BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 11 2011, pages 1537–1545
doi:10.1093/bioinformatics/btr177

Systems biology Advance Access publication April 8, 2011

Prediction of metabolic reactions based on atomic and molecular properties of small-molecule compounds

Fangping Mu[1,*], Clifford J. Unkefer[2], Pat J. Unkefer[2] and William S. Hlavacek[1]

[1]Theoretical Biology and Biophysics Group, Theoretical Division and [2]National Stable Isotope Resource, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Our knowledge of the metabolites in cells and their reactions is far from complete as revealed by metabolomic measurements that detect many more small molecules than are documented in metabolic databases. Here, we develop an approach for predicting the reactivity of small-molecule metabolites in enzyme-catalyzed reactions that combines expert knowledge, computational chemistry and machine learning.

Results: We classified 4843 reactions documented in the KEGG database, from all six Enzyme Commission classes (EC 1–6), into 80 reaction classes, each of which is marked by a characteristic functional group transformation. Reaction centers and surrounding local structures in substrates and products of these reactions were represented using SMARTS. We found that each of the SMARTS-defined chemical substructures is widely distributed among metabolites, but only a fraction of the functional groups in these substructures are reactive. Using atomic properties of atoms in a putative reaction center and molecular properties as features, we trained support vector machine (SVM) classifiers to discriminate between functional groups that are reactive and non-reactive. Classifier accuracy was assessed by cross-validation analysis. A typical sensitivity [TP/(TP + FN)] or specificity [TN/(TN + FP)] is ≈ 0.8. Our results suggest that metabolic reactivity of small-molecule compounds can be predicted with reasonable accuracy based on the presence of a potentially reactive functional group and the chemical features of its local environment.

Availability: The classifiers presented here can be used to predict reactions via a web site (http://cellsignaling.lanl.gov/Reactivity/). The web site is freely available.

Contact: fmu@lanl.gov

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

Received on June 28, 2010; revised on February 23, 2011; accepted on March 25, 2011

1 INTRODUCTION

Metabolomics, the study of a cell's complement of small-molecule metabolites, has recently been recognized as an important part of post-genomics science (Fiehn, 2002). Identification of metabolites, including their chemical structures, and reconstruction of the network of enzyme-catalyzed reactions in which metabolites participate are two of the important challenges for metabolomic analysis of a cell (Breitling *et al.*, 2006; Kell, 2004; Schreiber, 2005).

Large-scale metabolic network reconstructions currently available are based on literature data and genome annotation, which identifies genes encoding enzymes. Reconstructions intended to be suitable for flux analysis have been obtained in this way for a variety of organisms (Forster *et al.*, 2003; Reed *et al.*, 2003; Romero *et al.*, 2004). Organism-specific metabolic databases are also available [for an example, see Mueller *et al.* (2003)].

The sequence/knowledge-based approach to network reconstruction has a number of limitations. This approach can only account for reactions that have been characterized experimentally in some system, and large gaps in our knowledge of metabolism seem likely. More than 35% of known metabolic activities that are classified by Enzyme Commission (EC) numbers are not associated with known amino acid sequences (Chen and Vitkup, 2007). In addition, even for *Escherichia coli*, the quintessential model organism, novel metabolic reactions continue to be discovered (Fischer and Sauer, 2003; Loh *et al.*, 2006; Nakahigashi *et al.*, 2009). Xenobiotic metabolism, which ranges from biodegradation pathways in microorganisms (Ellis *et al.*, 2006) to drug metabolism in mammals (Boyer *et al.*, 2007; Rendic, 2002; Vaz *et al.*, 2010), is rarely considered in network reconstructions. An aspect that is also largely neglected is the broad substrate specificity of many enzymes, which has been exploited in synthesis of organic compounds (Muller, 2004) and in building libraries of drug candidates (Breinbauer *et al.*, 2002). Broad substrate specificity has been stated to be the major reason why many more metabolites will be present in a metabolome than can be deduced from genome sequence (Fischbach and Clardy, 2007; Nobeli *et al.*, 2009; Schwab, 2003).

To improve network reconstruction, we must generate and use new types of data, such as whole-cell metabolite profiles, which can be produced using techniques such as high-resolution mass spectrometry (MS) (Dunn *et al.*, 2005). These techniques have been used to detect many novel metabolites, both endogenous (Bhalla *et al.*, 2005; Saito *et al.*, 2010; van der Werf *et al.*, 2005) and xenobiotic (Anari and Baillie, 2005), but signatures of novel metabolites typically cannot be easily associated with chemical structures. As our ability to analyze complex mixtures obtained from cells grows, so does appreciation of our ignorance of the metabolic networks that produce the compounds in these mixtures (Harrigan and Goodacre, 2003). Although current MS-based techniques enable researchers to detect a large fraction of the metabolome with relatively low per-experiment cost, methods

*To whom correspondence should be addressed.

and software tools for analysis of metabolomic data have lagged behind the development of analytical chemistry tools, and the full potential of metabolomic data analysis has yet to be realized (Baran *et al.*, 2009; Fiehn, 2007; Kind and Fiehn, 2008; Moco *et al.*, 2007; Wishart, 2007)

Metabolites are linked through metabolic pathways, and novel metabolites are produced and consumed within pathways that are linked to known metabolites and pathways. Thus, as suggested by Anari and Baillie (2005), one possible strategy for metabolite structure elucidation and metabolic pathway reconstruction would be to predict reactions and compare properties of predicted reactants against experimental observations. This approach has been used in studies of drug metabolism (Anari and Baillie, 2005).

Current methods of reaction prediction are based on expert-defined rules that generalize the transformations of known reactions (Langowski and Long, 2002). Several software tools are available for the prediction of metabolic reactions, such as MetabolExpert (Darvas, 1998), METEOR (Greene *et al.*, 1999), META (Klopman *et al.*, 1994), the UM-BBD Pathway Prediction System (Hou *et al.*, 2003) and BNICE (Hatzimanikatis *et al.*, 2005). Predictions of these systems are characterized by high false positive rates (Boobis *et al.*, 2002). To mitigate this problem, some systems report priority indices based on expert knowledge to aid a user in assessing the likelihood of a predicted reaction (Payne, 2004). BNICE enables a user to consider the thermodynamic properties of predicted reactions (Hatzimanikatis *et al.*, 2005).

Recently, we investigated how machine learning can be used to extend expert systems for reaction prediction, and we developed a method for predicting potential substrates and products of oxidoreductase-catalyzed reactions (Mu *et al.*, 2006). In this work, we classified 1626 oxidoreductase reactions in the KEGG database (Goto *et al.*, 2002) into 12 classes of functional group transformations. Using seven empirical atomic properties, which are readily calculated from a compound's chemical structure, we trained classifiers to discriminate between positive and negative examples of substrates and products. Classifiers were found to be reasonably accurate on the basis of cross-validation (CV) analysis.

