BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 3 2013, pages 365–372
doi:10.1093/bioinformatics/bts700

Data and text mining

Advance Access publication December 5, 2012

ECOH: An Enzyme Commission number predictor using mutual information and a support vector machine

Yoshihiko Matsuta, Masahiro Ito and Yukako Tohsato*

Department of Bioinformatics, College of Life Sciences, Ritsumeikan University, Shiga, Kusatsu 525-8577, Japan

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The enzyme nomenclature system, commonly known as the enzyme commission (EC) number, plays a key role in classifying and predicting enzymatic reactions. However, numerous reactions have been described in various pathways that do not have an official EC number, and the reactions are not expected to have an EC number assigned because of a lack of articles published on enzyme assays. To predict the EC number of a non-classified enzymatic reaction, we focus on the structural similarity of its substrate and product to the substrate and product of reactions that have been classified.

Results: We propose a new method to assign EC numbers using a maximum common substructure algorithm, mutual information and a support vector machine, termed the Enzyme COmmission numbers Handler (ECOH). A jack-knife test shows that the sensitivity, precision and accuracy of the method in predicting the first three digits of the official EC number (i.e. the EC sub-subclass) are 86.1%, 87.4% and 99.8%, respectively. We furthermore demonstrate that, by examining the ranking in the candidate lists of EC sub-subclasses generated by the algorithm, the method can successfully predict the classification of 85 enzymatic reactions that fall into multiple EC sub-subclasses. The better performance of the ECOH as compared with existing methods and its flexibility in predicting EC numbers make it useful for predicting enzyme function.

Availability: ECOH is freely available via the Internet at http://www.bioinfo.sk.ritsumei.ac.jp/apps/ecoh/. This program only works on 32-bit Windows.

Contact: yukako@sk.ritsumei.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 15, 2012; revised on October 30, 2012; accepted on November 30, 2012

1 INTRODUCTION

In the automatic prediction of protein functions and their evolutionary relations on a large scale, the computational prediction of the catalytic function of enzymes has traditionally been based on sequence similarities (Yu *et al.*, 2009). The usefulness of this approach may be questioned because small changes in key residues may greatly alter enzyme function, which necessitates alternative approaches to assess enzyme similarities. Several researchers have focused on the similarity between functional motifs and protein structures and on the correlation between levels of protein expression in cells. Systematic annotation systems have been provided for generating and testing biological hypotheses, although one should note any propagation of functional mis-annotation in the systems (Furnham *et al.*, 2009). For example, Gene Ontology (GO) is a controlled vocabulary that can be used to describe gene products (Ashburner *et al.*, 2000). Structural Classification Of Proteins (SCOP) (Andreeva *et al.*, 2008) and Class, Architecture, Topology, Homologous superfamily (CATH) (Cuff *et al.*, 2011) regard any pair of enzymes that share protein domains as being similar. The Kyoto Encyclopedia for Genes and Genomes (KEGG) provides annotation of biochemical pathways for genes (Kanehisa *et al.*, 2012). Other enzyme classification systems, such as the enzyme commission (EC) classification scheme (IUPAC-IUBMB, 1999) and the reaction classification (RC) system (Kotera *et al.*, 2004b), are based on the chemical reaction catalysed by the enzyme under consideration. An EC number consists of the letters 'EC' followed by four digits separated by periods (e.g. EC 1.1.1.1). The first, second and third number are termed class, subclass and sub-subclass, respectively. The EC system consists of six classes: oxidoreductase (EC 1), transferase (EC 2), hydrolase (EC 3), lyase (EC 4), isomerase (EC 5) and ligase (EC 6). The subclass and sub-subclass specify the type of enzymatic reaction and its substrate requirements, respectively. In this manner, a hierarchically structured system is developed on the basis of the main EC classes; for example, EC 1 is divided into 23 EC subclasses (ECs 1.1–1.22 and EC 1.97; version 2011). Although the system is designed to classify enzymatic reactions based on their EC numbers, enzymes themselves can be hierarchically grouped by their function (Tohsato *et al.*, 2000). Owing to the fact that EC numbers are manually assigned by the nomenclature committee of the International Union of Biochemistry and Molecular Biology (IUBMB), many enzymatic reactions have either incomplete EC numbers or no EC number at all (Holliday *et al.*, 2012).

Several studies have proposed methods to predict EC numbers (Nath and Mitchell, 2012). A group of researchers attempted to use protein sequences and structures to predict EC numbers (Bray *et al.*, 2009; Dobson and Doig, 2005; Ferrari *et al.* 2012; Lu *et al.*, 2007). Bray *et al.* (2009) achieved an accuracy of 33% by using statistical analysis to predict the top EC class. Furthermore, Dobson and Doig (2005) and Lu *et al.* (2007) improved the accuracy to 35% and 86%, respectively, by using a support vector machine (SVM). Recently, Ferrari *et al.* (2012) achieved an accuracy of 98% by using a k-nearest neighbour algorithm ($k = 1$) and binary fingerprints, which indicate the presence or absence of specific sequence signatures, and focused on multi-label predictions. Further, the similarity of catalysed chemical transformations among protein families has been

*To whom correspondence should be addressed.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at Xiamen University on February 27, 2013

discussed (Almonacid and Babbitt, 2011; Babbitt and Gerlt, 1997). Some studies have proposed methods for automatically predicting EC numbers, by focusing on chemical transformations between substrates and products and ignoring protein sequences and structures.

