of reactions predicted with these probabilities decrease to 81.8%, 69.5%, and 45.8%.

Still the same experiment was performed for the test set 1 of partition 3, at the subclass level (where only ∼50% of the reactions had been correctly assigned). For probabilities higher than 0.5, 0.7, and 0.9 the subclass is correctly predicted in 84.2%, 85.8%, and 89.5% of the cases, respectively. However, the number of reactions predicted with these probability values decreases considerably when compared with the other experiments. In this case the percentages of reactions predicted with a probability of 0.5, 0.7, and 0.9 is 21.1%, 12.0%, and 3.3%. If the considered probability threshold is decreased to 0.4, then 32% of the data set is covered, with 79.8% of correct assignments at the subclass level.

In principle, a reaction is defined at the subsubclass level in the EC system, and the fourth digit is an identifier of the enzyme that does not only depend on the reaction. So, the prediction of the fourth digit from the reaction equation is not so meaningful. However, as many enzymes have several reactions reported, we attempted in those cases to predict also the last digit of the EC number. A data set was compiled with 1343 reactions corresponding to 110 different full EC numbers (only EC numbers with four or more reactions were considered). From these, one reaction of each full EC number (in both directions) was randomly selected to the test set (220 reactions), and the remaining were used as a training set (1122 reactions). The full EC number could be correctly assigned for 86% of the test set, and in the OOB estimation of the training set an error of 19% was observed. These numbers not only confirm the similarity, in general, between reactions catalyzed by the same enzyme (with the same EC number) but also indicate a relatively large number of reactions which are significantly different despite featuring the same EC number - and thus probably being catalyzed by the same enzyme.

RFs Assignment of EC Numbers from the Main Reactants and Products. MOLMAPs of reactions are "difference MOLMAPs" and therefore require, in principle, reaction equations to be balanced. However, reactions are often not balanced, and ideally a classification system would be able to classify reactions only from main reactants/products. We explored the possibility of training RF to predict EC numbers of reactions represented only by their main reactants and products (as described in the KEGG database).

Table 5. Relationship between the Prediction Accuracy and the Probability Associated with Each Prediction by RFs for the Test Set of Partition 2[a]

	probability					
	≥0.5		≥0.7		≥0.9	
	no. reactions	no. correct (%)	no. reactions	no. correct (%)	no. reactions	no. correct (%)
EC1	938	930 (99.2)	879	875 (99.5)	736	734 (99.7)
EC2	596	584 (98.0)	549	543 (98.9)	392	390 (99.5)
EC3	383	356 (93.0)	338	326 (96.5)	241	235 (97.5)
EC4	97	84 (86.6)	45	43 (95.6)	18	18 (100)
EC5	11	11 (100)	4	4 (100)	0	–
EC6	73	70 (95.9)	64	64 (100)	47	47 (100)
Total	2098	2035 (97.0)	1879	1855 (98.7)	1434	1424 (99.3)

[a] Classification according to the first digit of the EC number.

Table 6. Classification of Enzymatic Reactions by RFs Based on the Main Reactants and Products[a]

data sets		% incorrect predictions (number of reactions)		
		1st level	2nd level	3rd level
main compounds (main cpd)	all	23.48 (3850)	28.32 (3800)	32.76 (3632)
	training set	29.39 (2644)	36.71 (2596)	42.78 (2452)
	test set	21.56 (1206)	26.50 (1204)	37.37 (1180)
full reactions + main compounds	all	11.42 (11332)	14.81 (11242)	17.27 (10890)
	training set	14.54 (7890)	20.01 (7802)	22.73 (7498)
	test set	11.30 (3442)	14.97 (3440)	22.49 (3392)
		5.72 (2236 full rxns)	9.26 (2236 full rxns)	14.24 (2212 full rxns)
		21.64 (1206 main cpd)	25.58 (1204 main cpd)	37.97 (1180 main cpd)

[a] Main compounds (main cpd) - MOLMAPS encoded based on the main reactants and products of each reaction. Full reactions (full rxns) - MOLMAPS encoded based on the full equation.

Experiments were performed to train RFs with MOLMAPs of reactions calculated only from the main reactants and products and to get predictions for new reactions represented in the same way. The training set was built with one reaction of each EC number of the data set and the test set with the remaining reactions. Additionally, RFs were trained with a data set encompassing reactions represented by all reactants and products (full reaction) and reactions represented only by the main reactants and products (main compounds). The partition into training and test sets was also done by taking one reaction of each EC number into the training set. For the reactions represented by all reactants and products the training and test sets were those of partition 2 in Table 3. For the reactions represented by main reactants/products a new partition was generated because the structure of the data set was significantly changed. In this case, the predictions for the test set were separately calculated for full reactions and reactions with main compounds. The two directions of a full reaction were always included in the same set (training or test), and the same happened with reactions represented only by main compounds. However, in some cases, the full reaction was in the training set, while the same reaction represented by main compounds was in the test set and vice versa. The results are presented in Table 6.