Here, we report the extension of this method to all six EC classes of enzyme-catalyzed reactions, which is important for providing a complete reaction prediction system. Enzymes catalyze a limited number of biotransformations. We have defined 80 reaction classes, each of which is characterized by a functional group biotransformation (e.g. primary alcohol dehydrogenation). These reaction classes encompass the vast majority of reactions documented in KEGG. The reaction centers and surrounding local substructures of substrates and products are represented as molecular substructure patterns using SMARTS (http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html), which is a language for describing molecular substructures. The existence of a particular substructure is a prerequisite for reactivity; however, it is not sufficient. To model the reactivity of functional groups, we have built binary classifiers, support vector machines (SVMs), using positive and negative examples taken from the KEGG database for each SMARTS-defined substructure. The input to classifiers is a set of atomic and molecular properties, features that characterize potential reaction centers. CV shows average classification accuracy to be about 80%. A web site has been established for reaction prediction based on the classifiers presented here (http://cellsignaling.lanl.gov/Reactivity/). The web site is freely available.

2 METHODS

Data: we initially considered 6577 reactions documented in the KEGG database (Goto *et al.*, 2002) on June 8, 2006. (The work reported here began in 2006.) We also downloaded the whole COMPOUND section of the KEGG database, which included 12 027 chemicals. Of 6577 reactions, 5140 were associated with a complete EC number, 476 were associated with an incomplete EC number and 961 were unclassified (i.e. not associated with any EC number). The 5616 reactions assigned a full or partial EC number consist of 2009 oxidoreductase reactions, 1791 transferase reactions, 953 hydrolase reactions, 503 lyase reactions, 227 isomerase reactions and 200 ligase reactions. A few reactions are assigned multiple EC numbers. Out of the 12 027 KEGG compounds, 10 942 were linked to 2D chemical structures in MDL format. The 6577 reactions found in KEGG involve a total of 4792 distinct metabolites as substrates, products or cofactors: 4083 of the 4792 metabolites have 2D structures. More than half of the KEGG compounds are orphaned, i.e. not documented to participate in reactions. On November 23, 2008, we downloaded 755 newly documented reactions in the KEGG database (i.e. 755 reactions not available in 2006). These 755 reactions involve 1721 unique metabolites, of which 1619 are associated with 2D structures in KEGG. We used the 2006 data to define reaction classes, to train classifiers and to test classifiers through CV. We used the 2008 data only to test classifiers.

Reaction classification and rules: we manually classified 4843 of the 6577 reactions that we considered into 80 reaction classes. These reaction classes are fully defined in Supplementary Material S1. Each reaction class corresponds to a set of reactions involving a common functional group transformation and a common set of reaction centers in substrates and products. A reaction center is the substructure of a reactant directly affected by reaction. Among reactants in the 4843 reactions, we identified 170 reaction center patterns: 82 of these patterns correspond to substructures in substrates and 88 of these patterns correspond to substructures in products. Here, 'substrate' refers to a reactant on the left-hand side of a reaction, and similarly, 'product' refers to a reactant on the right-hand of a reaction. Each reaction in KEGG is associated with a direction, which is arbitrary in many cases. Directions are also associated with our reaction classes.

Figure 1 illustrates the basic concepts that guided our definitions of reaction classes and reaction center patterns. (Our approach for defining reaction center patterns, which are essentially functional groups, is explained in detail later.) Consider EC 1.1.-.- reactions, which involve dehydrogenation or oxidation of a CH-OH group and various cofactors. The CH-OH group can be a primary alcohol, secondary alcohol or acetal group. Thus, we defined three reaction classes for EC 1.1.-.- reactions, one for each possible type of CH-OH group. The reaction class for secondary alcohol dehydrogenation/oxidation is illustrated in Figure 1A. As can be seen, substrates in this reaction class are matched by a reaction center pattern that contains a secondary alcohol, and products are matched by a reaction center pattern that contains a ketone group. A second reaction class is illustrated in Figure 1B.

As suggested by Figure 1, our classification scheme was guided by the EC numbering system, which is used to classify enzymes by the reactions they catalyze. However, there is not a one-to-one relationship between our reaction classes and EC numbers. For reactions with a common EC number, we sometimes divided the reactions into several reaction classes if the functional group transformations among these reactions were found to be different (e.g. methyltransfer reactions, EC 2.1.1.- reactions, are divided into several classes based on whether the acceptor functional group is derived through *N*-methyltransfer, *O*-methyltransfer, *S*-methyltransfer or *C*-methyltransfer). A reaction class may also include reactions with diverse EC numbers (e.g. our alkyl hydroxylation reaction class includes EC 1.14.-.- and EC 1.17.-.- reactions). Our classification scheme differs from

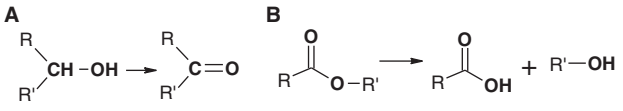

Fig. 1. Examples of reaction classes and reaction center patterns (in bold). (**A**) Dehydrogenation of a secondary alcohol, which is an EC 1 reaction (listed as Reaction class 2 in Supplementary Material S1). Reactions in this class have the form A → B; the majority of reaction classes have this form. The reaction center in the substrate (i.e. the reactant on the left-hand side of the reaction), a secondary alcohol, is a hydroxyl group, and the reaction center in the product (i.e. the reactant on the right-hand side of the reaction) is a ketone group (as shown) or acetal group. Reaction centers in substrates are matched by the following SMARTS pattern: [CX4H;!$(C([OX2H])[O,S,#15])][OX2H]. Reaction centers in products are matched by either [CX3](=[OX1])([#6,#7])[#6] or [CX4;R](C)(C)([Oh])OC. We define a total of 80 reaction classes and 170 reaction center patterns, 82 for 'substrate' reaction centers (e.g. CH-OH) and 88 for 'product' reaction centers (e.g. C=O). (**B**) Linear carboxylic ester hydrolyzation, which is an EC 3 reaction (listed as Reaction class 48 in Supplementary Material S1). Reactions in this class have the form A → B + C. A total of nine reaction classes have this form (Reaction classes 15, 33, 48, 52, 58, 64, 65, 68 and 69). Two reaction classes have the form A + B → C (Reaction classes 77 and 78).

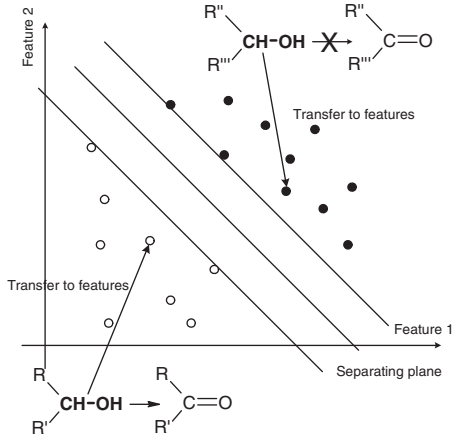

Fig. 2. Illustration of the classifier corresponding to the reaction center pattern that identifies potentially reactive hydroxyl groups in substrates of reactions of Reaction class 2 (Fig. 1A). Features (atomic and molecular properties) are calculated for compounds containing a substructure matched by the reaction center pattern, and the matches are labeled as either negative examples (filled circles) or positive examples (open circles) as explained in the Section 2. The classifier is simply a surface in feature space (labeled 'separating plane' in this figure) that divides negative and positive examples. This surface is found using standard SVM methods (Vapnik, 1998; Chang and Lin, 2001) as explained in the Section 2. There is a classifier for each of the 170 reaction center patterns.

that of the EC numbering system in two ways. First, our scheme does not capture all known enzyme-catalyzed reactions. Second, our scheme is free of idiosyncrasies, in that reactions within a class all involve a common transformation and a common set of reaction centers in reactants. Each reaction class can be viewed as corresponding to a rule that defines reactions, including a specification of the necessary properties of substrates and products in these reactions.

