BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 22 2009, pages 2975–2982
doi:10.1093/bioinformatics/btp507

Data and text mining

Metabolite and reaction inference based on enzyme specificities

M. J. L. de Groot[1,2,3,4,†], R. J. P. van Berlo[1,3,†], W. A. van Winden[2,3], P. J. T. Verheijen[2,3], M. J. T. Reinders[1,3,4] and D. de Ridder[1,3,4,*]

[1]The Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD Delft, [2]Bioprocess Technology Group, Department of Biotechnology, Delft University of Technology, Julianalaan 67, 2628 BC Delft, [3]Kluyver Centre for Genomics of Industrial Fermentation, P.O. Box 5057, 2600 GA Delft and [4]Netherlands Bioinformatics Center, 260 NBIC, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

Received on June 2, 2009; revised on August 7, 2009; accepted on August 9, 2009

Advance Access publication August 20, 2009

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Many enzymes are not absolutely specific, or even promiscuous: they can catalyze transformations of more compounds than the traditional ones as listed in, e.g. KEGG. This information is currently only available in databases, such as the BRENDA enzyme activity database. In this article, we propose to model enzyme aspecificity by predicting whether an input compound is likely to be transformed by a certain enzyme. Such a predictor has many applications, for example, to complete reconstructed metabolic networks, to aid in metabolic engineering or to help identify unknown peaks in mass spectra.

Results: We have developed a system for metabolite and reaction inference based on enzyme specificities (*MaRIboES*). It employs structural and stereochemistry similarity measures and molecular fingerprints to generalize enzymatic reactions based on data available in BRENDA. Leave-one-out cross-validation shows that 80% of known reactions are predicted well. Application to the yeast glycolytic and pentose phosphate pathways predicts a large number of known and new reactions, often leading to the formation of novel compounds, as well as a number of interesting bypasses and cross-links.

Availability: MATLAB and C++ code is freely available at https://gforge.nbic.nl/projects/mariboes/

Contact: d.deridder@tudelft.nl

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

1 INTRODUCTION

In biotechnology, much effort is spent on altering metabolism, mainly of industrially relevant microorganisms such as bacteria, yeasts and fungi. In most cases, the aim is to increase existing product yield or to introduce and optimize a pathway to a new product. To be able to perform such metabolic engineering, one needs a full description of the metabolism of the species of interest: to select desired functions (enzymes) needed to introduce a new pathway, to unravel metabolic regulation, to find bottlenecks in metabolism, and to reveal undesired bypasses. Missing functions or 'gaps' in this metabolic network description make metabolic engineering difficult; but even when the main pathways are known, missing bypasses or cross-links may pose problems. It is therefore essential to have a full overview of all possible metabolic reactions in the cell.

The metabolic networks of model organisms are mostly sufficiently characterized and annotated (e.g. Feist and Palsson, 2008; Herrgard *et al.*, 2008). For newly sequenced species, metabolic functions are usually derived by looking for genes homologous to known enzymes in other species (e.g. Pireddu *et al.*, 2006). At a certain stage, homology does not suffice to complete the metabolic network, i.e. to fill the remaining gaps in a network, to connect dead ends, or to create links between fragmented (sub)networks and pathways. In such cases one needs to perform an extensive manually search for functions or pathways (Feist *et al.*, 2009).

To complete metabolic networks, enzyme functions can also be inferred from metabolome data, such as mass spectra (MS). Although high-resolution techniques and advanced pathway extraction tools are available (Breitling *et al.*, 2006; Gipson *et al.*, 2008), it is still not always possible to uniquely identify compounds, as the MS 'peaks' are not sufficiently accurate. Even when measurements are perfect, structural isomers cannot be distinguished by mass alone.

Alternatively, metabolic networks can be completed by exploiting enzyme functionality information. The key idea is that (at least some) enzymes are known to be aspecific, i.e. able to perform the associated chemical transformation on compounds other than the one traditionally associated with that enzyme (D'Ari and Casadesús, 1998). Some enzymes can even perform slightly different transformations (O'Brien and Herschlag, 1999). Modeling this aspecificity is important for biotechnology and poses significant bioinformatics challenges; for example, predicting aspecificity based on mining the available enzyme characterization data (Nobeli *et al.*, 2009).

A number of researchers have explored the idea of predicting metabolic reactions based on an analysis of the basic biochemical transformations performed by enzymes (Hatzimanikatis *et al.*, 2005;

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

Li *et al.*, 2004). Specifically in the field of the biotransformation of xenobiotics (substances foreign to a biological system), several such systems have been developed. These tools, such as METEOR, META (Klopman *et al.*, 1999), CATABOL (Dimitrov *et al.*, 2004), UM-PPS (Ellis *et al.*, 2008), etc., mainly consist of manually supplied reaction rules and heavily depend on user selection of feasible predicted pathways.

In Oh *et al.* (2007), instead of manually created rules, xenobiotic reaction possibilities were derived using measures of structural similarity between compounds, which were represented as graphs. Chemical transformations were captured in so-called reaction patterns (RDM, or Reaction centre-Difference-Matched patterns). A given query compound is assumed to be converted by an enzyme when its RDM pattern is present and the compound is sufficiently similar to known substrates of all enzymes with the same RDM pattern. To develop their system, Oh *et al.* (2007) used the KEGG database (Kanehisa and Goto, 2000), which describes only enzymatic reactions with well-known metabolic function. The limited amount of data available means that generalization can only take place per RDM pattern, rather than per enzyme. The latter would be desirable, as enzymes with identical RDM patterns can be specific to different types of substrates.