The RPAIR database was constructed to map information on atom types in three regions of the molecular structures of substrate–product pairs (Kotera *et al.*, 2004a). Using the RPAIR database, E-zyme predicts possible EC number of a target reaction by comparing the correlation between vectors that represent the occurrence of atom mapping patterns (Yamanishi *et al.*, 2009). These researchers achieved a prediction accuracy of 82.5%. E-zyme is a rule-based method. In two of their studies, Gasteiger's group converted reactions only within classes EC 1 and EC 3 to physicochemical descriptors relating to the reacting bonds and classified them into EC sub-subclasses by using a self-organizing neural network, a SVM, and hierarchical clustering (Hu and Garcia, 2010; Sacher *et al.*, 2009). Their SVM gave an accuracy of 97.7%. Leber *et al.* (2009) proposed a method for the mapping of atom and bond types between substrates and products in a reaction using matrices, and they obtained unique matrices for each EC sub-subclass. The use of a reaction graph kernel (RGK) along with a random walk kernel was proposed for performing approximate matching between atom mapping patterns (Saigo *et al.*, 2010). The matching accuracy was 82.5%. Although this method also uses the RPAIR database to achieve high prediction performance, it is used to identify the main substrate–product pairs in a reaction. Latino and Aires-de-Sousa (2009) proposed physicochemical and topological descriptors, named the MOLMAP descriptor, encoding bonding changes during chemical reactions using matrices. Their random forest predictors, a semi-supervised learning algorithm, gave an accuracy of 82.4% for EC sub-subclasses for a genome-wide set of reactions. Their work showed that the predictions were reliable if a full balanced description of the reaction is used. Egelhofer *et al* (2010) used typical chemoinformatics approaches to EC sub-subclasses, that is, a string-type descriptor and a Tanimoto similarity metric. Nath and Mitchell (2012) evaluated a combination of five descriptors and three machine learning algorithms; they used a SVM, random forest and k-nearest neighbour algorithm. They achieved accuracies of 74.4% and 83.7% in the case of the SVM and random forest, respectively. However, their tanning set was limited to 260 reactions, and it was derived from the MACiE database (Holliday *et al.*, 2012).

In this study, we investigated the automatic selection of important chemical transformations from reactions for predicting the EC numbers of the reactions; in the selection process, we removed outliers and noise from chemical transformations without any preconditions. We propose a new method to flexibly predict EC numbers from the chemical structure of substrates and products of unclassified enzymatic reactions by the use of a maximal common substructure (MCS) algorithm, mutual information (MI) and SVM through comparison of the structural fingerprint to a list of classified reactions.

2 METHODS

To predict the EC number for an enzymatic reaction, the user enters the chemical structures of substrates and products in the MDL Molfile format, and Enzyme COmmission numbers Handler (ECOH) outputs the predicted EC sub-subclasses. The ECOH algorithm consists of the following three steps: (i) extraction of substructures from the substrates and products by using a MCS algorithm; (ii) calculation of MI for the extracted substructures by comparison with a standard set of EC-classified reactions, and subsequent generation of a candidate list of EC numbers for the target reaction; and (iii) prediction of the EC number using an SVM for the target reaction.

2.1 Extraction of substructures

We have focused on the difference in chemical structure between the substrate and product of an enzymatic reaction, and extracted their shared/unshared substructures by using the MCS algorithm (Tonnelier *et al.*, 1991). The chemical structures of the compounds were represented as graphs in which the nodes represented atoms and the edges represented bonds. The MCS algorithm was based on that implemented in the chemistry development kit (CDK) ver. 1.2.5 (Steinbeck *et al.*, 2003), an open source Java toolkit for manipulating small molecules using Java 6. A main issue with using MCS is that it is non-deterministic polynomial time (NP)-complete. The MCS algorithm based on CDK was therefore limited by both the number of searches and the computational time. If no MCS was found within these limits, then the program terminated the search and generated the MCS from the solution at that point.

In this study, we improved the original MCS algorithm implemented in CDK by cache management of the results and accelerating bit-string operations. Furthermore, the criteria confirming whether two chemical bonds were matched were modified by subjecting them to the following matching standards: (i) the electrical charge on the atom, (ii) the number of hydrogens, (iii) the bond order of both atoms, (iv) its cis- or trans-isomerism, and, for the whole molecule, (v) the number of atoms involved in bonds that are part of ring structures. This modified MCS algorithm improved the sensitivity of our method as it improved the likelihood of finding equivalent atoms or bonds between two chemical structures.

By applying the modified MCS algorithm, we extracted four types of substructure pairs from the target enzymatic reaction (Fig. 1a): a 'conserved substructure pair', which is a combination set of matching bonds extracted from the substrates and products of an enzymatic reaction by using the MCS algorithm (Fig. 1b); a 'changed substructure pair', which is a combination of the remaining bond sets and their adjacent atoms that were not in the conserved substructure pair (Fig. 1c); a 'neighbouring substructure pair', which is a combination of adjacent atom sets belonging to the conserved substructures and located adjacent to the changed substructures (Fig. 1d); and a 'small substructure pair', which is a combination of substructures having three or less atoms (except hydrogen) (Fig. 1e). As a result, water, oxygen, ammonia and metal ions were categorized as small substructures in the ECOH algorithm.

2.2 Calculation of mutual information and generation of the candidate EC numbers

Feature selection is the process of selecting a subset of the informative terms to solve a given inference problem. MI, which represents the measure of the statistical dependence between two variables, has previously been used as the feature selection strategy with promising results (Guyon and Elisseeff, 2003; Wang and Liu, 2011). We measured MI with a function $I(s, e)$, which was formally defined by Manning *et al.* (2008) according to the equation

$$I(s,e) = \sum_{r_s \in \{1,0\}} \sum_{r_e \in \{1,0\}} p(s = r_s, e = r_e) \log_2 \frac{p(s = r_s, e = r_e)}{p(s = r_s)p(e = r_e)} \quad (1)$$

where s is a random variable that assumes the value $r_s = 1$ when the reaction contains substructure pair s and assumes $r_s = 0$ when the reaction does not contain the substructure pair s. Furthermore, e is another

Downloaded from http://bioinformatics.oxfordjournals.org/ at Xiamen University on February 27, 2013

