JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Metabolic Pathway Predictions for Metabolomics: A Molecular Structure Matching Approach

Mai A. Hamdalla,[†,‡] Sanguthevar Rajasekaran,[†] David F. Grant,*[,§] and Ion I. Măndoiu*[,†]

[†]Computer Science and Engineering Department and [§]Pharmaceutical Sciences Department, University of Connecticut, Storrs, Connecticut 06269, United States

[‡]Computer Science Department, Helwan University, Cairo, Egypt

ABSTRACT: Metabolic pathways are composed of a series of chemical reactions occurring within a cell. In each pathway, enzymes catalyze the conversion of substrates into structurally similar products. Thus, structural similarity provides a potential means for mapping newly identified biochemical compounds to known metabolic pathways. In this paper, we present TrackSM, a cheminformatics tool designed to associate a chemical compound to a known metabolic pathway based on molecular structure matching techniques. Validation experiments show that TrackSM is capable of associating 93% of tested structures to their correct KEGG pathway class and 88% to their correct individual KEGG pathway. This suggests that TrackSM may be a valuable tool to aid in associating previously unknown small molecules to known biochemical pathways and improve our ability to link metabolomics, proteomic, and genomic data sets. TrackSM is freely available at http://metabolomics.pharm.uconn.edu/?q=Software.html.

INTRODUCTION

Metabolic pathways are characterized as a series of enzyme catalyzed reactions where the products of a previous reaction serve as substrates for a subsequent reaction. Understanding these pathways is essential to understanding the machinery of life.[1] The reconstruction of the metabolic network of an organism based on its genome sequence is a key challenge in systems biology.[2] Predicting which metabolic pathways are present in the organism based on the annotated genome of the organism is a possible strategy to address this issue.[3] Such metabolic pathways are selected from a reference database of known pathways. Other strategies use data mining techniques to correlate protein annotations to pathway templates so that organism-specific pathways can be derived. Some of the commonly used tools include PathComp,[4] Pathway Analyst,[1] Rahnuma,[5] Pathway Tools,[6] UM-BBD Pathway Prediction System,[7] and PathPred.[8]

Pathway prediction can involve predicting pathways that were previously known in other organisms or predicting novel pathways that have not been previously observed (pathway discovery).[3] The work presented here is focused on methodologies that do the former, predicting pathways from a curated reference database.

A number of databases containing biological pathway information are available. One of the most commonly used is the Kyoto Encyclopedia of Genes and Genomes (KEGG) database.[9] KEGG contains a collection of manually curated pathway maps representing molecular interaction and reaction networks. KEGG defines 11 metabolic pathway classes that are strongly associated with the biological function of compounds:[10] Carbohydrate Metabolism, Energy Metabolism, Lipid Metabolism, Nucleotide Metabolism, Amino Acid Metabolism, Metabolism of Other Amino Acids, Glycan Biosynthesis and Metabolism, Metabolism of Cofactors and Vitamins, Metabolism of Terpenoids and Polyketides, Biosynthesis of Other Secondary Metabolites, and Xenobiotics Biodegradation and Metabolism. Each of these classes contains several individual pathways. Some compounds serve as intermediates in multiple pathways and appear in multiple KEGG pathways.

Several recent advances in analytical and computational metabolomics techniques[11−14] will potentially improve our ability to identify the structures of previously unknown metabolites that do not belong to any known metabolic pathways. Placing these molecules in the context of known metabolic pathways might aid in understanding their biological function and will shed light on the presence of yet unidentified gene products that may be catalyzing relevant reactions.[15] Thus, the aim of the work described here is to develop and assess a model to predict pathway classes and individual pathways for a previously unknown query molecule.

Previous attempts to annotate metabolites with metabolic pathway information have been performed by Nobeli and Thornton.[16] Further investigations performed by Cai et al.[17] utilized functional group composition of compounds to represent small molecules. They proposed a Nearest Neighbor Algorithm to map small chemical molecules to a metabolic pathway class. After excluding all compounds that belonged to two or more metabolic pathway classes, a set of 2764 compounds from 11 classes of metabolic pathways obtained from KEGG were selected for that study. An overall successful prediction rate of 73.3% was observed. Since the authors