Reaction center patterns and classifiers for reaction prediction: for each reaction center in substrates and products of the 80 reaction classes, we used SMARTS, a notational system for representing molecular substructures, to define the local structures of reaction centers as previously described (Mu *et al.*, 2006). Examples of SMARTS specifications of reaction center patterns are given in the caption of Figure 1. We defined a total of 170 reaction center patterns (82 for substrates and 88 for products), which are provided in a spreadsheet (Supplementary Material S2). A SMARTS pattern may include not only a specification of a reaction center but also a specification of a local structure that must occur or is necessarily absent based on our best understanding of the relevant biochemistry (Silverman, 2000). Each SMARTS pattern is associated with a rule derived from 1 of the 80 reaction classes. There are 170 reaction center patterns and 80 rules all together (Supplementary Material S1).

For each SMARTS-defined reaction center pattern, we developed a binary classifier via the SVM approach (Vapnik, 1998). The classifier can be used for reaction prediction. The basic concept is illustrated in Figure 2. A reaction center pattern and its associated classifier serve as a model for a reactive functional group. The classifier is trained to distinguish between functional groups that are reactive and functional groups that are non-reactive (the details of classifier training are given below). The rule associated with a pattern together with the structure of a compound found to contain a reactive functional group matched by the pattern serve to define a particular reaction. Patterns identify compounds containing potentially reactive functional groups, classifiers identify functional groups that are likely to be reactive and rules identify the reactions of compounds with reactive functional groups.

Labeling of training data: to identify negative and positive examples for classifier training, we used JOELib (http://sourceforge.net/projects/joelib/) (JOELib, 2004) to match each of the 170 SMARTS patterns against the structures of 10 942 metabolites, the set of metabolites with 2D chemical structures in the KEGG database in 2006. Molecular symmetry was considered to avoid multiple matches of the same reaction center (Mu *et al.*, 2006; Steinbeck *et al.*, 2003). For each SMARTS pattern, we divided the set of matching substructures in metabolites (i.e. the matches) into negative and positive examples. A match to a reaction center pattern was considered to be a positive example if the match was recorded in KEGG to be a reaction center in the type of reaction associated with the reaction center pattern. Otherwise, the match was considered to be a negative example. Thus, for lack of a better approach, we assume that absence of evidence is evidence of absence. It should be noted that more than half of the 10 942 metabolites are orphaned, meaning that these metabolites are not documented to participate in reactions. Orphan metabolites contribute to negative training data but not positive training data. There are 170 sets of negative and positive examples, one for each reaction center pattern/classifier.

Features—atomic and molecular properties: in general, an SVM classifier is trained to distinguish negative and positive examples on the basis of features. The features considered here are atomic properties of the atoms in a potential reaction center (i.e. the atoms in a substructure matched by a reaction center pattern) and molecular properties (i.e. properties that depend on the overall structure of a molecule). We consider 54 atomic properties, which are listed and defined in Supplementary Material S3, and 81 molecular properties, which are listed and defined in Supplementary Material S4. Thus, the total number of features associated with each example is 54 times the number of atoms in a potential reaction center (1, 2 or 3) plus 81. Below, we provide a brief overview of the atomic and molecular properties. We also describe how the properties are determined.

The 54 atomic properties can be divided into empirical properties (Properties 1–26) and theoretical/semiempirical properties (Properties 27–54). The properties can also be divided into the following six categories: (i) electrostatic properties (e.g. charge distribution), which include Properties 5–10 and 27–49; (ii) inductive properties (e.g. charge transmission), which include Properties 12–16; (iii) energetic properties, which include Properties 50–54; (iv) topological properties (e.g. graph potential), which include Properties 2–4; (v) steric properties (e.g. surface area exposed to solvent), which include Properties 11, 17, 18 and 20; and (vi) distance properties (e.g. the Euclidean distance between an atom in a molecule and the center of mass

of the molecule), which include Properties 1, 19 and 21–26. In earlier work (Mu *et al.*, 2006), we considered only seven atomic properties. Here, we consider more atomic properties largely because we are considering more reaction classes (80 versus 12). The properties influencing reactivity seem to depend on the class of reaction under consideration (see Section 3).

The 81 molecular properties characterize the shape, surface, energy and charge distribution of a molecule. We introduce molecular properties to define the limits of applicability of our classifiers, as we expect naturally occurring metabolites, which comprise our training data, to have molecular properties distinct from those of many xenobiotic compounds.

The atomic and molecular properties are derived from 2D and 3D chemical structures as described in detail in Supplementary Materials S3 and S4. Starting from the 2D structures obtained from the COMPOUND section of the KEGG database, we used the Marvin Beans software package (http://www.chemaxon.com/marvin/help/applications/molconvert.html) (MolConverter, 2009) to add explicit hydrogen atoms for valence balance and to generate 3D atomic coordinates. The generated 3D structures were then optimized using MOPAC2007 and the PM3 parameter set (Stewart, 2007). The optimized 3D structures were then used to determine atomic and molecular properties using JOELib (JOELib, 2004), CDK (Steinbeck *et al.*, 2003), MOPAC2007 (Stewart, 2007) and simple in-house codes. In-house codes were used to calculate several properties that we introduce in this study (Properties 19–26) and to process MOPAC2007 output files. Detailed definitions of the newly introduced properties are available in Supplementary Material S3.

Classifier training: each training example is associated with a vector of features. We scaled the numerical values of each feature to lie in the range $[-1, +1]$.

Given l training examples $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_l\}$, where $\boldsymbol{x}_i$ is the vector of features associated with example i, and l labels $\boldsymbol{y} = \{y_1, y_2, \ldots, y_l\}$, which identify each example as either a negative or positive example, we used LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm) (Chang and Lin, 2001) to train a soft margin SVM classifier by solving the following dual optimization problem:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T \boldsymbol{H}\boldsymbol{\alpha} - \boldsymbol{\alpha}^T 1 \tag{1}$$

subject to

$$\begin{aligned} &0 \le \alpha_i \le C_+, \quad \text{if } y_i = 1, \\ &0 \le \alpha_i \le C_-, \quad \text{if } y_i = -1, \\ &\boldsymbol{y}^T \boldsymbol{\alpha} = 0 \end{aligned} \tag{2}$$

where $\boldsymbol{\alpha}$ is a vector of Lagrange multipliers, C_+ and C_- are penalty parameters introduced to balance the numbers of positive and negative examples, and $\boldsymbol{H}$ is an $l \times l$ matrix. The elements of $\boldsymbol{H}$ are given by

$$H_{ij} = y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{3}$$

where $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a kernel. We initially considered a linear kernel, as in our earlier work (Mu *et al.*, 2006), but we obtained better results with a radial basis function (RBF) kernel, a type of kernel commonly used. This kernel is given by

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \, \| \boldsymbol{x}_i - \boldsymbol{x}_j \|^2)$$

where $i, j = 1, \ldots, l$. The values of γ, C_+ and C_- are selected as described below and training yields values for the Lagrange multipliers $\boldsymbol{\alpha}$ (Chang and Lin, 2001). The decision function for a new feature vector is the sign of the classifier's raw score, which is given by

$$f(x) = \sum_{i=1}^{l} \boldsymbol{\alpha}_i y_i K(x, x_i) + b \tag{4}$$

where b is a parameter that can be obtained given the solution of the optimization problem described above (Chang and Lin, 2001).