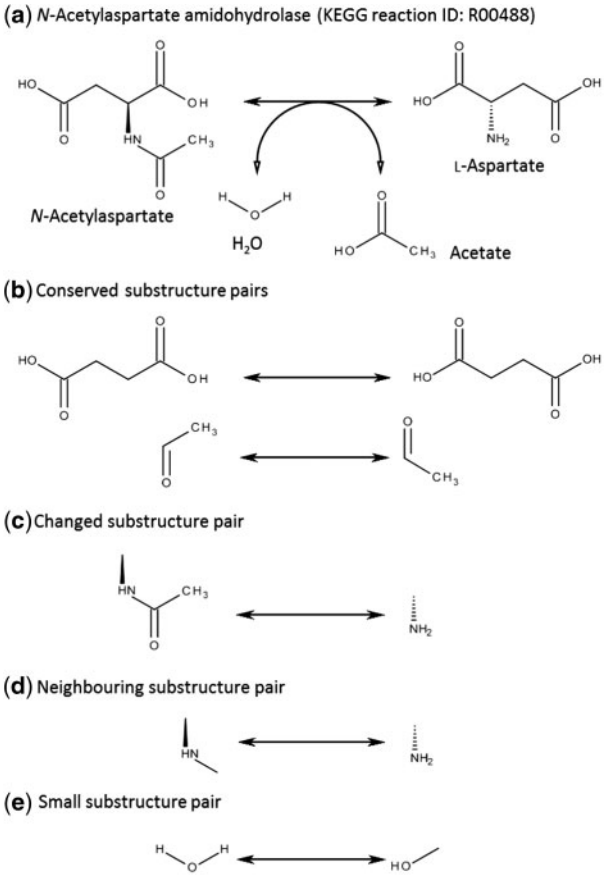


Fig. 1. Substructure pairs extracted by the MCS algorithm from the compounds involved in the reaction catalysed by *N*-acetylaspartate amidohydrolase. (**a**) The reaction catalysed by *N*-acetylaspartate amidohydrolase, (**b**) the extracted two conserved substructure pairs, (**c**) the extracted changed substructure pair, (**d**) the extracted neighbour substructure pair and (**e**) the extracted small substructure pair

random variable that takes the values $r_e = 1$ when the reaction is labelled with EC sub-subclass e, and $r_e = 0$ when the reaction is not labelled with the EC sub-subclass e. When a substructure pair s and an EC sub-subclass e are independent, $I(s, e)$ is zero, whereas MI becomes large when s is biased to one particular EC sub-subclass. For example,

$$p(s=1,e=0)\log_2\frac{p(s=1,e=0)}{p(s=1)p(e=0)} = \frac{N_{10}}{N}\log_2\frac{N\times N_{10}}{N_{1.}\times N_{.0}} \quad (2)$$

where Ns are the number of reactions, N_{10} is the number of reactions that contain substructure pair s and that are not labelled with the EC sub-subclass e, N_1. ($= N_{10} + N_{11}$) is the number of reactions that contain substructure pairs s ($r_s = 1$), $N_{.0}$ ($= N_{01} + N_{00}$) is the number of reactions that are not categorized into the EC sub-subclass e and N ($= N_{00} + N_{01} + N_{10} + N_{11}$) is the total number of reactions. $p(s = 1, e = 0)$ is balanced by N_{10}, $N_{1.}$, and $N_{.0}$. Consequently, on the basis of Eq. (1), it can be expected that the substructure pair with high MI in a particular EC sub-subclass e shares important structural information with that class.

For a set of substructure pairs generated from a target reaction r, the candidate score $W(S_r, e)$ for an EC sub-subclass e was heuristically defined as follows:

$$W(S_r,e) = \sum_{s\in S_r}\exp\left(I(s,e) - \frac{1}{|E|}\sum_{e_i\in E} I(s,e_i)\right) \quad (3)$$

where the sigmoid function is used as a weight function and $W = 0$ if $I(s, e) = 0$. All EC sub-subclasses e were sorted by their candidate score $W(S_r, e)$ to arrive at the candidate list for the target reaction.

When a substructure pair extracted from a target reaction did not correlate with any of the substructure pairs of the reactions included in the training set, the most similar substructure pair was selected. Although several researchers have performed similarity measurements between chemical structures based on the MCS approach, the algorithm used was considered too time-consuming for the purpose of this study (Cao *et al.*, 2008). The molecular fingerprinting method was therefore adapted to perform similarity measurements between substructure pairs. Molecular fingerprints are bit string representations of molecular structures in which each bit represents the presence or absence of a specific structural feature, and are commonly used for structure similarity searching (Tohsato *et al.*, 2000). In this study, we used the MACCS key, which is the most widely known molecular fingerprint (McGregor and Pallai, 1997).

Using the MACCS key fingerprint, an extracted substructure was converted into a 166-bit string. All extracted substructure pairs, corresponding to substrates and products from an irreversible enzymatic reaction, were further converted into a single 332-bit string by joining the two 166-bit strings in the order of substrate and product. Here, the 332-bit string for a conserved substructure is redundant when substrates and products contain the same substructure. By considering all reactions as reversible, 332-bit strings were generated in both directions. These joined 332-bit strings were pre-generated from the query reaction and all extracted substructure pairs of reactions in the database. For each 332-bit string from a query reaction, it was confirmed whether the same bit-string pattern existed in the training set. When the corresponding 332-bit string was absent in the training set, the most similar 332-bit string was selected based on the Tanimoto coefficient given by the following equation:

$$T(x,y) = N_z/(N_x + N_y - N_z) \quad (4)$$

where N_x and N_y are the number of binary values 1 of the 332-bit string x and y, respectively, and N_z is the number of binary value 1 shared by both. The number $T(x, y)$ is a measure for the degree of similarity between two 332-bit strings, where a value of 1 indicates full similarity.