Received: August 24, 2014
Published: February 10, 2015

DOI: 10.1021/ci500517v
J. Chem. Inf. Model. 2015, 55, 709−718

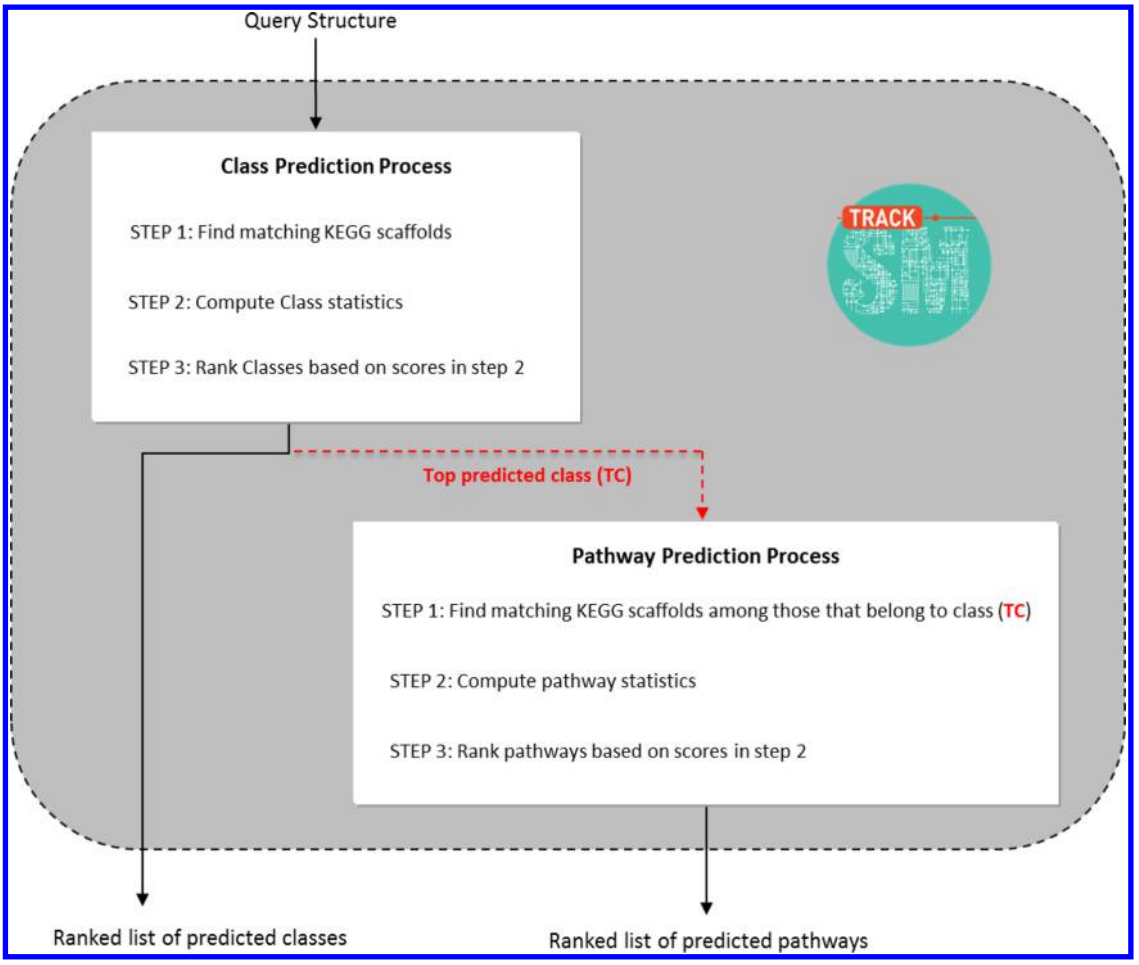

Figure 1. Schematic of TrackSM's predictive process.

focused on addressing the single-label classification problem, their methods could not be used to deal with "multifunction" compounds, i.e., compounds that belong to more than one pathway class. Macchiarulo et al.[15] used 32 physiochemical and topological descriptors to derive a quantitative structure activity relationship model and estimate the proximity of any small molecule to a given pathway class. When classifying 681 small molecules into 7 KEGG pathway classes using a random forest classifier,[18] they reported an average Matthews correlation coefficient of 0.73. They expanded their investigation to predict individual pathways to which these small molecules would belong. When classifying those metabolites into 52 individual KEGG pathways, they were able to predict the correct pathway for 31% of the molecules.