We present a novel system for metabolite and reaction inference based on enzyme specificities (*MaRIboES*), building on the work of Oh *et al.* (2007). We generalize an enzymatic transformation by training a classifier on the list of activities found for that enzyme in the Braunschweig Enzyme Database (BRENDA; Barthelmes *et al.*, 2007). This unique database contains a large number of enzyme activities reported in literature, found by detailed enzyme characterization including non-metabolically relevant compounds, toxic compounds, etc. We demonstrate the potential of our method by performing both an internal validation and an application to extend the glycolysis and pentose phosphate pathways of *Saccharomyces cerevisiae*, in which we predict a number of reactions (bypasses and cross-links), many of which lead to the formation of novel compounds.

2 METHODS

Our system takes four steps to get from the enzyme activity data in BRENDA to a trained classifier for each enzymatic transformation. First, reaction and compound data are extracted from BRENDA and preprocessed. Second, enzymatic reactions are defined in terms of reaction patterns (changes in molecular structure), which are then used to derive enzymatic transformations between reactant pairs. Third, sets of candidate compounds for each transformation (having the reaction pattern) are characterized by a set of structural, stereo and fingerprint features. Finally, a selection of these features is used to build a classifier for each enzymatic transformation, predicting for novel compounds whether that transformation is likely to occur.

2.1 Extracting data from BRENDA

We downloaded version 0702 of BRENDA (August 22, 2008) and selected all information available for 3016 enzymes present in *S.cerevisiae*. In this way, we obtained a list of reactions known to be catalyzed by each enzyme. A total of 3360 compounds were involved in these reactions, 1146 of which are known as natural substrates or products. We also gathered the chemical structures (mol-files) of the substrates and products of all enzymatic reactions; these were available for 2399 of the 3360 compounds (August 2008).

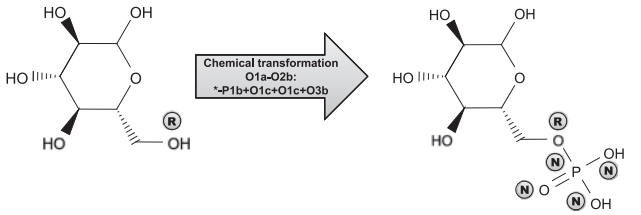

Fig. 1. Illustration of an RN pattern. Here, a phosphate group is added to D-glucose, obtaining D-glucose 6-phosphate. The RN pattern, O1a-O2b:*-P1b+O1c+O1c+O3b, fully describes the chemical transformation, using the atom types defined in Hattori *et al.* (2003). R- and N-atoms are indicated; all other atoms are I-atoms.

BRENDA contains a number of duplicates, i.e. compounds with the same composition and structure but different names (e.g. glucose and D-glucose). To remove these, we employed the following strategy. First, we examined which pairs of compounds had the same binary and discrete fingerprints (see Section 2.3). For these pairs, we subsequently calculated similarity in structure and stereochemistry. If two compounds were completely equal in structure and stereochemistry, we treated these compounds as identical. This left 1914 unique molecular structures, 753 of which correspond to natural substrates or products.

Of several compounds (particularly monosaccharides), we found both the linear and the ring variant in the database. This poses problems for the definition of reactant pairs based on compound similarity (see Section 2.2), particularly if a substrate is in linear form and the corresponding product in ring form. Therefore, we first automatically transformed all compounds in linear form which are likely to be in ring form in solution, into their ring variants. To accomplish this, we first detected whether a compound had a ketone or aldehyde group. If so, we assessed whether it was indeed a Fisher projection of a saccharide. If this was also the case, we adjusted the bonds such that the compound became cyclic, and subsequently generated the Haworth projection using MarvinView (ChemAxon, 2009).

2.2 Inferring enzymatic transformations

2.2.1 Defining enzymatic transformations We defined chemical structure transformation patterns similar to the proposal in Oh *et al.* (2007). They represented molecules as graphs and looked for differences between the substrate and product of an enzymatic reaction (the reactant pair). First, these graphs were aligned (see below), to obtain a matched and unmatched part. The boundary atom between the matched and unmatched part is called the reaction center (R-atom); the atom(s) adjacent to the R-atom in the unmatched parts are the difference atom(s) or D-atom(s); and the atom(s) adjacent to the R-atom in the matched region are the matched atom(s) or M-atom(s).

Unlike Oh *et al.* (2007), we focus on describing *all* changes in molecular structure due to a reaction, i.e. all unmatched atoms rather than just those connected to the reaction center. Therefore, we distinguish two different atom types besides the reaction center (R-atom): identical or I-atoms, all matched atoms except the R-atom; and non-identical or N-atoms, all unmatched atoms (not to be confused with iodine (I) or nitrogen (N) atoms). The RN pattern describing the entire transformation between a reactant pair (Fig. 1) then consists of an RN pattern for the substrate (RNs) and one for the product (RNp).

2.2.2 Inferring enzymatic transformations The representation of chemical and biological molecules by means of graphs permits the use of a maximum common subgraph (MCS) algorithm to identify the chemical structure transformation pattern between a reactant pair (Gardiner *et al.*, 1997). Many existing algorithms convert the MCS problem into a maximum clique finding problem, by introducing an association graph (Hattori *et al.*, 2003). Due to the nature of chemical structures, this association graph usually is very

dense, making clique finding computationally prohibitive. Cao *et al.* (2008) proposed an algorithm that directly operates on the chemical structure graphs themselves. Still, although much more computationally efficient than algorithms based on clique finding, it cannot always infer the MCS in reasonable computation time without progressive optimization (human intervention), especially when the MCS is large. The main bottleneck of the algorithm is that the common subgraph is extended by only one node at a time. As a consequence, many different ways exist in which the same common subgraph can be constructed. We adjusted their method to speed up the process (van Berlo *et al.*, 2009), enabling us to detect *all maximal* substructures common to a pair of molecules, rather than only the *maximum* one (a *maximal* common subgraph is any complete common subgraph not contained in any other complete common subgraph; the MCS is the largest of these).