2.3 Predictions of EC numbers with a support vector machine

SVM is a supervised machine learning technique that is widely used in pattern recognition and classification problems because of its high-performance prediction ability (Vapnik, 1998). SVM performs a classification by constructing a multidimensional hyperplane that optimally discriminates between two classes by maximizing the margin between two data clusters. Query sample orientation relative to this hyperplane gives the predicted class (Vapnik, 1998). As MI does not consider the dependency between EC sub-subclasses, we introduced an SVM to predict an EC sub-subclass for a target enzymatic reaction from the two EC sub-subclasses that ranked first and second in the candidate list. The SVM with the Gaussian kernel as a function of the SVM was implemented by using LibSVM version 3.0 (Chang and Lin, 2011). A grid search (Hsu *et al.*, 2003) found the best combination of the two tuning parameters, C (error penalty) and γ (kernel function), on Gaussian kernels of the SVM in the range of Log $C = [-5, \ldots, 15]$ and Log $\gamma = [-15, \ldots, -1]$ by calculating the sensitivity, precision and accuracy (details of performance measurements are provided in the next section). The resulting optimized parameters $C = 128$ and $\gamma = 0.0004$ were achieved at a sensitivity, precision and accuracy of 95.3%, 93.8% and 99.9% (see Supplementary File S1), respectively, by the jack-knife cross-validation procedure, where the reactions alternate with each other in being the query reaction and the EC sub-subclass is predicted on the basis of the remaining reactions in the dataset.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Xiamen University on February 27, 2013

Although LibSVM supports multi-class classification using the 'one-against-one' technique by combining all binary classifiers (Vapnik, 1998), we finally used an SVM as a binary classifier after considering the results of experiments. Thus, the top two predicted EC sub-subclasses in the candidate list of EC sub-subclasses covered 92.0% of the correct predictions (Supplementary File S2). In practice, the prediction performance for our data when the multi-classifier SVM from top n predicted EC sub-subclasses ($2 < n \leq 10$) is lower than that in the case of a binary SVM from top two predicted EC sub-subclasses (Supplementary File S3). As we will see later in the dataset section, the EC classification represents an imbalanced distribution, i.e. the number of reactions in one class is much smaller than that in another. For example, in the data, EC 1.1.1 contains 528 reactions, whereas EC 6.6.2 covers two reactions. This imbalance problem is of concern to many researchers who analyse datasets (Hsu *et al.*, 2003). Indeed, a highly imbalanced data distribution generally results in a poor classification performance for unseen samples belonging to the minority class in a conventional SVM classifier; this is because the classifier may be strongly biased towards the majority class (He and Garcia, 2009). Cost-sensitive sampling and kernel-based methods are all designed to address this problem of imbalanced classification (Japkowicz and Stephen, 2002). For multi-class SVMs, the distribution is even more imbalanced (You *et al.*, 2011). Therefore, we formally used a binary SVM to predict EC sub-subclasses for enzymatic reactions in the ECOH.

In the study, we consider all reactions as reversible, and used 1328-bit string representations of irreversible enzymatic reactions as the input and training data for the SVM. To generate the 1328-bit string, if more than one substructure pair are extracted from an irreversible reaction (e.g. 'conserved' in Fig. 1b), all the pairs are encoded in the same 332-bit string. The 332-bit strings representing the substructure pairs extracted from an irreversible reaction were further converted into a single 1328-bit string by joining the bit strings of each type of substructure pair in the order of 'conserved', 'changed', 'neighbouring' and 'small'. Using this 1328-bit string representation, the EC sub-subclass of the target reaction can be predicted from the similarity of its bit string to the substructure pairs extracted from the reactions in the training set, and each type of substructure pair can be simultaneously assessed. Because two types of 1328-bit strings are generated from a reaction by the previous procedures, two types of outputs are provided by the SVM for a reaction. When the outputs conflict with each other, the EC sub-subclass that is ranked first in the candidate list generated by the MI is used as the prediction result by the SVM for the targeted reaction.

3 RESULTS AND DISCUSSION

3.1 Dataset and performance measurement

The REACTION database of the KEGG database (version 58.1) (Kanehisa *et al.*, 2012) contains 7976 individual reactions. From the original dataset, reactions with no EC sub-subclass assigned (1255 entries) or reactions with an incomplete EC sub-subclass classification (487 entries) were removed. Of the remaining 6234 reactions, 5643 reactions covering the chemical structures of substrates and products in the MDL Molfile format were targeted. All reactions in the data were considered to be reversible reactions. We extracted 2744 conserved substructure pairs, 5860 changed substructure pairs, 3391 neighbouring substructure pairs and 388 small substructure pairs from the targeted reactions. Following the aforementioned procedure, 1328-bit string representations were obtained for the target reactions, which covered 162 EC sub-subclasses. The minimum number of reactions in 26 EC sub-subclasses is 1.

As EC sub-subclasses are more difficult to predict than EC subclasses and the last digit in the identifier is merely a serial number, an attempt was made to predict the EC sub-subclass for the query reaction (Latino and Aires-de-Sousa, 2009; Saigo *et al.*, 2010; Yamanishi *et al.*, 2009). All evaluations were conducted by a simple jack-knife (leave-one-out) cross-validation procedure. Namely, the reactions alternate with each other in being the query reaction, and the EC sub-subclass is predicted on the basis of the remaining reactions in the dataset. The *positive* (the query reaction is labelled with the predicted EC sub-subclass, and for reactions that have more than one EC sub-subclass, at least one of the sub-subclasses is predicted) or *negative* (the query reaction is not labelled) result of the prediction is recorded. This is repeated for all reactions. The performance of the ECOH method was measured by assessing its *precision* $\{[TP/(TP+FP)] \times 100\%\}$, *sensitivity* $\{[TP/(TP+FN)] \times 100\%\}$ and *accuracy* $\{[(TP+TN)/(TP+FP+TN+FN)] \times 100\%\}$; *TP*, *FP*, *TN* and *FN* represent the number of enzymatic reactions labelled as true positive, false positive, true negative and false negative, respectively.