A multitarget model for predicting which of the 11 KEGG metabolic pathway classes a query compound may belong was proposed by Hu et al.[10] Their model was built using chemical–chemical interactions retrieved from STITCH,[19] a database containing known and predicted interactions of chemicals and proteins derived from experiments, literature, and other databases. In their model, an interaction unit consists of two chemicals, and their interaction weight (confidence score) representing the probability that the interaction occurs between the two chemicals being compared. Their overall success rate was approximately 78% using a 5-fold cross-validation test on a benchmark data set consisting of 3137 compounds.

Gao et al.[20] extended Hu et al.'s work by integrating interactions among chemicals and proteins in yeast. Their work included not only chemical–chemical interactions but also protein–protein interactions and chemical–protein interactions

Table 1. Distribution of 3190 KEGG Compounds among the 11 KEGG Metabolic Pathway Classes and 137 Individual Pathways

	pathway class name	pathway class code	number of individual pathways	number of compounds
1	carbohydrate metabolism	CM	15	575
2	energy metabolism	EM	7	193
3	lipid metabolism	LM	16	444
4	nucleotide metabolism	NM	2	137
5	amino acid metabolism	ACM	13	580
6	metabolism of other amino acids	MOAA	9	170
7	glycan biosynthesis and metabolism	GBM	5	48
8	metabolism of cofactors and vitamins	MCV	12	365
9	metabolism of terpenoids and polyketides	MTP	18	541
10	biosynthesis of other secondary metabolites	BOSM	20	555
11	xenobiotics biodegradation and metabolism	XBM	20	796
	total		**137**	**4404**