A natural score for a common subgraph thus found would be its size, the number of matched atoms (R- and I-atoms). However, to reflect the prior knowledge that most enzymes affect a molecule at only one point, we can assign a lower weight to R-atoms than to I-atoms. Furthermore, as many reactions add a phosphate group to a molecule as a single, elementary unit (by extracting it from ATP), it would be desirable to count this group as a single atom rather than four. Hence, we adopted the following similarity score (SS) for weighting the different maximal common subgraphs found between graphs G_1 and G_2:

$$\mathrm{SS}(G_1,G_2)=w_1|\mathrm{R}|+w_2|\mathrm{I}|+w_3|\mathrm{PO_3}| \quad (1)$$

Here, |R| denotes the number of R-atoms, |I| the number of I-atoms and $|PO_3|$ the number of aligned phosphate groups. For the reasons given above, the weight vector $\mathbf{w}=(w_1,w_2,w_3)$ was set to $(0.5,1,-3)$, as this favors substructures that (i) include long backbones (and/or few phosphate groups) and (ii) contain few reaction centers. As a result, the MCS will not necessarily lead to the highest similarity score. This emphasizes the need for an algorithm that can identify all *maximal* common subgraphs.

Like Hattori *et al.* (2003), we used the Jaccard coefficient (also known as the Tanimoto coefficient) to adjust the similarity score for the size of the two aligned graphs G_1 and G_2:

$$\mathrm{JC}(G_1,G_2)=\frac{\|G^{\mathrm{opt}}(G_1,G_2)\|}{\|G_1\|+\|G_2\|-\|G^{\mathrm{opt}}(G_1,G_2)\|} \quad (2)$$

where $G^{\mathrm{opt}}(G_1,G_2)$ is the highest scoring maximal subgraph common to G_1 and G_2 according to (1) and $\|G\|$ indicates the number of nodes in G.

2.2.3 Defining reactant pairs As an enzymatic reaction usually involves multiple substrates and products, we employed an iterative procedure to find all reactant pairs. We first selected from all possible substrate–product combinations the one resulting in the highest JC. Second, we removed the corresponding G^{opt} from both the substrate and product of this reactant pair. This procedure was iterated until all atoms in all substrates and products were part of some G^{opt}. Figure 2 illustrates the procedure for the hexokinase reaction with D-glucose as substrate, showing that the highest JC is obtained when ATP is aligned with ADP. The corresponding G^{opt} is removed from both molecules. In the second step, the reactant pair consisting of D-glucose and D-glucose 6-phosphate yields the highest JC. In the final step, the remaining phosphate groups of ATP and D-glucose 6-phosphate are aligned. In this way, we infer a complete reaction definition.

Note that we can obtain multiple RN patterns for a single enzyme, as enzymes can catalyze similar reactions and reverse reactions are taken into account as well. Let t_e denote a chemical transformation (i.e. an RN) as accomplished by enzyme e. We define T_e as the set of all possible chemical transformations that can be accomplished by enzyme e, and the set E_t as the set of all enzymes that can perform the same chemical transformation t.

2.3 Characterizing candidate compounds

In general, the RN patterns in T_e are quite consistent between the different reactions listed for e in BRENDA. This allowed us to infer possible

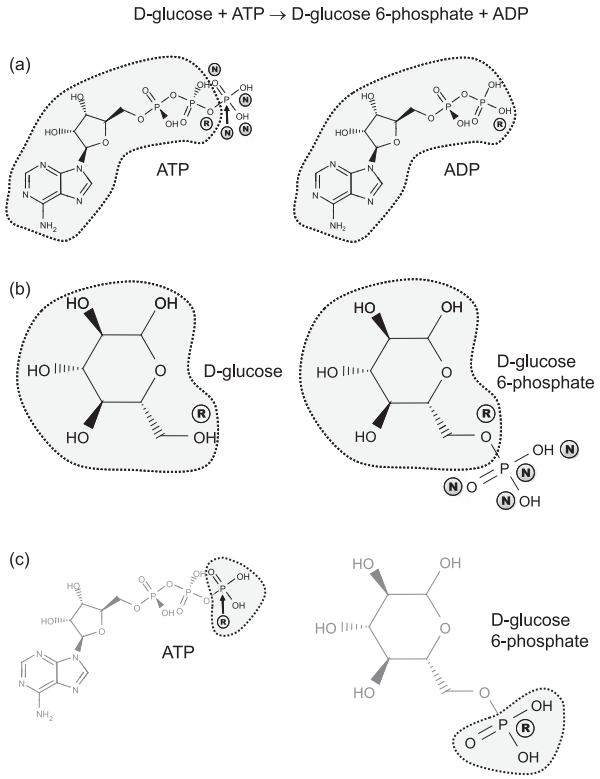

Fig. 2. Reactant pairs inferred for a hexokinase reaction in which D-glucose is converted into D-glucose 6-phosphate. (**a**) The best alignment is between ATP and ADP, (**b**) the next best between D-glucose and D-glucose 6-phosphate and (**c**) finally, the remaining parts of ATP and D-glucose 6-phosphate are aligned (atoms already matched are no longer considered), resulting in G^{opt} being the phosphate group. R- and N-atoms are indicated, all other atoms are I-atoms.