3.2 Jack-knife prediction

The effectiveness of SVM in the ECOH approach was demonstrated by a comparison between the predicted EC sub-subclass based on MI (taking the highest candidate score in the generated candidate list), and the EC sub-subclass predicted by SVM from the top two EC sub-subclasses (MI + SVM). As shown in Table 1, the experimental results show that overall, the SVM based on the selection of target EC sub-subclass by MI has a superior performance (sensitivity = 86.1%, precision = 87.4% and accuracy = 99.8% by the jack-knife test; see the dataset section). Of the correct predictions, 180 predictive efforts matched well with the candidate list ordered by MI, rather than the SVM, which resulted in a run-off between the #1 and #2 candidates. In contrast, 298 predictions matched better with the list ordered by the SVM compared with the extent of its agreement with the list ordered by MI. The results also reveal that accuracy is not an appropriate measure to evaluate the detailed performance of the ECOH method because it spans only a narrow range from 99.7% to 99.9% (Table 1). This narrowness derives from the fact that all reactions in EC sub-subclasses, excluding the target reaction, tend to be categorized into TN. Therefore, in the following analyses, we mainly measured performance by sensitivity and precision, except when comparing the ECOH method

Table 1. Prediction performance (%) for the EC sub-subclasses

Method	Number of reactions (*n*)	Total (5643)	EC 1 (2015)	EC 2 (1816)	EC 3 (917)	EC 4 (535)	EC 5 (213)	EC 6 (202)
MI	Sensitivity	84.0	81.5	92.4	81.7	77.3	60.8	85.9
	Precision	85.3	81.4	90.0	84.1	87.7	87.2	80.1
	Accuracy	99.8	99.8	99.8	99.8	99.7	99.8	99.9
MI + SVM	Sensitivity	86.1	87.1	88.9	86.0	80.3	67.0	85.9
	Precision	87.4	87.2	92.0	84.7	79.7	89.3	81.6
	Accuracy	99.8	99.9	99.8	99.9	99.7	99.9	99.9

Total number of reactions categorized into each EC sub-subclass is greater than the number of reactions in total because a single reaction corresponds to multiple EC sub-subclasses.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Xiamen University on February 27, 2013

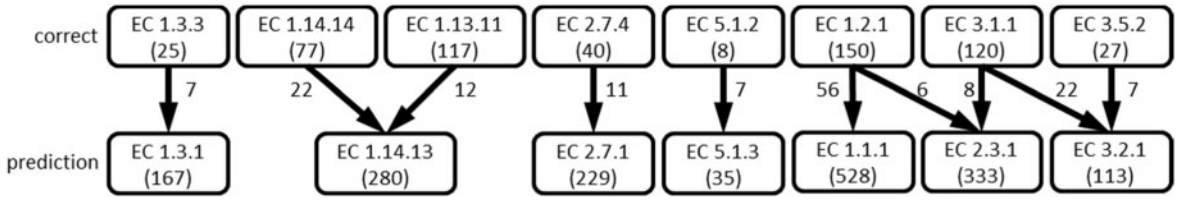


Fig. 2. The top 10 combinations of misclassified and correct EC sub-subclasses. The value besides each arrow indicates the total number of misclassified reactions. The number of reactions categorized into each EC sub-subclass is given in brackets

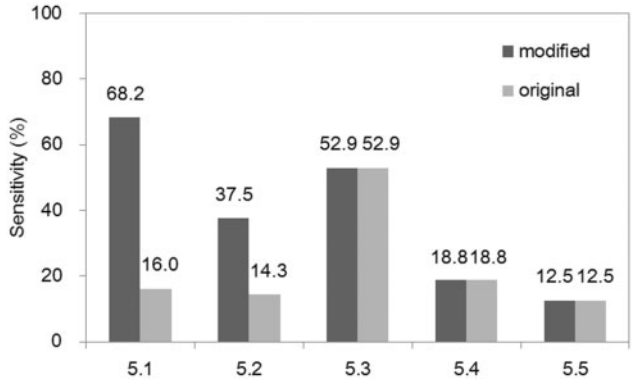


Fig. 3. Prediction performance of the algorithms for the EC 5 subclass of reactions

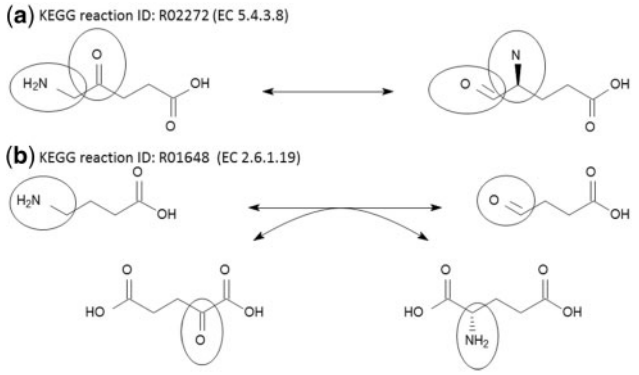


Fig. 4. Difference in the position within the molecule where atoms are transferred between (**a**) isomerases that transfer amino groups, as shown for the reaction catalysed by glutamate-1-semialdehyde aminotransferase, and (**b**) transaminases, as shown for the reaction catalysed by 4-aminobutyrate aminotransferase. The changed substructures are indicated with circles

with methods used by others (Latino and Aires-de-Sousa, 2009; Saigo *et al.*, 2010; Yamanishi *et al.*, 2009).