Sensitivity and specificity of each classifier are calculated as part of a 3-fold CV procedure, in which the whole training dataset is divided randomly into three equal sets and each of the three sets is used for testing one by one while the other two sets are used for training. Sensitivity Q_p is the fraction of positive examples (TP) that are predicted to be positives [TP/(TP + FN)], and specificity Q_n is the fraction of negative examples (TN) that are predicted to be negatives [TN/(TN + FP)]. The quantities FP and FN are the numbers of false positives and negatives. For each SVM classifier, three parameters, γ, C_+ and C_-, need to be determined. We used a grid-search of γ, C_+ and C_- values to optimize

$$Q = \sqrt{Q_p \times Q_n} \tag{5}$$

This procedure is intended to balance sensitivity and specificity. A good classifier does not sacrifice sensitivity to obtain specificity or vice versa, and it is desirable for Q_p and Q_n to each be close to 1. Two grid-searches were performed for each SVM classifier. First, we performed a coarse grid search over the following grid points: $C_+ = [2^{-5}, 2^{-3}, \ldots, 2^{15}]$, $C_- = [2^{-5}, 2^{-3}, \ldots, 2^{15}]$ and $\gamma = [2^{-15}, 2^{-13}, \ldots, 2^{3}]$. We then performed fine grid searches around local optima with exponentially decaying steps: 2, $2^{0.5}$, $2^{0.125}$, etc.

Feature importance ranking: to rank the relative importance of features for a given classifier (Guyon *et al.*, 2002), we calculate a ranking coefficient, which is given by the change in the value of the objective function [Equation (1)] when a feature is essentially removed from $\boldsymbol{H}$ [Equation (3)]. Thus, for a feature with index r in the vector of features, we calculate,

$$\mathrm{DJ}(-r) = \frac{1}{2}\boldsymbol{\alpha}^T \boldsymbol{H}\alpha - \frac{1}{2}\boldsymbol{\alpha}^T \boldsymbol{H}(-r)\boldsymbol{\alpha} \tag{6}$$

where $\boldsymbol{H}(-r)$ is $\boldsymbol{H}$ recalculated such that the feature with index r in each vector of features $\boldsymbol{x}_i$ $(i = 1, \ldots, l)$ is set equal to zero.

For a given set of feature indices R, we define the contribution of the corresponding features to the output of a classifier as,

$$\mathrm{CT}(-R) = 1 - \frac{\boldsymbol{\alpha}^T \boldsymbol{H}(-R)\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \boldsymbol{H}\alpha} \tag{7}$$

where $\boldsymbol{H}(-R)$ is $\boldsymbol{H}$ recalculated such that the features with indices in R in each vector of features $\boldsymbol{x}_i$ $(i = 1, \ldots, l)$ are set equal to zero.

Feature selection: we modified the classifier training protocol described above to include a simple recursive feature selection algorithm suitable for SVM classifiers (Guyon and Elisseeff, 2003). In the algorithm, a classifier is trained using a given set of features as usual. We start with the complete set of atomic and molecular features. Next, the features are ranked as described above, the least important feature is removed to obtain a smaller set of features, and the classifier is retrained using the new, smaller feature set. The procedure is repeated.

3 RESULTS

3.1 Classification of enzyme-catalyzed reactions

As detailed in the Section 2, we manually classified reactions obtained from the KEGG database (Goto *et al.*, 2002) in 2006 into 80 reaction classes based on the biotransformations occurring in reactions (Fig. 3; Supplementary Material S1). Of 5616 reactions obtained from KEGG with a full or partial EC number, 4251 (75.7%) involve a transformation captured in one of the 80 reaction classes defined here (Fig. 3B). Of the 5616 reactions, 826 (14.7%) could not be included in our analysis because structural information was missing in KEGG for substrates and/or products, and 539 (9.6%) involve exotic transformations unaccounted for in our 80 reaction classes (Fig. 3B). With more data, it may be possible to recognize regularities among these exotic reactions, but for now, we do not consider them further. Of 961 reactions obtained from KEGG without an assigned EC number, 592 (61.6%) have well-defined functional group transformations and these reactions are included in the 80 reaction classes.

For each reaction center in substrates and products, we used SMARTS to define the local structures of reaction centers as

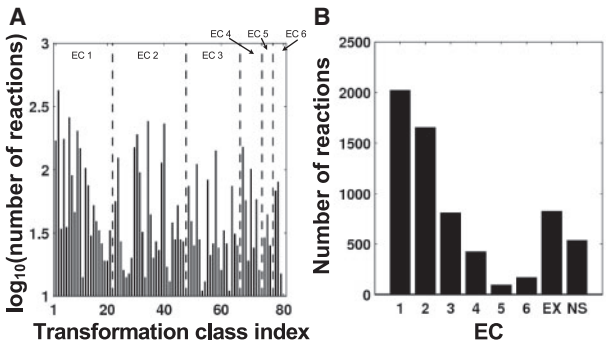

Fig. 3. Summary of reaction classification. (**A**) Number of reactions included in each of the 80 reaction classes. Reaction classes 1–21 are typically subclasses of oxidoreductase-catalyzed reactions (EC 1). We defined more reaction classes than in our earlier study of oxidoreductase-catalyzed reactions (Mu *et al.*, 2006). For example, alcohol dehydrogenation reactions considered in our earlier work were divided into dehydrogenation reactions of primary and secondary alcohols. Classes 22–47 are typically subclasses of transferase-catalyzed reactions (EC 2). Classes 48–66 are typically subclasses of hydrolase-catalyzed reactions (EC 3). Classes 67–73 are typically subclasses of lyase-catalyzed reactions (EC 4). These subclasses were defined based on the type of chemical bonds cleaved and the type of new bonds formed. Classes 74–76 are typically subclasses of isomerase-catalyzed reactions (EC 5). Classes 77–80 are typically subclasses of ligase-catalyzed reactions (EC 6). These subclasses were defined based on the type of chemical bond formed. (**B**) Number of reactions included among the 80 reaction classes for each major EC class. EX is the number of reactions documented in KEGG but not included in our analysis because they involve exotic transformations. NS is the number of reactions not included because structural information is missing in KEGG for substrates and/or products.

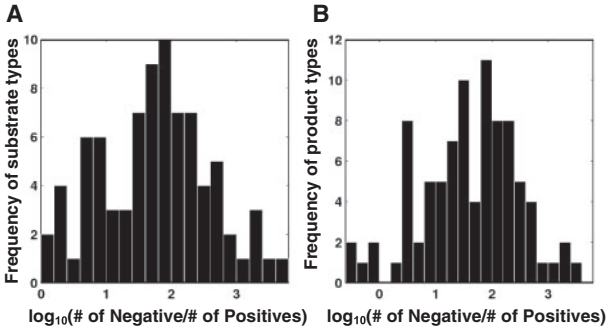

Fig. 4. Summary of training data for (**A**) the 82 'substrate' classifiers corresponding to reaction center patterns that match substructures in substrates of reaction rules and (**B**) the 88 'product' classifiers corresponding to reaction center patterns that match substructures of products in reaction rules. The number of negative examples is usually much larger than the number of positive examples.