The results show that the accuracy of the predictions is decreased in ca. 20% if only main reactants and products are used. Intermediate results are obtained if incomplete and complete reactions are mixed. In that case, if accuracies are separately calculated for the incomplete and complete reactions of the test set, the results for the complete reactions are similar to those obtained from RF trained only with complete reactions. The same happens with incomplete reactions. These observations suggest that the patterns of incomplete and complete reactions are processed independently by the RF, even for those belonging to a common classification.

Finally, RF trained only with complete reactions were applied to predict EC numbers for incomplete reactions. The prediction set included reactions that were available in the training set in the complete form. The incorrect predictions were 59% for the class level and 69% for the subclass and subsubclass levels. The opposite experiment (training with incomplete reactions and testing with complete reactions) yielded errors of 39%, 52%, and 55% for the class, subclass, and subsubclass levels.

CONCLUSIONS

The results demonstrate the possibility of applying Random Forests to the automatic assignment of EC numbers with better accuracy than the unsupervised method (Kohonen Self-Organizing Maps) used in previous studies. The accuracies of predictions reached 92%, 85%, and 83% for the class, subclass, and subsubclass level if several examples of reactions belonging to the same subsubclass are available in the training set. RF advantageously associate a probability to each prediction that correlated well with the observed accuracy for independent test sets, although probabilities enabling a significant increase in correct classification rates could only be attained for a reduced number of reactions. In the absence of information for reactions belonging to the same subsubclass, accuracies drop to 70% in the prediction of the class. This illustrates the heterogeneity of subclasses within the same subclass for a large number of cases.

This study also shows the possibility of training a single RF with complete reactions together with reactions

represented only by their main reactants and products and to still obtain accurate predictions for complete reactions, while predictions for incomplete reactions are ~20% less accurate. The results show that (even with high percentages of errors) it is easier to predict EC numbers for complete reactions from information on incomplete reactions than vice versa.

ACKNOWLEDGMENT

Diogo A. R. S. Latino acknowledges Fundação para a Ciência e a Tecnologia (Ministério da Ciência, Tecnologia e Ensino Superior, Lisbon, Portugal) for financial support under a Ph.D. grant (SFRH/BD/18347). The authors thank ChemAxon Ltd. (Budapest, Hungary) for access to JChem and Marvin software and Kyoto University Bioinformatics Center (Kyoto, Japan) for access to the KEGG database.