new substrates for an enzyme e, by searching for candidate compounds, containing a particular RNs- or RNp-pattern. Let c_e^t denote a particular candidate compound c for the chemical transformation t as accomplished by enzyme e, and C_e^t the set of all such candidates. We divided this set into substrates or products involved in reactions listed in BRENDA as catalyzed by e, P_e^t ('positive' examples), and possible new candidates N_e^t ('negative' examples). The end goal is to construct a classifier to predict whether or not a particular candidate compound c can be transformed by a chemical transformation t_e of an enzyme e, based on a number of features of that compound. We expect that c is more likely to be transformed if it is similar to the compounds in the set of positive examples P_e^t. Therefore, we characterize c by the following potentially useful features:

- *Structural similarity*: just as the MCS algorithm was applied to determine the chemical transformation of a reactant pair, we used it to infer the G^{opt} between c and each of the (other) positive examples. We used the largest JC to construct a structural similarity feature, i.e. we employ the similarity to the closest positive example as a feature.
- *Stereo dissimilarity*: we believe that for several enzymes, stereo (3D) information can be an important feature for determining whether or not a compound can be transformed by an enzyme e. Therefore, we inferred whether the stereochemistry of c matched that of the positive examples. To this end, for all positive examples we used the alignment as obtained by the MCS algorithm and counted the number of times a stereo bond differs, as illustrated in Figure 3. We selected

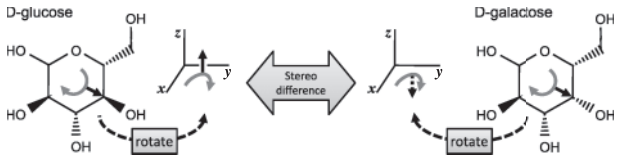

Fig. 3. Comparison of D-glucose and D-galactose results in one stereo difference. MCSs are rotated in 3D such that their major axes are aligned in the $x-y$ plane; stereo differences can then be easily identified by checking the z-coordinates of atoms.

the minimum of these for constructing a stereo-based distance feature, i.e. we employ the difference with the closest positive example as a feature.

- *Binary and discrete fingerprints*: fingerprinting is one of the most widely applied methods for measuring similarity. A fingerprint of a compound is usually a binary vector in which each element is nonzero if the corresponding feature (e.g. 'it is an alcohol') holds for that compound, and zero otherwise. Sometimes these fingerprints are discrete instead of binary. In this case, each element denotes the *number of times* a particular feature (e.g. a certain type of atom) is present in the compound. We used the 204 binary and 57 discrete fingerprints as defined by Checkmol (Haider, 2003).

2.4 Constructing a classifier

As a final step of our algorithm, we constructed a classifier based on the available features, which outputs a posterior probability $p_e^{t,q}$ of a compound q being transformed by transformation t_e. For this, we used a Parzen kernel density estimation-based classifier (Duin, 2009) as, in our experience, it works well in situations where there are only few examples.

First, all 263 features (1+1+204+57) were individually normalized by applying mean-variance normalization. Subsequently, they were ordered based on their mutual information with the labels (negative or positive), i.e. the amount of information that can be inferred about the labels by observing the features. We prefer this option over using a wrapper approach (Wessels *et al.*, 2005), since wrapper approaches have exponential time complexity and often do not improve classification performance (Lai *et al.*, 2006). Stereo dissimilarity is only predictive in conjunction with any of the other features. For example, if two molecules have exactly the same 2D-structure (e.g. glucose and galactose), then stereo similarity enables us to distinguish these compounds. Therefore, although mutual information between stereo similarity and the labels is often low, we always included it as a feature.

Next, the optimal number of features k^{opt} was found using leave-one-out cross-validation (LOOCV). For $k=1,\ldots,263$, we trained a classifier on the first k of the ordered features, determined the ROC (receiver operator characteristic; Duda *et al.*, 2001) curve and used the area under this curve (AUC) to evaluate the performance of the classifier. The smoothing parameter s of the Parzen density estimate was also optimized on the training data, using the leave-one-out Lissack and Fu estimate (Duin, 2009). This means that at least three positive examples and three negative example should be available, to allow estimation of the two parameters k^{opt} and s. This was the case for 137 RN pattern-enzyme combinations, of 78 unique RN patterns and 57 enzymes. To prevent overfitting, we employed the same smoothing parameter s for both the negative and positive class and for each feature.

3 EXPERIMENTAL SETUP

3.1 Validation

In a first experiment, we assessed whether we could predict the correct enzymatic transformation t_e for a particular query compound q. To test prediction performance, we first determined whether this compound could be transformed by enzyme e; that is, whether it contains one of its RN patterns, corresponding to one of its enzymatic transformations. If so, then q serves as either a positive example for t_e (if q is listed in BRENDA as the substrate or product of a reaction catalyzed by e) or a negative example (if not). We then removed q from the training set, learned the prediction rule based on the remaining training samples ($P_e^t \backslash q$ and $N_e^t \backslash q$), and applied the resulting rule to query compound q to assess whether it is likely to be transformed by that particular enzymatic transformation of that enzyme. It was necessary that, after removing the query compound q, at least three positive and three negative examples remained for training (see Section 2.4).

More formally, let t_e^q denote a chemical transformation t accomplished by enzyme e, that can be applied to query compound q, and for which the training set except query compound q contains at least three positive as well as negative examples. Let T_E^q denote the full set of chemical transformations that can be applied to query compound q by any of the enzymes in set E. Note that this set may contain multiple chemical transformations for one enzyme but also a single chemical transformation for multiple enzymes. Let $l_e^{t,q}$ be a label equal to one if query compound q is a positive example for the chemical transformation t as performed by enzyme e and zero otherwise, and let $p_e^{t,q}$ be the posterior probability as calculated by applying the corresponding prediction rule to the query compound. We can then order the posterior probabilities to analyze whether transformations predicted as the most likely indeed correspond to transformations for which q is a positive example, i.e. if $p_e^{t,q}$ correlates with $l_e^{t,q}$.