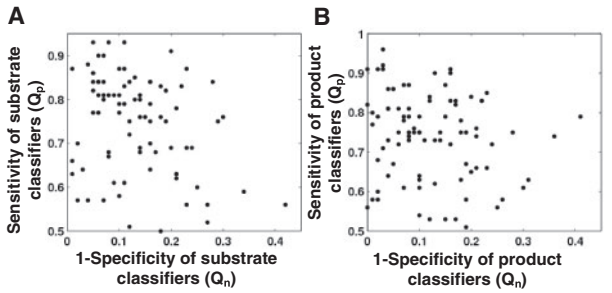

Fig. 5. Sensitivity (Q_p) and specificity (Q_n) of (**A**) the 82 'substrate' classifiers and (**B**) the 88 'product' classifiers. The average sensitivity is 0.74, with a SD of 0.11. The average specificity is 0.87, with a SD of 0.08.

previously described (Mu *et al.*, 2006), and we defined a total of 170 reaction center patterns (Supplementary Materials S1 and S2). Each reaction center is associated with a class of reaction. The reaction center patterns can be used to determine if a compound is a potential reactant in an enzyme-catalyzed reaction. The existence of a reaction center in a given metabolite is a prerequisite for a reaction.

To identify negative and positive examples of reactive functional groups, we matched each SMARTS pattern against 10 942 metabolites, which is the set of metabolites with 2D chemical structures documented in the KEGG database in 2006. A match to a reaction center pattern was considered to be a positive example if the match was recorded in KEGG to be a reaction center in the class of reaction associated with the reaction center pattern. Otherwise, the match was considered to be a negative example. Figure 4 shows the numbers of positive and negative examples identified in this manner for the 170 reaction center patterns, 82 of which correspond to substructures of substrates in our reaction classes and 88 of which correspond to substructures of products in our reaction classes. Each reaction class has a direction, with substrates on the left side and products on the right side. For each of the reaction classes, we found that the reaction centers of substrates and products are widely distributed among known metabolites (Fig. 4). The number of negative examples is typically much larger than the number of positive examples, although there are exceptions (e.g. thiol oxidation, RSH + RSH → RS-SR). This finding demonstrates that generalized transformation rules derived from reaction classification schemes can only be applied selectively to a specific set of metabolites. Metabolites are not generated in a cell through processes akin to random combinatorial synthesis.

3.2 Validation of classifiers

Using atomic and molecular properties as features, as discussed in the Section 2, we built classifiers, SVMs (Vapnik, 1998), to distinguish positive examples from negative examples for 170 sets of examples. The examples in each set are metabolites that all contain a substructure matched by one of the 170 reaction center patterns. We used a RBF kernel. The kernel parameters and other hyperparameters of the SVMs were determined through grid searching, as discussed in the Section 2. The hyperparameters found for each of the 170 classifiers are available in Supplementary Materials S1. Details about classifiers and classifier training are available in Section 2.

We measure the performance of each SVM using two statistics: sensitivity Q_p, which is the fraction of positive examples that are predicted to be positives, and specificity Q_n, which is the fraction of negative examples that are predicted to be negatives. We calculated Q_p and Q_n for each classifier in two ways: (i) in CV using data obtained from KEGG in 2006 and (ii) in blind testing using data obtained from KEGG in 2008. Figure 5 shows average values of sensitivity and specificity for each of the 170 classifiers based on 10 replications of 3-fold CV. The sensitivities and specificities shown in Figure 5 and SDs are given in Supplementary Material S1.

The average sensitivity for all 170 classifiers is 0.74 with a SD of 0.11, and the average specificity is 0.87 with a SD of 0.08. These results suggest that the local environment of atoms in a potential reaction center plays an important role in reactivity, as the features that provide the basis for classification characterize the local environment of a potential reaction center.

To further test the performance of classifiers, we used information about reactions now documented in KEGG but not used at all in training. These reactions were added to KEGG after initial versions of our classifiers were built. A total of 755 new reactions were obtained from KEGG in 2008 for use in blind testing. The 755 reactions involve 1721 metabolites. We found 2D structural information for 1619 of these metabolites in the KEGG database. After removing reactions that are not associated with reactant structure information and that involve ambiguous or exotic transformations (e.g. KEGG reaction R07744), we classified the remaining 624 reactions into our 80 reaction classes. These reactions involve reactants that match 71 of the 82 reaction center patterns for substrates and reactants that match 75 of the 88 reaction center patterns for products. For each reaction center pattern, we then identified positive and negative examples of reactive functional groups among the 1619 metabolites. Results from blind testing are similar to those obtained via CV. For example, the average sensitivity for the 71 substrate classifiers is 0.81 with a SD of 0.22. For the 75 product classifiers, it is 0.82 with an SD of 0.21. These results reinforce those of our CV analysis and further indicate that our classifiers generalize well and can be used to predict enzyme-catalyzed reactions with reasonable accuracy. These results also suggest that our classifiers, which were trained using data obtained from KEGG in 2006, will not quickly become obsolete. Detailed results from blind testing are available in Supplementary Materials S1.

In the above validation analysis, we examined the reactivity of a functional group in different molecular environments. We asked if our classifiers could identify the environments in which a functional group is reactive. We now examine different functional groups in the same molecule and ask if our classifiers can identify the reactive ones.

We can apply our classifiers to predict reactions for a metabolite as follows (Fig. 6). For each candidate reaction center in the metabolite, the raw output score of the corresponding classifier is determined. (The classifier decision function for the candidate reaction center is the sign of the raw output score, and the raw output score measures the location of the candidate reaction center in feature space relative to the separation line of the SVM classifier.) For a given compound, the 80 reaction classes discussed above serve as rules that determine the possible reactions of the compound. Each classifier that corresponds to a reaction center pattern that matches a substructure in the compound provides a raw output score for a possible reaction. The possible reactions identified in this way are ranked using the raw output scores of classifiers. Top ranked reactions (the most likely) have greater raw output scores than the bottom ranked reactions (the least likely). For the 1619 metabolites involved in the reactions considered in blind testing, 650 (731) metabolites are involved in 755 reactions as substrates (products). Based on the reaction rules that follow from our 80 reaction classes, the average number of possible reactions that each metabolite can be involved in as a substrate or product is 149.7 and 123.2, respectively. Among these possible reactions, we identified