Supporting Information Available: List of bond descriptors used to implement the MOLMAP descriptors. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Goto, S.; Nishioka, T.; Kanehisa, M. LIGAND: Chemical database for enzyme reactions. *Bioinformatics* **1998**, *14*, 591–599.
(2) Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30.
(3) Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **2004**, *32*, D277−D280.
(4) Kanehisa, M. A database for post-genome analysis. *Trends Genet.* **1997**, *13*, 375–376.
(5) Caspi, R.; Foerster, H.; Fulcher, C. A.; Kaipa, P.; Krummenacker, M.; Latendresse, M.; Paley, S.; Rhee, S.; Shearer, A. G.; Tissier, C.; Walk, T. C.; Zhang, P.; Karp, P. D. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **2008**, *36*, D623−D631.
(6) Matthews, L.; Gopinath, G.; Gillespie, M.; Caudy, M.; Croft, D.; de Bono, B.; Garapati, P.; Hemish, J.; Hermjakob, H.; Jassal, B.; Kanapin, A.; Lewis, S.; Mahajan, S.; May, B.; Schmidt, E.; Vastrik, I.; Wu, G.; Birney, E.; Stein, L.; D'Eustachio, P. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **2009**, *37*, D619−D622.
(7) Pinter, R. Y.; Rokhlenko, O.; Yeger-Lotem, E.; Ziv-Ukelson, M. Alignment of metabolic pathways. *Bioinformatics* **2005**, *21*, 3401–3408.
(8) Barrett, A. J.; Canter, C. R.; Liebecq, C.; Moss, G. P.; Saenger, W.; Sharon, N.; Tipton, K. F.; Vnetianer, P.; Vliegenthart, V. F. G. *Enzyme Nomenclature*; Academic Press: San Diego, U.S.A., 1992.
(9) Tipton, K.; Boyce, S. History of the enzyme nomenclature system. *Bioinformatics* **2000**, *16*, 34–40.
(10) Yamanishi, Y.; Mihara, H.; Osaki, M.; Muramatsu, H.; Esaki, N.; Sato, T.; Hizukuri, Y.; Goto, S.; Kanehisa, M. Prediction of missing enzyme genes in a bacterial metabolic network - reconstruction of the lysine-degradation pathway of pseudomonas aeruginosa. *FEBS J.* **2007**, *274*, 2262–2273.
(11) Devos, D.; Valencia, A. Practical limits of function prediction. *Proteins* **2000**, *41*, 98–107.
(12) Orengo, C. A.; Pearl, F. M.; Bray, J. E.; A. E, T.; Martin, A. C.; Conte, L. L.; Thornton, J. M. The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.* **1999**, *27*, 275–279.
(13) Shaknovich, B. E.; Harvey, J. M. Quantifying structure-function uncertainty: A graph theoretical exploration into the origins and limitations of protein annotation. *J. Mol. Biol.* **2004**, *337*, 933–949.
(14) Todd, A. E.; Orengo, C. A.; Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **2001**, *307*, 1113–1143.
(15) Li, C. H.; Henry, C. S.; Jankowski, M. D.; Ionita, J. A.; Hatzimanikatis, V.; Broadbelt, L. J. Computational discovery of biochemical routes to specialty chemicals. *Chem. Eng. Sci.* **2004**, *59*, 5051–5060.
(16) 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.
(17) Ridder, L.; Wagener, M. SyGMa: Combining expert knowledge and empirical scoring in the prediction of metabolites. *ChemMedChem* **2008**, *3*, 821–832.
(18) Kotera, M.; Okuno, Y.; Hattori, M.; Goto, S.; Kanehisa, M. Computational assignement of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.* **2004**, *126*, 16487–16498.
(19) Kotera, M.; McDonald, A. G.; Boyce, S.; Tipton, K. F. Eliciting possible reaction equations and metabolic pathways involving orphan metabolites. *J. Chem. Inf. Model.* **2008**, *48*, 2335–2349.
(20) Faulon, J.-L.; Misra, M.; Martin, S.; Sale, K.; Sapra, R. Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. *Bioinformatics* **2008**, *24*, 225–233.
(21) Rose, J. R.; Gasteiger, J. HORACE: An automatic system for the hierarchical classification of chemical reactions. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 74–90.
(22) 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.
(23) Gasteiger, J. Modeling chemical reactions for drug design. *J. Comput.-Aided Mol. Des.* **2007**, *52*, 21–33.
(24) 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.
(25) Latino, D. A. R. S.; Aires-de-Sousa, J. Genome-scale classification of metabolic reactions: A chemoinformatics approach. *Angew. Chem., Int. Ed.* **2006**, *45*, 2066–2069.
(26) 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.
(27) JATOON applets. http://www.dq.fct.unl.pt/staff/jas/jatoon (accessed March 2009).
(28) Aires-de-Sousa, J. JATOON: Java tools for neural networks. *Chemom. Intell. Lab. Syst.* **2002**, *61*, 167–173.
(29) Atalay, V.; Cetin-Atalay, R. Implicit motif distribution based hybrid computational kernel for sequence classification. *Bioinformatics* **2005**, *21*, 1429–1436.
(30) Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Statist. Assoc.* **1963**, *58*, 236–244.
(31) Murtagh, F. *Review of Fast Techniques for Nearest Neighbour Searching*; Physica-Verlag: Vienna, Austria, 1984.
(32) El-Hamdouchi, A.; Willett, P. Hierarchic Document Clustering Using Ward's Method. In *SIGIR'86*. Proceedings of the Ninth International ACM SIGIR Conference on Research and Development in Information Retrieval; Pisa, Italy, September 8–10, 1986; ACM, 1986; pp 149–156.
(33) Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally-related subfamilies. *Protein Eng.* **1996**, *9*, 1063–1065.
(34) Breiman, L. Random forests. *Machine Learn.* **2001**, *45*, 5–32.
(35) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
(36) *R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing*; Vienna, 2004. ISBN 3-900051-07-0, http://www.R-project.org (accessed March 2009).
(37) Fortran original by Leo Breiman and Adele Cutler, R port by Andy Liaw and Mathew Wiener, 2004. http://www.stat.berkeley.edu/users/breiman/ (accessed March 2009).

CI900104B