Considering the first digit of the target EC sub-subclasses in Table 1, transferases (EC 2) are predicted with the highest performance (sensitivity = 88.9% and precision = 92.0%). Here, the performance is high because of the sample size, i.e. the number of reactions categorized into EC 2. However, making predictions is not so straightforward for EC 1. These results indicate that the performance also depends on how easily the classification criteria for the EC sub-subclass to be targeted are detected by the ECOH algorithm. Figure 2 shows the 10 most frequently encountered combinations of misclassified EC sub-subclasses along with their correct classification. The 56 reactions with EC 1.2.1 (i.e. a reaction catalysed 3,4-dihydroxyphenylacetaldehyde) are wrongly predicted as EC 1.1.1 because the reaction with EC 1.1.1 includes enzymes 'acting on the CH-OH group of donors, with nicotinamide adenine dinucleotide (NAD+) or nicotinamide adenine dinucleotide phosphate (NADP+) as an acceptor'. However, it can also be assigned EC 1.2.1, which covers enzymes 'acting on the aldehyde or oxo-group of donors, with NAD+ or NADP+ as an acceptor' (Egelhofer *et al.*, 2010).

For reactions catalysed by isomerases (EC 5), it is difficult to predict the EC sub-subclasses (sensitivity = 67.0% and precision = 89.3%) because EC 5 contains indistinguishable classifications such as intramolecular oxidoreductases, transferases and lyases that usually involve only minor structural changes. We further divided the EC 5 classified reactions into their EC subclasses and calculated the average prediction performance for each EC subclass by using the modified MCS algorithm and the original MCS algorithm, in which our five matching conditions between two bonds are not added (Fig. 3). The prediction performances for reactions with EC 5.4 (intramolecular transferases) and EC 5.5 (intramolecular lyases) were relatively low with sensitivities of 18.8% and 12.5%, respectively. The enzymatic reaction corresponding to EC 5.4.3 (isomerases), for instance, was predicted to belong to EC 2.6.1 (transaminases) (Fig. 4), indicating that for further improvement of the performance, differences in the position where atoms are transferred within the target molecule need to be considered. In contrast to EC 5.4 and EC 5.5, the sensitivity performances for EC 5.1 (racemases and epimerases) and EC 5.2 (cis-trans isomerases) were 68.2% and 37.5%, respectively. We can see significant improvement in the performances compared with that achieved by the original MCS algorithm. These relatively good performances appeared to result from the identification of the connectivity between the molecular graphs of substrates and products, with only stereochemical changes, and the modification of the matching conditions between bonds in the MCS algorithm.

3.3 Performance of the SVM approach

On the basis of the performance measures of accuracy, the ECOH method shows the best performance compared with the methods proposed by previous related studies focusing on chemical transformations in reactions, in which the highest accuracy was 82.5% when all EC sub-subclasses were targeted (for each method, see Section 1). In the ECOH algorithm, when a

Downloaded from http://bioinformatics.oxfordjournals.org/ at Xiamen University on February 27, 2013

Table 2. Performance (%) for EC number prediction of target reactions with unregistered substructure pairs

Method	Number of reactions (*n*)	Total (2406)	EC 1 (550)	EC 2 (990)	EC 3 (440)	EC 4 (266)	EC 5 (72)	EC 6 (88)
MI + SVM	Sensitivity	80.3	75.6	88.2	81.5	67.3	42.3	85.4
	Precision	80.6	74.4	89.0	77.8	69.4	71.4	79.2

Table 3. Prediction performance (%) for compounds categorized as main reactant pairs for reactions taken from the RPAIR database

Method	Number of reactions (*n*)	Total (5520)	EC 1 (1965)	EC 2 (1776)	EC 3 (899)	EC 4 (525)	EC 5 (213)	EC 6 (197)
MI + SVM	Sensitivity	76.4	70.0	88.4	77.1	66.4	64.2	68.2
	Precision	77.6	69.3	89.1	77.8	73.7	78.6	66.5

substructure pair extracted from a test reaction did not correlate with any of the substructure pairs of reactions in the training set, the most similar substructure pair was selected. We evaluated the prediction performance for reactions with such unregistered substructure pairs and found a slight decrease in sensitivity and precision performance of ~5% (sensitivity = 80.3% and precision = 80.6%) (Table 2). In addition, we explored whether the ECOH algorithm could predict EC sub-subclasses solely on the main substrates and products in the target reaction (as described in the KEGG/RPAIR database). As shown in Table 3, the sensitivity and precision of the predictions decreased by ~10% (sensitivity = 76.4% and precision = 77.6%), indicating the greater effectiveness of the ECOH method compared with MOLMAP (Latino and Aires-de-Sousa, 2009) and RGK (Saigo *et al.*, 2010), both of which exhibit decreases of ~20%. Furthermore, in the RGK and E-zyme, the number of reactions registered in the RPAIR database is used. It is known that E-zyme is a rule-based method and can only predict similar reactions (Saigo *et al.*, 2010).

3.4 Prediction for multi-functional enzymatic reactions

As an additional application of candidate lists of EC sub-subclasses, we collected 85 reactions that were assigned to multiple EC sub-subclasses. Using these reactions as a query, we evaluated whether the correct EC sub-subclasses for a target reaction would both rank high in the EC sub-subclasses candidate list (Fig. 5). We observed that for 53 reactions (62.3%) both correct EC sub-subclasses were present at the top of the candidate lists. For instance, the reaction catalysed by 3-hydroxyisobutyryl-CoA hydrolase is classified into the EC 3.1.2 and EC 6.2.1, which ranked first and second, respectively, in the predicted candidate list of EC sub-subclasses (Fig. 6a). At least one correct EC sub-subclass was correctly predicted for 25 reactions (29.4%) classified in multiple EC sub-subclasses. The reaction displayed in Figure 6b was classified into this 'at least one' group because the correct EC classification EC 3.1.3 and EC 4.6.1 appeared as the first and ninth entry in the candidate EC sub-subclasses list, respectively. However, this is a similar reaction, as the reaction classified as EC 3.1.4 (KEGG reaction ID: R03435), which serves as an example that the inconsistent assignment of EC sub-subclasses worsens the prediction performance of the ECOH algorithm. In total, at least one correct EC sub-subclass ranked high in the candidate EC sub-subclasses list in 91.7% of the multi-functional enzymatic reactions.