3.2 Application to central metabolism of *S.cerevisiae*

In a second experiment, we focused on the well-described central metabolism of *S.cerevisiae*, specifically glycolysis and the pentose phosphate pathway. We used all compounds known to be involved in these pathways as input for all chemical transformation classifiers. For each compound-transformation pair (q,t_e), this leads to a posterior probability $p_e^{t,q}$ of q being a substrate for t_e (note that if q does not contain the RN-pattern, no probability can be calculated). Each reaction for which this posterior exceeded 0.9 was then automatically predicted to occur.

It is hard to base decisions on these posteriors alone, as some transformations yield many more predictions than others, i.e. are far less specific. Therefore, for each transformation t_e, we also ranked the n predicted compounds by the posterior probability $p_e^{t,q}$. We then applied a hypergeometric test with the null hypothesis that the top j was *not* enriched for true positives, using the remaining $n-j$ compounds as background, for $j=1,\ldots,n$. Compound j was then predicted to be transformed by t_e if the corresponding null hypothesis was rejected (i.e. if the Bonferroni-corrected $P<0.05$).

At the end of this prediction step, we checked whether each predicted product (found by applying the chemical transformation to the substrate) is already listed in BRENDA or KEGG, by looking for an identical compound (see Section 2.1). If no match was found, the predicted new compound was given a unique identifier (*new...*). The compounds were translated to SMILE strings (ChemAxon, 2009) and searched for using the ChemSpider search engine (ChemZoo, 2007). If found, the relevant compound name was assigned;

otherwise, the compound was annotated manually or given a standard IUPAC name (ChemAxon, 2009).

We re-iterated this entire procedure twice, using the compounds predicted to be formed in the previous iteration as new inputs.

4 RESULTS AND DISCUSSION

4.1 Validation

4.1.1 Most predictions are reliable For any compound q, we would like transformations t_e for which $l_e^{t,q}=1$ (i.e. for which q is known substrate of enzyme e) to be predicted with high probability. To verify this, q can be left out during classifier training. After training, we can then test whether the correct transformations are indeed highly ranked. Figure 4 lists the 130 compounds thus tested. For each compound q, a bar indicates a ranked list of all transformations T_E^q, from high (left) to low (right) probability transformations. Black elements indicate transformations for which q is a known substrate or product; dark gray elements indicate either different transformations of enzyme e (members of T^e, see Section 2.2), or identical transformations t by a different enzyme (members of E^t).

For 77 compounds, an actual enzyme reaction annotated in BRENDA to perform this particular conversion (i.e. a black element) is found as the most likely transformation. For an additional 27 compounds, this is the second most likely transformation and for 26 compounds, the true conversion ranks lower. For 13 of the latter 53 compounds, related enzymes or transformations are predicted as most likely (dark gray elements). In the 26 compounds for which the actual transformations were not ranked highly, cofactors (e.g. ADP, UDP, etc.) are overrepresented. These compounds play a very generic role in many transformations, and the RN patterns hence occur in many transformation classifiers, increasing the chances of misclassification.

We conclude that, as 80% of the results (77 + 27 out of 130) are good, our system produces reliable predictions even given the relatively limited amount of training data available.

4.1.2 Features selected reflect enzyme specificities In a second experiment, we investigated whether the features used for building the prediction rule differed between different enzymatic transformations. For this, we compared all prediction rules corresponding to the same transformation t but accomplished by different enzymes, E_t. Four of these results are shown in Supplementary Figure 1. The figures show that not only the *type*, but also the *number* of features used in the prediction rule can be quite different between enzymatic transformations.

Structural similarity based on MCS frequently seems to be most relevant, and hence is a highly predictive feature. This agrees with the findings of Oh *et al.* (2007). We do find that for different enzymes that perform the same conversion, different subsets of fingerprint features are selected. This indicates that specificity is governed by different structural features for each individual enzyme. It also demonstrates that it is apparently beneficial (in terms of predictive power) to use more than just structural similarity.

4.2 Application to central metabolism of *S.cerevisiae*

To demonstrate how our system can yield practical predictions, we applied it repeatedly to the glycolysis and pentose phosphate pathways of *S.cerevisiae*, to see which new reactions and possibly new compounds would be predicted. The reactions in these pathways were used in training and hence not considered as new predictions. We performed three iterations; in the second and third, we used only compounds predicted to be produced in the previous iteration as input. Table 1 gives an overview of the number of predicted novel reactions and compounds; Supplementary Figure 2 gives a detailed overview of all outcomes. As there are far too many predictions to discuss, below we focus on some key findings.