DOI: 10.1021/ci500517v
J. Chem. Inf. Model. 2015, 55, 709–718

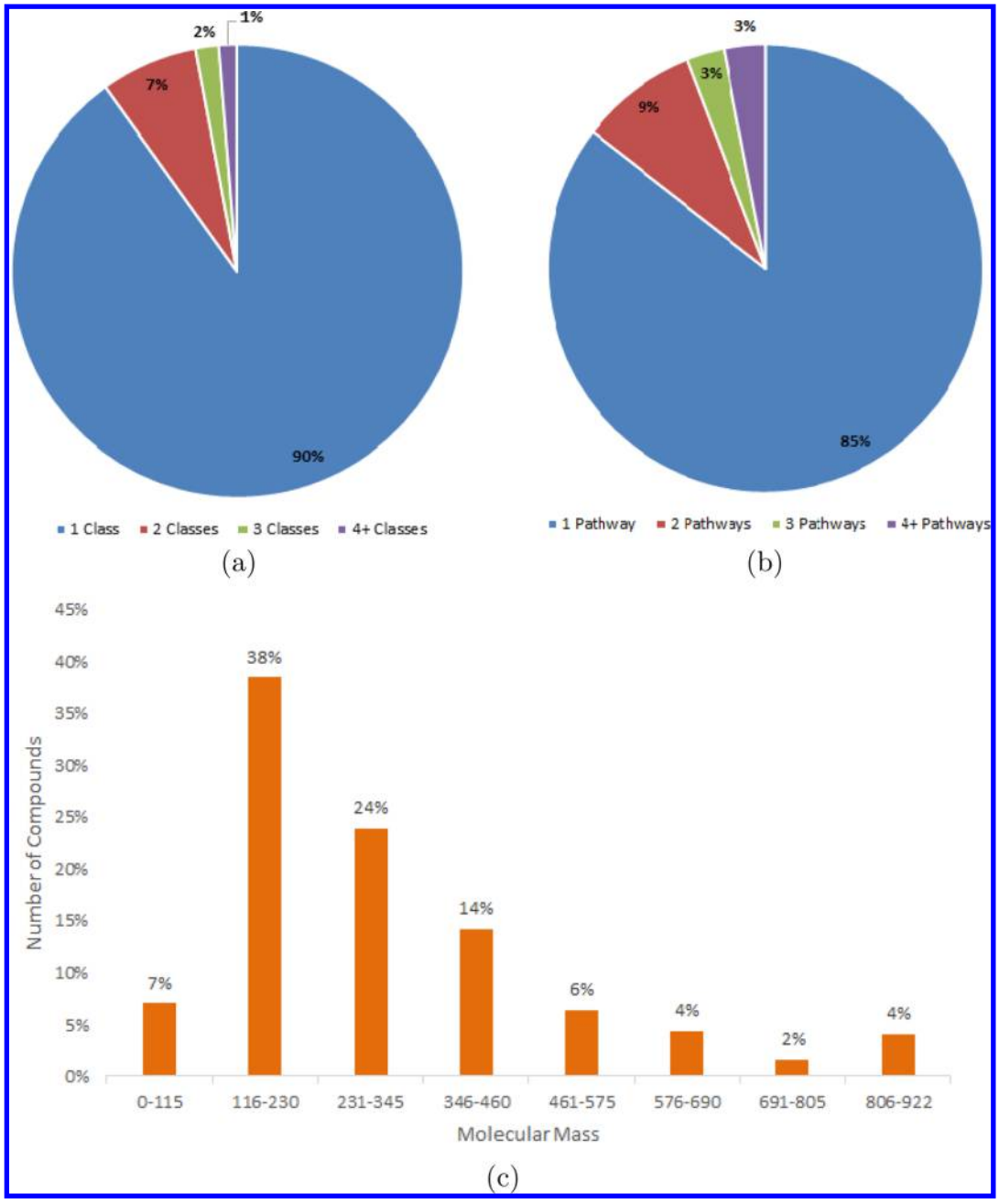

Figure 2. Distribution of 3,190 scaffolds based on (a) the number of classes they belong to and (b) the number of individual pathways they belong to. Panel (c) shows the mass distribution of 3,190 scaffolds into 8 mass bins ranging from 0 to 922 Da.

to predict metabolic pathways in which small molecules and enzymes participate. The protein−protein interaction data was retrieved from STRING.[21] They constructed a hybrid interaction network having small molecules and enzymes as nodes and edges between two nodes if and only if there was data showing that they can interact with one another. Results of a leave-one-out cross validation method showed that the first order prediction accuracy was 77.12% for the 3348 small molecules. This was not an improvement over the approach of Hu et al. described above. One of the major limitations in the approaches proposed in Hu et al. and Gao et al. is their dependency on interaction information. Hu et al. reported they were unable to process 1,229 compounds due to the lack of interaction information with other compounds within their data set.

In the work described here, we develop and assess TrackSM, a cheminformatics tool designed to predict the metabolic pathway class as well as the individual pathways to which previously unknown small molecules might be associated, based only on their molecular structures. TrackSM is guided by structural similarity information acquired from a set of compounds, hereafter referred to as scaffolds. In this context, the term scaffolds refers to common core features (i.e., substructures) that are shared among structurally related compounds and, hence, among compounds in biochemically related pathways. In other words, TrackSM represents pathways using the scaffolds they comprise. The method described was inspired by the fact that small molecules within a typical pathway tend to look similar because they are related to each other through stepwise chemical transformations.

DOI: 10.1021/ci500517v
J. Chem. Inf. Model. 2015, 55, 709−718

Table 2. SENS of Each Ranking Method when TrackSM Predicts 1, 2, or 3 Classes per Candidate Compound

	ranking method					
classes predicted	*SsScCo*	*ScCoSs*	*ScSsCo*	*SsCoSc*	*CoSsSc*	*CoScSs*
1	84.92%	84.73%	83.76%	64.70%	50.47%	50.41%
2	92.82%	92.76%	92.23%	81.38%	73.17%	73.13%
3	95.39%	95.27%	95.14%	89.78%	86.36%	86.36%