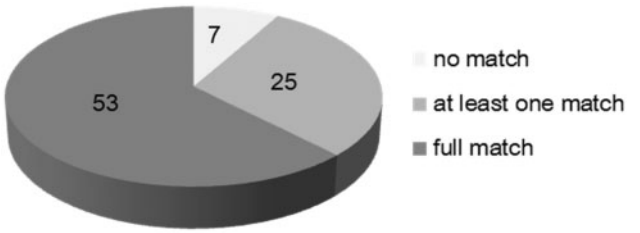


Fig. 5. Prediction result for target reactions that are classified in multiple EC sub-subclasses

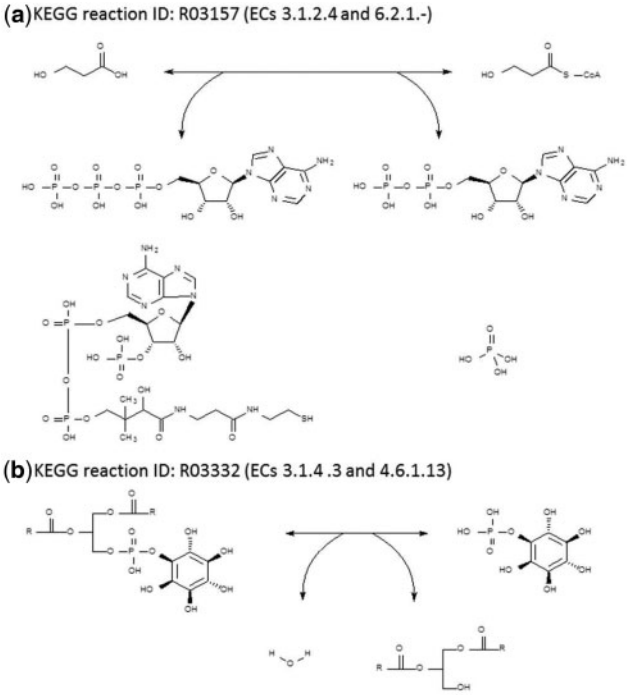


Fig. 6. Examples of enzymatic reactions that are classified into multiple EC sub-subclasses: (**a**) the reaction catalysed by 3-hydroxyisobutyryl-CoA hydrolase, and (**b**) the reaction catalysed by 1-phosphatidyl-D-myo-inositol inositolphosphohydrolase

3.5 Computational time for the MCS extraction

As previously mentioned, we improved the original MCS algorithm implemented in CDK. We compared the processing time of both the original MCS algorithm and the improved one by measuring the average processing time (in seconds) for an MCS with an increasing size in 20-bond intervals (Fig. 7). All time

Downloaded from http://bioinformatics.oxfordjournals.org/ at Xiamen University on February 27, 2013

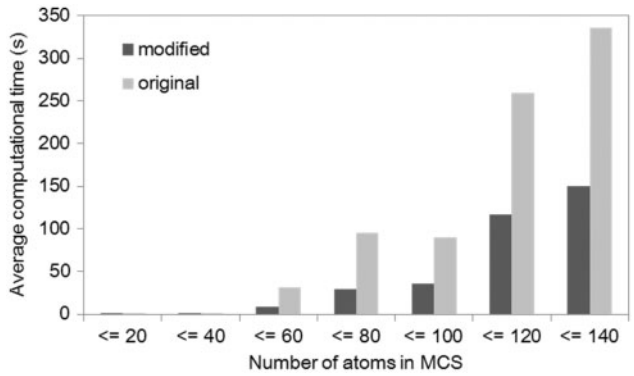


Fig. 7. Correlation of MCS size with computational time

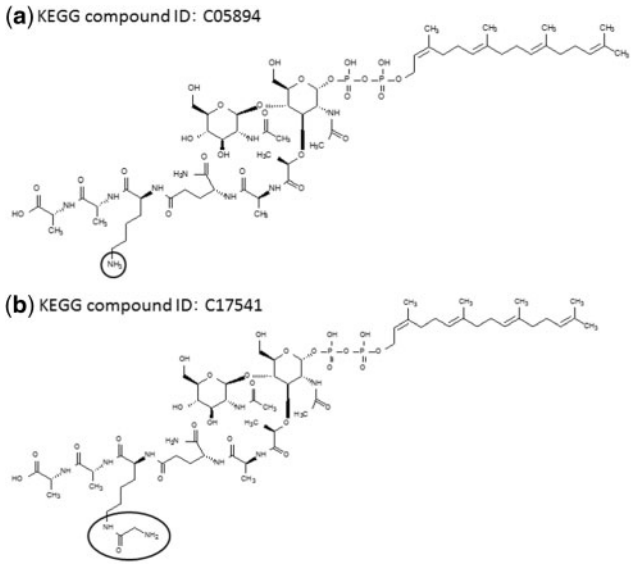


Fig. 8. Example of a large MCS. The changed substructures between (**a**) C05894 and (**b**) C17541 are indicated with circles. The remaining substructures are common MCSs between them

measurements were obtained on a Windows 7 system with 4.0 GB RAM and an Intel Core 2 Duo i7-860 processor running at 2.80 GHz. The modified algorithm was faster than the original, and it extracted relatively small MCSs of $\leq$20 atoms within 0.02 s. On setting an upper limit for the number of searches (5 million) and computational time (10 min), only 4.2% did not fall within the set search limit (320 out of 7678); this number was 0.05% (4 out of 7678) for the set time limit. The correct MCS was still predominantly generated in instances where a large MCS was extracted (Fig. 8).

4 CONCLUSIONS

We proposed a method to predict EC sub-subclasses of enzymatic reactions using a MCS algorithm based on the comparison of substructure pairs, the output of which was a measure of the change in the chemical structure between the substrate and the product. The correlation between EC sub-subclasses and substructure pairs was defined by MI values. We adopted SVM to estimate the MI of an individual substructure pair more rigorously within the same period.