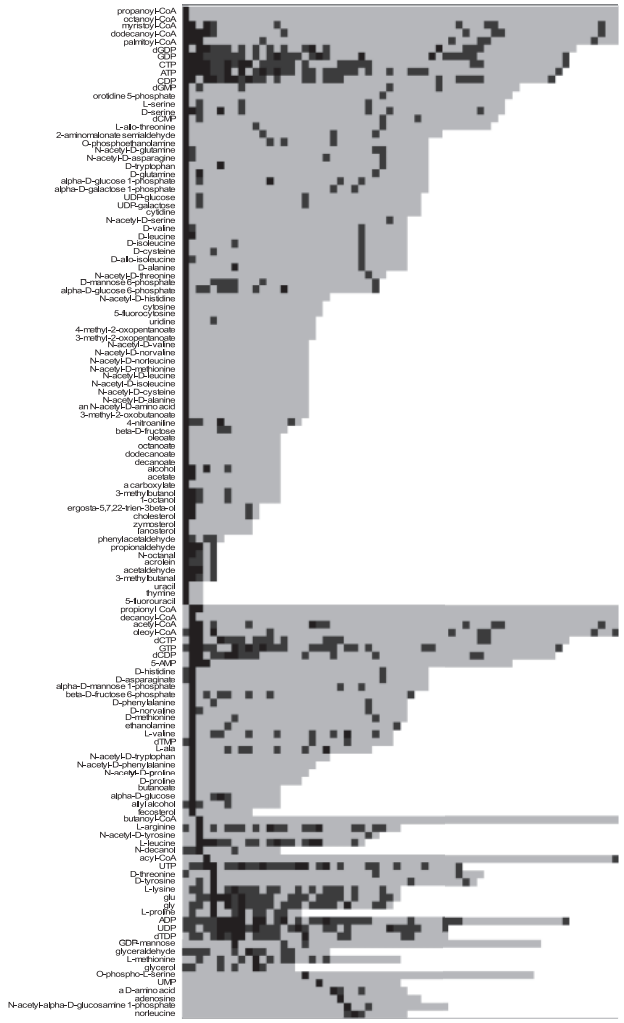

Fig. 4. Classification performance for different query compounds q. Each row represents the results for a particular q, showing the different candidate transformations in T_E^q ordered from the most likely (left) to the most unlikely (right). Black elements: chemical transformations t_e^q listed in BRENDA ($l_e^{t,q}=1$). Dark gray: either different transformations of q by the same enzyme e (in T^e) or identical transformations t by a different enzyme (in E^t); light gray: chemical transformations t_e^q that could in principle be applied to compound q, but for which we have no biological evidence ($l_e^{t,q}=0$).

4.2.1 Many predicted reactions are known to exist We predict a large number of new reactions. The corresponding chemical transformations were not present in the dataset we used to train our system, either because BRENDA does not list them or because

Table 1. Number of novel reactions and compounds predicted by the system when applied to the glycolysis and pentose phosphate pathways in *S.cerevisiae*, in three iterations

Iteration	Novel reactions predicted	Of which in KEGG	Novel compounds predicted	Of which in BRENDA or KEGG
1	70	17	62	11
2	109	37	58	7
3	84	8	41	1
Total	263	62	161	19

it lists less than three conversions with accompanying structure information (mol-file), too little to train the classifier on. Strikingly, a large number of these predicted reactions are listed in KEGG (see Table 1, left columns and the green arrows in Supplementary Figure 2). This demonstrates that our system is able to generalize well, and indicates that it is potentially useful in, for example, metabolic network reconstruction and metabolic engineering.

4.2.2 Enzymatic alternatives to autocatalytic reactions are found Some predicted reactions occur in pathways described in literature as autocatalytic, i.e. not requiring enzymes. For example, an autocatalytic pathway has been reported from dihydroxyacetone phosphate and glyceraldehyde 3-phosphate to hydroxypyruvaldehyde (Thornalley *et al.*, 1984). However, we find leads suggesting possible enzyme catalyzed conversions (Fig. 5a). In the first iteration of our system, dihydroxyacetone phosphate is predicted to be transformed into dihydroxyacetone (an existing reaction); in the second iteration, this is further transformed into hydroxypyruvaldehyde. A similar path is predicted from glyceraldehyde 3-phosphate via glyceraldehyde to hydroxypyruvaldehyde. Perhaps the corresponding enzymes are required to decrease the activation energy for these reactions only under *in vivo* conditions and hence may have been missed in Thornalley *et al.* (1984).

4.2.3 Interesting bypasses and cross-links are suggested Some existing bypasses are predicted. For example, an isoenzyme conversion from fructose 1,6-bisphosphate to fructose 6-phosphate is predicted, as well as a longer bypass via fructose 1-phosphate and fructose (Fig. 5b). The reactions involved are all found in KEGG (RPAIRS RP00242, RP00680 and RP00210). Interestingly, an even longer bypass is predicted through 1-keto-D-fructose and 1-keto-D-fructose 1-phosphate. Another example is the predicted formation of lactate and subsequently lactoyl-CoA from both pyruvate and phosphoenolpyruvate. Although this is not an annotated path in *S.cerevisiae*, this cross-link may be interesting for sterol biogenesis and propionate metabolism, in which lactoyl-CoA is involved (KEGG PATHWAYS KO00643 & KO00640). Missing such bypasses or cross-links could cause problems when applying flux analysis, as the flux through the known reactions may be overestimated, leading to incorrect conclusions on possible bottleneck reactions.