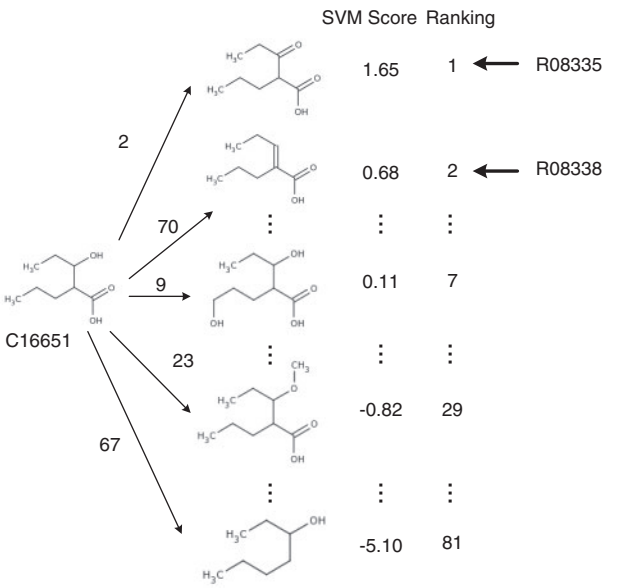

Fig. 6. Illustration of how classifiers can be used to rank the reactivity of functional groups within a compound. We note that the structure of compound C16651 (KEGG ID) was not used in classifier training, i.e. we downloaded information about this compound in 2008. Using our 82 'substrate' reaction center patterns, we identified 81 potential reaction centers in C16651. These reaction centers correspond to 81 possible reactions that consume C16651, 5 of which are illustrated in the figure. These reactions are instances of Reaction classes 2, 70, 9, 23 and 67 (Supplementary Material S1). For each of the 81 reactions, we calculated a raw SVM score, as described in the Section 2. Functional groups that receive a positive score are classified as reactive. Only 7 of the 81 functional groups are classified as reactive. The raw SVM scores are used to rank the 81 possible reactions, from most likely to least likely. We associate the most likely reaction with the greatest score (1.65) and the least likely reaction with the least score (−5.10). Among the 81 possible reactions, there are two reactions documented in KEGG. These reactions have KEGG IDs R08335 and R08338 and they are ranked 1 and 2. The enrichment factors for these *bonda fide* reactions are 81/1 (for R08335) and 81/2 (for R08338). If desired, the 88 'product' reaction center patterns and corresponding classifiers can be used in a similar manner to evaluate possible reactions that produce, rather than consume, compound C16651.

the rank of *bona fide* reactions, those reported in KEGG. The median ranks of the *bona fide* reactions involving substrates or products are 10 and 4, respectively, and the average rank is 23.24 and 11.14, respectively. For each *bona fide* reaction, we defined an enrichment factor, the total number of possible reactions involving a substrate or product divided by the rank of the *bona fide* reaction among all possible reactions. Multiple reactions for the same metabolite are computed separately. The average enrichment factors for reactions involving substrates or products are 25.44 and 50.37, respectively. These results indicate that our classifiers can reasonably identify the truly reactive functional groups among the potential reactive functional groups in a given compound.

3.3 Insights into reactivity

To gain insights into the chemical properties of metabolites that affect their reactivity, we used the metric of Equation (6) to determine the importance of each atomic and molecular feature for

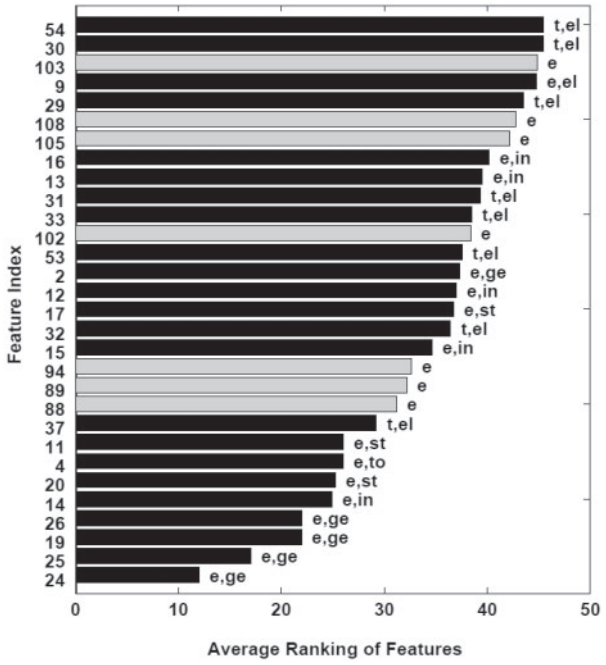

Fig. 7. Average feature importance rank across 170 classifiers. Feature indices, which are defined in Supplementary Material S5, are given along the *y*-axis. Average rank, which is between 1 (best possible) and 135 (worst possible), is given along the *x*-axis. The most important feature on average is the feature with index 24. Only the 30 highest ranked features are included in this figure. Black bars correspond to atomic properties; gray bars correspond to molecular properties. All bars are labeled to identify empirical (e) and theoretical/semiempirical (t) properties. Black bars are labeled to identify the class of atomic property (see Section 2): electrostatic (el), inductive (in), topological (to), steric (st) or distance (ge). No energetic properties appear in the top 30.

each classifier. For each classifier, relative importance was estimated for each of the 135 features (54 atomic properties and 81 molecular properties) as follows. For a given classifier, we translated the importance assigned by Equation (6) to a feature ranking in the range of 1–135, with rank 1 corresponding to the most important feature and rank 135 corresponding to the least important feature. For each feature, an average ranking was calculated based on ranking across all 170 classifiers. The average importance ranks are given in Supplementary Material S5. The importance rank for each of the 135 properties for each of the 170 classifiers is given in Supplementary Material S6. We also assessed the relative importance of atomic versus molecular properties using the metric of Equation (7). The results are given in Supplementary Material S7.

Figure 7 shows the top 30 features based on average ranking. The four most important features are distance properties, properties that reflect the geometrical location of a potential reaction center in a molecule. It seems that functional groups far away from the molecular center, but near the surface of a molecule, tend to be more reactive than those that are near the center of the molecule. Among the top 30 properties, the majority of features are electrostatic properties, consistent with our intuition about biochemistry. Also, most of the properties are atomic rather than molecular properties. Atomic properties generally have more importance than molecular properties. As assessed by the metric of Equation (7), the average importance contribution of molecular properties is 0.39 (out of 1.0) with a SD of 0.15 (Supplementary Material S7). We note that three of the eight properties defined in this study are among the top 30 features.

Supplementary Material S8 is a heat map showing the 135×170 ranks calculated using Equation (6)—this heat map visualizes the information in Supplementary Material S6. As can be seen from the heat map, there is no subset of features that is important across all classifiers. In other words, there are no atomic or molecular properties that govern reactivity across all 80 classes of reactions. One consequence of this finding is that calculation of a large number of properties is a necessary feature of our prediction system, although we are not claiming that the set of properties considered here is optimal or minimal.

We investigated the effect of modifying classifier training to include a simple feature selection algorithm, which is described in the Section 2. In short, no dramatic improvement in classifier performance was obtained by introducing feature selection. The impact of feature selection on specificity, sensitivity and Q [Equation (5)] is illustrated in Supplementary Material S9 for six classifiers. The results shown are typical. In some cases, a slight performance gain can be achieved.

4 DISCUSSION

It seems that numerous metabolites unknown to science exist in Nature, perhaps especially in the plant kingdom (Fiehn, 2002). The KEGG database includes information about $<16\,000$ compounds, and <5000 of these compounds are documented to participate in a reaction of some kind. Considering also man-made chemicals, the number of unknown enzyme-catalyzed reactions is surely large. How can we fill gaps in our knowledge of metabolites and metabolic reactions? Metabolites are linked through metabolic pathways, and novel metabolites are produced and consumed by pathways connected in some way to known metabolites and pathways. Also, it seems likely that novel pathways will use the known repertoire of biotransformations which enzymes have evolved to catalyze. Thus, reaction prediction methods, which rely on rules that generalize known reactions, may be helpful in the discovery of novel enzyme-catalyzed reactions and in assignment of structures to novel metabolites (Anari and Baillie, 2005; Soh and Hatzimanikatis, 2010). Such methods may have other applications as well (Soh and Hatzimanikatis, 2010).