The proposed method achieved high performance and was flexible in the sense that by using the MACCS fingerprint and the Tanimoto coefficient for unclassified substructure pairs in a target reaction, the most similar substructure pair was obtained and substituted. In addition, because of the high percentage of high-ranking corrected EC sub-subclasses in the candidate list, this study also revealed the possibility of using the generated candidate list of EC sub-subclasses for the further study of multi-functional enzymatic reactions. It should be noted that the sensitivity and precision for the EC 5 class of isomerases were lower than those for the other classes. Improving the representation of substructure patterns and the introduction of methods proposed to address the imbalance problem for the input dataset are both considered effective future approaches for the improvement of the prediction performance of EC numbers.

Funding: This work was supported by a Grant-in-Aid for Young Scientists (B) (No. 23700353) and a grant of Strategic Research Foundation Grant-aided Project for Private Universities (No. S1001042) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Conflict of Interest: none declared.

REFERENCES

Almonacid,D.E. and Babbitt,P.C. (2011) Toward mechanistic classification of enzyme functions. *Curr. Opin. Chem. Biol.*, **15**, 435–442.

Andreeva,A. *et al.* (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Babbitt,P.C. and Gerlt,J.A. (1997) Understanding enzyme superfamilies. Chemistry as the fundamental determinant in the evolution of new catalytic activities. *J. Biol. Chem.*, **272**, 30591–30594.

Bray,T. *et al.* (2009) Sequence and structural features of enzymes and their active sites by EC Class. *J. Mol. Biol.*, **386**, 1423–1436.

Cao,Y. *et al.* (2008) A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics*, **24**, i366–i374.

Chang,C.C. and Lin,C.J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27.

Cuff,A.L. *et al.* (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res.*, **39**, D420–D426.

De Ferrari,L. *et al.* (2012) EnzML: Multi-label prediction of enzyme classes using InterPro signatures. *BMC Bioinformatics*, **13**, 61.

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

Egelhofer,V. *et al.* (2010) Automatic assignment of EC numbers. *PLoS Comput. Biol.*, **6**, e1000661.

Furnham,N. *et al.* (2009) Missing in action: Enzyme functional annotations in biological databases. *Nature Chemical Biology*, **5**, 521–525.

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157–1182.

He,H.B. and Garcia,E.A. (2009) Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.*, **21**, 1263–1284.

Holliday,G.L. *et al.* (2012) MACiE: Exploring the diversity of biochemical reactions. *Nucleic Acids Res.*, **40**, D783–D789.

Hsu,C.W. *et al.* (2003) A practical guide to support vector classification. *Bioinformatics*, **1**, 1–16.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Xiamen University on February 27, 2013

Hu,X. and Garcia,E.A. (2010) Similarity perception of reactions catalyzed by oxidoreductases and hydrolases using different classification methods. *J. Chem. Inf. Model.*, **50**, 1089–1100.

IUPAC-IUBMB. (1999) IUPAC-IUBMB Joint Commission on Biochemical Nomenclature (JCBN) and Nomenclature Committee of IUBMB (NC-IUBMB), Newsletter 1999. *Eur. J Biochem.*, **264**, 607–609.

Japkowicz,N. and Stephen,S. (2002) The class imbalance problem: a systematic study. *Intel. Data Anal.*, **6**, 429–449.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Kotera,M. *et al.* (2004a) RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions. *Genome Informatics*, **15**, P062.

Kotera,M. *et al.* (2004b) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.*, **126**, 16487–16498.

Latino,D.A. and Aires-de-Sousa,J. (2009) Assignment of EC numbers to enzymatic reactions with MOLMAP reaction descriptors and random forests. *J. Chem. Inf. Model.*, **49**, 1839–1846.

Leber,M. *et al.* (2009) Automatic assignment of reaction operators to enzymatic reactions. *Bioinformatics*, **25**, 3135–3142.

Lu,L. *et al.* (2007) ECS: an automatic enzyme classifier based on functional domain composition. *Comput. Biol. Chem.*, **31**, 226–232.

Manning,C.D. *et al.* (2008) *Introduction to Information Retrieval*. Cambridge University Press, Cambridge.

McGregor,M.J. and Pallai,P.V. (1997) Clustering of large databases of compounds: using MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.*, **37**, 443–448.

Nath,N. and Mitchell,J.B. (2012) Is EC class predictable from reaction mechanism? *BMC Bioinformatics*, **13**, 60.

Sacher,O. *et al.* (2009) Investigations of enzyme-catalyzed reactions based on physicochemical descriptors applied to hydrolases. *J. Chem. Inf. Model.*, **49**, 1525–1534.

Saigo,H. *et al.* (2010) Reaction graph kernels predict EC numbers of unknown enzymatic reactions in plant secondary metabolism. *BMC Bioinformatics*, **11**, S31.

Steinbeck,C. *et al.* (2003) The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.*, **43**, 493–500.

Tohsato,Y. *et al.* (2000) A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2000**, 376–383.

Tonnelier,C. *et al.* (1991) Machine learning of generic reactions: 3. An efficient algorithm for maximal common substructure determination. *Tetrahedron Comput. Methodol.*, **3**, 351–358.

Vapnik,V.N. (1998) *Statistical Learning Theory*. Wiley, New York.

Wang,Y. and Liu,X. (2011) Prediction of silicon content in hot metal based on SVM and mutual information for feature selection. *J. Inf. Comput. Sci.*, **8**, 4275–4283.

Yamanishi,Y. *et al.* (2009) E-zyme: Predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. *Bioinformatics*, **25**, i179–186.

You,M.Z. *et al.* (2011) MAPLSC: A novel multi-class classifier for medical diagnosis. *Int. J. Data Min. Bioinf.*, **5**, 383–401.

Yu,C. *et al.* (2009) Genome-wide enzyme annotation with precision control: Catalytic families (CatFam) databases. *Proteins*, **74**, 449–460.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Xiamen University on February 27, 2013