4.2.4 Predictions of obscure metabolites make sense Our system predicts the production of a number of compounds that are

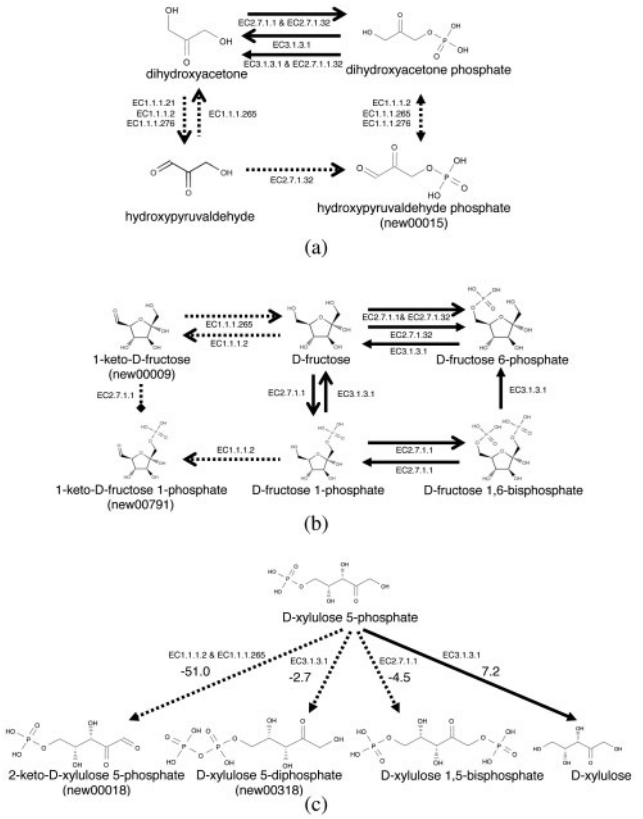

Fig. 5. Example pathways constructed using the MaRIboeS algorithm. (**a**) Pathway containing reactions originally reported as autocatalytic. (**b**) Prediction of possible pathways around phosphofructokinase. (**c**) Predictions from D-xylulose 5-phosphate. The direction of the arrows indicates the reaction direction. Predictions from the first, second and third iterations are indicated by closed, open and square arrowheads, respectively. EC numbers denote the enzyme for which an associated classifier predicted the reaction. In (**c**), numbers indicate the estimated Gibbs free energy of the reaction ($\Delta_r G'^o$, kcal/mol).

not yet listed as part of a traditional pathway (Table 1, right columns), but that have been described in literature. This suggests that the corresponding enzymatic transformations may have been overlooked. Figure 5c shows an example of D-xylulose 5-phosphate, which is predicted to be transformed into four compounds: D-xylulose (a known reaction, KEGG RP01652), D-xylulose 1,5-bisphosphate, 2-keto-D-xylulose 5-phosphate and a diphosphate. The latter three compounds are not known to exist, but the reactions seem possible from an energy point of view (the estimated Gibbs free energy of each reaction (Jankowski *et al.*, 2008) is shown in the figure). For D-ribulose 5-phosphate a reaction similar to the second one, to D-ribulose 1,5-bisphosphate, is predicted as well. This compound is a known substrate for glyoxylate and dicarboxylate metabolism (KEGG pathway KO00630).

This suggests that the predicted formation of D-xylulose 1,5-bisphosphate may be valid. It has been described before as being produced from D-xylose in algae (Wu *et al.*, 1970); as a side product of an enzymatic reaction involving a misprotonation (Edmondson *et al.*, 1990); as an inhibiting factor for growth on D-xylose and a strong competitive inhibitor of Rubisco (Andersson, 2008).

In light of metabolic engineering efforts in which the fermentation of pentoses is engineered in *S.cerevisiae* (Hahn-Hägerdal *et al.*, 2007; Wisselink *et al.*, 2009), it may be wise to attempt to circumvent production of this compound.

Another example of a predicted obscure metabolite is hydroxypyruvaldehyde phosphate, produced directly from dihydroxyacetone phosphate and through a reaction involving a kinase acting on hydroxypyruvaldehyde (Fig. 5a). This compound has been reported as a substrate of glyoxalase in *S.cerevisiae* (Weaver and Lardy, 1961); as a product of transaldolase (Christen and Gasser, 1976); and in erythrocytes when provided with glucose (Cogoli-Greuter and Christen, 1981). It has also been reported to react with hydrogen peroxide acting as an antioxidant (Cogoli-Greuter and Christen, 1981), which may point to an interesting application.

Not all predictions are easily explained. For example, phosphorylation of phosphate groups is performed on nucleotides; our system predicts this to occur as well on, for example, D-xylulose 5-phosphate, D-ribulose 5-phosphate, phosphoenolpyruvate and glycerol 1-phosphate. Although these compounds have not yet been described, perhaps they play a (minor) role in metabolism. Other predicted compounds are seemingly instable (e.g. containing two neighboring keto groups).

5 CONCLUSIONS

We have described MaRIboES, a system to predict possible enzymatic transformations as well as the resulting output compounds, given a set of input compounds. Our work significantly extends that of Oh *et al.* (2007). First, we generalize chemical transformations based on experimental data available in BRENDA, rather than pathway descriptions in KEGG. This allows us to include non-metabolically relevant conversions and to predict for each individual enzyme rather than for classes of enzymes. Second, we added both stereochemistry similarity and molecular fingerprints. Stereo information is essential, as many enzymes are known for their chiral specificity. Fingerprint features were often selected for prediction, indicating they are useful.

Our system was validated using a metabolome-wide leave-one-out procedure. For over 80% of the compounds, we predict the enzyme associated with the compound as the first or second most likely one. Next, we applied it to metabolites in the glycolysis and pentose phosphate pathways of *S.cerevisiae*. Besides reactions leading to well-annotated metabolites, we predict formation of novel compounds, for which we can find some confirmation in other organisms. We also predict enzymatic alternatives for reactions thought to be autocatalytic, interesting bypasses within and cross-links between pathways.

We foresee a number of applications besides the ranking of possible substrates for enzyme characterization. First, automated metabolic network reconstruction could be improved, by using MaRIboES to calculate function-based rather than sequence-based similarity between enzymes. Second, prediction of possible bypasses and cross-links can benefit metabolic engineering, by charting alternative routes and identifying potentially competing compounds. Finally, the predicted compounds may help to interpret metabolomics data, by listing possible candidates for unidentified masses.