To improve on currently available methods for reaction prediction, we have developed a prediction system, covering all six EC classes of reactions, which is based on a unique combination of expert knowledge, computational chemistry and machine learning. We classified metabolic reactions based on functional group chemistry and identified 80 reaction classes, which serve as rules for metabolic transformations. We then built SVM classifiers, which serve as models for reactive substructures in small-molecule compounds. Classifiers take as input a set of atomic and molecular properties, which can be derived from a chemical structure using computational chemistry tools. Our rules and classifiers are not comprehensive but they account for the majority of reactions documented in KEGG. Lack of training data is the main reason for not including more rules and classifiers at this time.

CV analysis indicates that our reaction prediction system is accurate (Fig. 5), in that a typical classifier is both sensitive and specific. The average sensitivity is 0.74, and the average specificity is 0.87. (A classifier with sensitivity and specificity of 1 is perfect.) The results of CV analysis were confirmed in blind testing, i.e. tests with data not considered in classifier training. We also found, via the approach illustrated in Fig. 6, that *bona fide* reactions are highly enriched among the set of reactions predicted for a given small-molecule compound. To achieve greater accuracy, our prediction approach could potentially be combined with aspects of other prediction systems designed to limit the combinatorial explosion that occurs when rules only are applied to predict reactions. Thermodynamic analysis (Hatzimanikatis *et al.*, 2005) and reasoning rules (Fenner *et al.*, 2008) seem promising in this respect.

We emphasize that our prediction system does not make predictions about the activities of particular enzymes, which can be highly specific (Boernke *et al.*, 1995), or the metabolic capabilities of particular organisms. Rather, it identifies functional groups that have properties similar to those of known reactive functional groups. Alternatively, our prediction system can be viewed as identifying functional groups that are *not* likely to be reactive in any biochemical system given the knowledge used to build the system, i.e. the information in the KEGG database.

One application of our system could be discovery of novel metabolic reactions through analysis of metabolomic data. In metabolite profiling experiments, especially those involving ultra high-resolution MS techniques, many more metabolites are detected in cell lysates than are documented to exist in pathway databases, such as KEGG (Fiehn and Weckwerth, 2003). The challenge in reconstructing a metabolic network from this type of data is to assign chemical structures to the signatures of novel metabolites and to elucidate substrate–product relationships. We believe reconstruction can be aided by reaction prediction in the following way. Consider a large set of MS peaks, most of which cannot be associated easily with chemical structures, and consider one peak to which a chemical structure can be assigned. For this structure, apply reaction prediction, which will identify a set of compounds that are potentially connected to the known compound via enzyme-catalyzed reactions, i.e. a set of possible precursors and products. Look for peaks corresponding to these predicted precursors and products. If such a peak is found, the structure obtained via reaction prediction can then be tentatively assigned to the peak. This process can be iterated as many times as desired. We have not attempted to determine if this idea can actually contribute to network reconstruction based on metabolite profiling, but it seems worth mentioning as a possible application. Note that this suggested application of our reaction prediction system would only be useful if there are reactions taking place in cells that are not currently documented in available databases.

It may be helpful to briefly discuss some of the machine learning aspects of the work reported here. The present work is an extension of our earlier work (Mu *et al.*, 2006), in which we developed 12 classifiers (for predicting EC 1 reactions) that take as input a set of only seven atomic properties, which are empirical. These classifiers are linear SVMs. Here, we report non-linear SVMs with RBF kernels, which we found to perform better in terms of sensitivity and specificity than linear SVMs (data not shown). However, the performance improvement is slight. Thus, use of the RBF kernel is not an essential ingredient of our prediction system.

In contrast, use of more than seven atomic properties is an essential ingredient. The main reason is the greater number of reaction classes considered in the present work (80 versus 12). This conclusion is based on our study of feature importance. Although we can point to some general aspects of feature importance, which are evident in Figure 7, there is no subset of features that governs reactivity for all reaction classes (Supplementary Materials S6 and S8).

We performed an exploratory investigation of the impact of feature selection (Guyon and Elisseeff, 2003) on classifier performance. Representative results are shown in Supplementary Material S9. We found that the performance of some classifiers can be improved slightly by incorporating a simple feature selection algorithm into the classifier training protocol. However, feature selection did not yield dramatic improvements.

Our present consideration of features extends our earlier consideration of features in three ways: (i) more empirical properties, (ii) theoretical/semiempirical properties for the first time and (iii) molecular properties for the first time. We found that molecular properties are less important than atomic properties (Supplementary Material S7). However, in our analysis, we only considered endogenous metabolites. We feel that molecular properties would be more important if we also considered xenobiotic compounds, as xenobiotic compounds do not necessarily resemble naturally occurring metabolites.

By considering more features than in earlier work (Mu *et al.*, 2006), we increased the breadth of our reaction prediction system (from EC 1 reactions to all six EC reaction classes). Our consideration of additional features also yielded improved classifier performance. For example, for dehydrogenation of CH-CH to C=C (Reaction class 6), the sensitivity and specificity scores of the 'substrate' classifier based on a linear kernel and seven atomic properties as features are 0.72 and 0.69, respectively (Mu *et al.*, 2006). In contrast, the sensitivity and specificity scores for the updated classifier based on an RBF kernel and 54 atomic and 81 molecular properties as features are 0.81 and 0.86, respectively (Supplementary Material S1).

As a final comment on machine learning aspects of our work, we wish to point out that the approach presented here and in our earlier work (Mu *et al.*, 2006) is a general approach that can be applied to develop application-specific reaction prediction systems. For example, this approach combined with training data taken from, say, a database for biodegradation pathways (Ellis *et al.*, 2006) could be used to develop a system for predicting reactions involved in the biodegradation of man-made chemicals.

Our reaction prediction system is available via a web-based application. This application takes as input a chemical structure and uses a variety of software tools identified in Section 2 to calculate the various atomic and molecular properties that must be determined to make predictions about the reactivity of the compound. For applications, such as metabolomic data analysis, additional software development will be needed. The software that we have developed so far was built only with the intention of evaluating the concept of combining expert knowledge, computational chemistry and machine learning to improve the accuracy of reaction prediction.

ACKNOWLEDGEMENTS

We thank Marian Anghel, Ingo Steinwart and Robert F. Williams for helpful technical discussions and the three anonymous reviewers for their detailed and constructive critiques. We also thank James P. Freyer for encouragement and support.

Funding: National Institutes of Health (grants GM080216, ES016920 and CA132629); DOE contract (DE-AC52-06NA25396).

Conflict of Interest: none declared.

REFERENCES

Anari,M.R. and Baillie,T.A. (2005) Bridging chemoinformatic metabolite predic-tion and tandem mass spectrometry. *Drug Discov. Today*, **10**, 711–717.

Baran,R. *et al.* (2009) Mass spectrometry based metabo-lomics and enzymatic assays for functional genomics. *Curr. Opin. Microbiol.*, **12**, 547–552.

Bhalla,R. *et al*. (2005) Metabolomics and its role in understanding cellular repsonse in plants. *Plant Cell Rep.*, **24**, 562–571.

Boernke,W.E. *et al*. (1995) Stringency of substrate specificity of Escherichia coli malate dehydrogenase. *Arch. Biochem. Biophys.*, **322**, 43–52.

Boobis,A. *et al*. (2002) In silico prediction of ADME and pharmacokinetics report of an expert meeting organised by COST B15. *Eur. J. Pharm. Sci.*, **17**, 183–193.

Boyer,S. *et al*. (2007) Reaction site mapping of xenobiotic biotransformations. *J. Chem. Inform. Model.*, **47**, 583–590.

Breinbauer,R. *et al*. (2002) From protein domains to drug candidates – natural products as guiding principles in the design and synthesis of compound libraries. *Angew. Chem. Int. Ed.*, **41**, 2878–2890.

Breitling,R. *et al*. (2006) Precision mapping of the metabolome. *Trends Biotechnol.*, **24**, 543–548.

Chang,C.-C. and Lin,C.-J. (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen,L. and Vitkup,D. (2007) Distribution of orphan metabolic activities. *Trends Biotechnol.*, **25**, 343–348.

Darvas,F. (1998) Predicting metabolic pathways by logic programming. *J. Mol. Graph.*, **6**, 80–86.

Dunn,W.B. *et al.* (2005) Measuring the metabolome: current analytical technologies. *Analyst*, **130**, 606–625.

Ellis,L.B.M. *et al*. (2006) The University of Minnesota Bioca-talysis/Biodegradation Database: the first decade. *Nucleic Acids Res.*, **34**, D517–D521.

Fenner,K. *et al*. (2008) Data-driven extraction of relative reasoning rules to limit combinatorial explosion in biodegradation pathway prediction. *Bioinformatics*e, **24**, 2079–2085.

Fiehn,O. (2002) Metabolomics - the link between genotypes and phenotypes. *Plant Mol. Biol.*, **48**, 155–171.

Fiehn,O. (2007) Cellular Metabolomics: the qwest for pathway structure. In Lindon,J.C., Nicholson,J.K. and Holmes,E. (eds) *The Handbook of Metabonomics and Metabolomics*. Elsevier, Amsterdam, Oxford, pp. 35–54.

Fiehn,O. and Weckwerth,W. (2003) Deciphering metabolic networks. *Eur. J. Biochem.*, **270**, 579–588.

Fischbach,M.A. and Clardy,J. (2007) One pathway, many products. *Nat. Chem. Biol.*, **3**, 353–355.

Fischer,E. and Sauer,U. (2003) A novel metabolic cycle catalyzes glucose oxidation and anaplerosis in hungry Escherichia coli. *J. Biol. Chem.*, **278**, 46446–46451.

Forster,J. *et al*. (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res*., **13**, 244–253.

Goto,S. *et al*. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res*., **30**, 402–404.

Greene,N. *et al*. (1999) Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR, and METEOR. *SAR QSAR Environ. Res*., **10**, 299–313.

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

Guyon,I. *et al.* (2002) Gene selection for cancer classification using support vector machine. *Mach. Learn.*, **46**, 389–422.

Harrigan,G.G. and Goodacre,R. (eds) (2003) *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*. Kluwer Academic Publishers, Boston.

Hatzimanikatis,V. *et al*. (2005) Exploring the diversity of complex metabolic networks. *Bioinformatics*, **21**, 1603–1609.

Hou,B.K. *et al.* (2003) Microbial pathway prediction: a functional group approach. *J. Chem. Inf. Comput. Sci.*, **43**, 1051–1057.

JOELib (2004) August 27, 2004 Version. Availabe at http://sourceforge.net/projects/joelib/ (last accessed date August, 2008).

Kell,D.B. (2004) Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.*, **7**, 296–307.

Kind,T. and Fiehn,O. (2008) Hardware and software challenges for the near future: structure elucidation concepts via hyphenated chromatographic techniques. *LCGC North America*, **26**, 176–187.

Klopman,G. *et al.* (1994) META. 1. A program for the evaluation of metabolic transformation of chemicals. *J. Chem. Inf. Comput. Sci.*, **34**, 1320–1325.

Langowski,J. and Long,A. (2002) Computer systems for the prediction of xenobiotic metabolism. *Adv. Drug Deliv. Rev.*, **54**, 407–415.

Loh,K.D. *et al.* (2006) A previously undescribed pathway for pyrimidine catabolism. *Proc. Natl Acad. Sci.*, **103**, 5114–5119.

Moco,S. *et al.* (2007) Metabolomics technologies and metabolite identification. *Trends Anal. Chem.*, **26**, 855–866.

MolConverter (2009) *Marvin beans 5.3.3*. ChemAxon Ltd, Budapest.

Mu,F. *et al*. (2006) Prediction of oxido-reductase-catalyzed reactions based on atomic properties of metabolites. *Bioinformatics*, **22**, 3082–3088.

Muller,M. (2004) Chemical diversity through biotransformations. *Curr. Opin. Biotechnol.*, **15**, 591–598.

Mueller,L.A. *et al*. (2003) AraCyc: a biochemical pathway database for Arabidopsis plant physiology. *Plant Physiol*., **132**, 453–460.

Nakahigashi,K. *et al*. (2009) Systematic phenome analysis of Escherichia coli multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Mol. Syst. Biol.*, **5**, Article no. 306.

Nobeli,I. *et al.* (2009) Protein promiscuity and its implications for biotechnology. *Nat. Biotech.*, **27**, 157–167.

Payne,M.P. (2004) Computer-based methods for the prediction of chemical metabolism and biotransformation within biological organisms. In Cronin,M.T.D. and Livingstone,D.J. (eds) *Predicting Chemical Toxicity and Fate*. CRC Press, Baca Raton, FL, pp. 205–227.

Reed,J.L. *et al*. (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol.*, **4**, R54.51-12.

Rendic,S. (2002) Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Metab. Rev.*, **34**, 83–448.

Romero,P. *et al.* (2003) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.*, **4**, R54.

Saito,N. *et al.* (2010) Unveiling cellular biochemical reactions via metabolomics-driven approaches. *Curr. Opin. Microbiol.*, **13**, 358–362.

Schreiber,S.I. (2005) Small molecules: the missing link in the central dogma. *Nat. Chem. Biol.*, **1**, 64–66.

Schwab,W. (2003) Metabolome diversity: too few genes, too many metabolites? *Phytochemistry*, **62**, 837–849.Silverman,R.B. (2000) *The Organic Chemistry of Enzyme-Catalyzed Reactions*. Academic Press, San Diego, CA.

Soh,K.C. and Hatzimanikatis,V. (2010) DREAMS of metabolism. *Trends Biotechnol.*, **28**, 501–508.

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.

Stewart,J.J.P. (2007) Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. *J. Mol. Model.*, **13**, 1173–1213.

van der Werf,M.J. *et al*. (2005) Microbial metabolomics: replacing trial-and-error by the unbiased selection and ranking of targets. *J. Ind. Microbiol. Biotechnol.*, **32**, 234–252.

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

Vaz,R.J. *et al*. (2010) The challenges of in silico contributions to drug metabolism in lead optimization. *Exp. Opin. Drug Metab. Toxicol.*, **6**, 851–861.

Wishart,D.S. (2007) Current Progress in computational metabolomics. *Brief. Bioinformatics*, **8**, 279–293.