MaRIboES' performance could be improved by including some estimate of stability (not yet readily available) of the compound predicted to be formed, and by comparing estimates of the activation energy of a chemical reaction to those of natural substrates. However, it would benefit most from a proper characterization of more enzyme activities (see Supplementary Figure 3). Although current research invests heavily in high-throughput analyses of genome expression, proteome levels and modifications, physical interactions and metabolites, it would be wise not to forget that these all rely on basic principles unraveled by looking at details rather than the big picture.

Funding: The Netherlands Bioinformatics Center (to MdG) and the Kluyver Centre for Genomics of Industrial Fermentation (to RvB), both supported by the Netherlands Genomics Initiative.

Conflict of Interest: none declared.

REFERENCES

Andersson,I. (2008) Catalysis and regulation in Rubisco. *J. Exp. Bot.*, **59**, 1555–1568.

Barthelmes,J. *et al.* (2007) BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res.*, **35**, D511–D514.

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

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

ChemAxon (2009) Marvinview. Available at http://www.chemaxon.com/marvin (last accessed date January, 2009).

ChemZoo (2007) Chemspider. Available at http://www.chemspider.com/ (last accessed date May, 2009).

Christen,P. and Gasser,A. (1976) Oxidation of the carbanion intermediate of transaldolase by hexacyanoferrate (III). *J. Biol. Chem.*, **251**, 4220–4223.

Cogoli-Greuter,M. and Christen,P. (1981) Formation of hydroxypyruvaldehyde phosphate in human erythrocytes. *J. Biol. Chem.*, **256**, 5708–5711.

D'Ari,R. and Casadesús,J. (1998) Underground metabolism. *Bioessays*, **20**, 181–186.

Dimitrov,S. *et al.* (2004) Predicting the biodegradation products of perfluorinated chemicals using CATABOL. *SAR QSAR Environ. Res.*, **15**, 69–82.

Duda,R. *et al.* (2001) *Pattern Classification*, 2nd edn. Wiley, New York, NY.

Duin,R. (2009) PRTools, a MATLAB pattern recognition toolbox. Available at http://prtools.org/ (last accessed date February, 2009).

Edmondson,D.L. *et al.* (1990) Slow inactivation of ribulosebisphosphate carboxylase during catalysis is caused by accumulation of a slow, tight-binding inhibitor at the catalytic site. *Plant Physiol.*, **93**, 1390–1397.

Ellis,L.B.M. *et al.* (2008) The University of Minnesota pathway prediction system: predicting metabolic logic. *Nucleic Acids Res.*, **36**, W427–W432.

Feist,A.M. and Palsson,B.O. (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat. Biotechnol.*, **26**, 659–667.

Feist,A.M. *et al.* (2009) Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.*, **7**, 129–143.

Gardiner,E. *et al.* (1997) Clique-detection algorithms for matching three-dimensional molecular structures. *J. Mol. Graphics Model.*, **15**, 245–253.

Gipson,G. *et al.* (2008) Assignment of MS-based metabolomic datasets via compound interaction pair mapping. *Metabolomics*, **4**, 94–103.

Hahn-Hägerdal,B. *et al.* (2007) Towards industrial pentose-fermenting yeast strains. *Appl. Microbiol. Biotechnol.*, **74**, 937–953.

Haider,N. (2003) Checkmol/Matchmol. Available at http://merian.pch.univie.ac.at/~nhaider/cheminf/cmmm.html (last accessed date January, 2009).

Hattori,M. *et al.* (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.*, **125**, 11853–11865.

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

Herrgard,M.J. *et al.* (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.*, **26**, 1155–1160.

Jankowski,M. *et al.* (2008) Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.*, **95**, 1487–1499.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Klopman,G. *et al.* (1999) META. 4. Prediction of the metabolism of polycyclic aromatic hydrocarbons. *Theor. Chem. Acc. Theory Comput. Model. (Theor. Chim. Acta)*, **102**, 33–38.

Lai,C. *et al.* (2006) A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC Bioinformatics*, **7**, 235.

Li,C. *et al.* (2004) Computational discovery of biochemical routes to specialty chemicals. *Chem. Eng. Science*, **59**, 5051–5060.

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

O'Brien,P. and Herschlag,D. (1999) Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.*, **6**, R91–R105.

Oh,M. *et al.* (2007) Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. *J. Chem. Inf. Model.*, **47**, 1702–1712.

Pireddu,L. *et al.* (2006) The Path-A metabolic pathway prediction web server. *Nucleic Acids Res.*, **34**, W714–W719.

Thornalley,P. *et al.* (1984) The autoxidation of glyceraldehyde and other simple monosaccharides under physiological conditions catalysed by buffer ions. *Biochim. Biophys. Acta*, **797**, 276–287.

van Berlo,R. *et al.* (2009) Efficient calculation of compound similarity based on maximum common subgraphs and its application to prediction of gene transcript levels. *Technical Report ICT-2009-01*, Information & Communication Theory Group, Delft University of Technology. Available at http://ict.ewi.tudelft.nl/ (last accessed date August, 2009).

Weaver,R.H. and Lardy,H.A. (1961) Synthesis and some biochemical properties of phosphohydroxypyruvic aldehyde and of 3-phosphoglyceryl glutathione thiol ester. *J. Biol. Chem.*, **236**, 313–317.

Wessels,L.F.A. *et al.* (2005) A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, **21**, 3755–3762.

Wisselink,H. *et al.* (2009) Novel evolutionary engineering approach for accelerated utilization of glucose, xylose, and arabinose mixtures by engineered *Saccharomyces cerevisiae* strains. *Appl. Environ. Microbiol.*, **75**, 907–914.

Wu,M. *et al.* (1970) On the mechanism of the inhibition of growth by xylulose in *Chlorococcum echinozygotum*. *J. Phycol.*, **6**, 57–61.