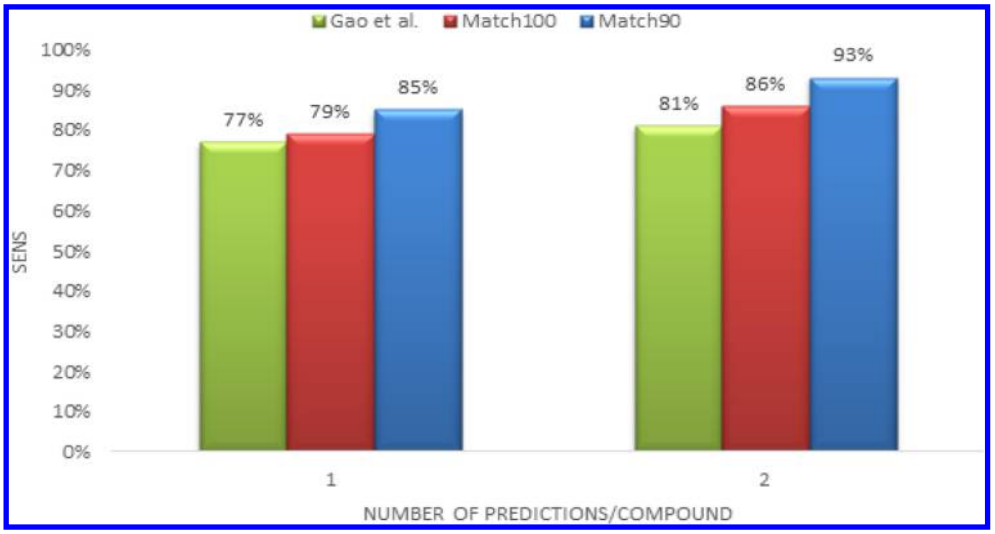

Figure 3. LOOCV prediction accuracy of the 1st and 2nd orders of predictions made by Gao et al.,[20] TrackSM with *Match100*, and TrackSM with *Match90* when predicting metabolic pathway classes.

METHODS AND MATERIALS

In TrackSM, a query compound is predicted to belong to a metabolic pathway based on how similar its chemical structure is to the structures associated with that pathway. As in our previous work on metabolite structure identification,[11] molecular similarity searches in TrackSM are carried using the SMSD Toolkit[22] for finding the maximum common subgraph (MCS) between small molecules. Although we have not directly evaluated similarity searches based on structural descriptors, the MCS search implemented in SMSD has the advantage of incorporating chemical knowledge (e.g., atom type match with bond sensitive information), thus resulting in improved structure matching sensitivity.[22]

Molecular Structure Similarity Score. In this study, we define two ways for matching molecular structures: *Match100* and *Match90*. In *Match100*, two molecular structures are considered a match if and only if the smaller structure is an exact substructure (atom and bond types) of the larger structure being compared. *Match90* considers two molecular structures as similar if at least 90% of the smaller structure's atoms match the larger structure being compared. Regardless of the matching method, if two molecular structures r and q were found to be a match, a similarity score is defined by

$$Sc = \frac{AC(r)}{AC(q)} \quad (1)$$

where $AC(r)$ represent the number of atoms in compound r and $AC(q)$ represent the number of atoms in compound q if r is a substructure of q. Clearly, a candidate molecule may match more than one scaffold structure, some as substructures and others as superstructures, resulting in several similarity scores computed for each candidate compound.

As previously mentioned, molecules within a typical pathway tend to have similar structures since they are related to each other through stepwise chemical transformations. Our hypothesis is that for a query compound c_q, the larger the number of compounds that are structurally similar to c_q within a given metabolic pathway the more likely that c_q is a member of that pathway. Also, if at least one of those scaffolds matches c_q as a substructure and another as a superstructure, then that might be more evidence of c_q belonging to that pathway. TrackSM identifies the biochemical pathway of a molecular compound in two steps. It first predicts a metabolic class to which the molecule is likely to belong, based on information from structurally similar scaffolds. Then it uses the predicted metabolic class with scaffold similarity information to predict an individual pathway to which the molecule is likely to belong. Figure 1 shows a general overview of the TrackSM prediction process.

Computational Prediction Algorithms. We propose an algorithm to predict the metabolic class and an individual metabolic pathway of a small compound based only on its molecular structure. First we will start by defining some notations followed by an explanation of the computational model behind TrackSM.

Let c_q be the molecular structure of a query compound, $S = \{s_1, s_2..., s_n\}$ be a set of n small compounds (scaffolds), $M = \{M_1, M_2, ..., M_l\}$ be a set of metabolic pathway classes, and $P = \{p_1, p_2, ..., p_m\}$ be the set of individual pathways. Let s_xM_y indicate that the scaffold s_x belongs to the metabolic pathway class M_y and let $s_x \rightarrow P_z$ indicate that s_x belongs to the individual metabolic pathway P_z. Let $CL(c_q)$ represent the list of candidate pathway classes to which c_q is predicted to be associated with. Similarly, let $PL(c_q)$ represent the list of candidate individual pathways to which c_q is predicted to be associated with.

DOI: 10.1021/ci500517v
J. Chem. Inf. Model. 2015, 55, 709−718

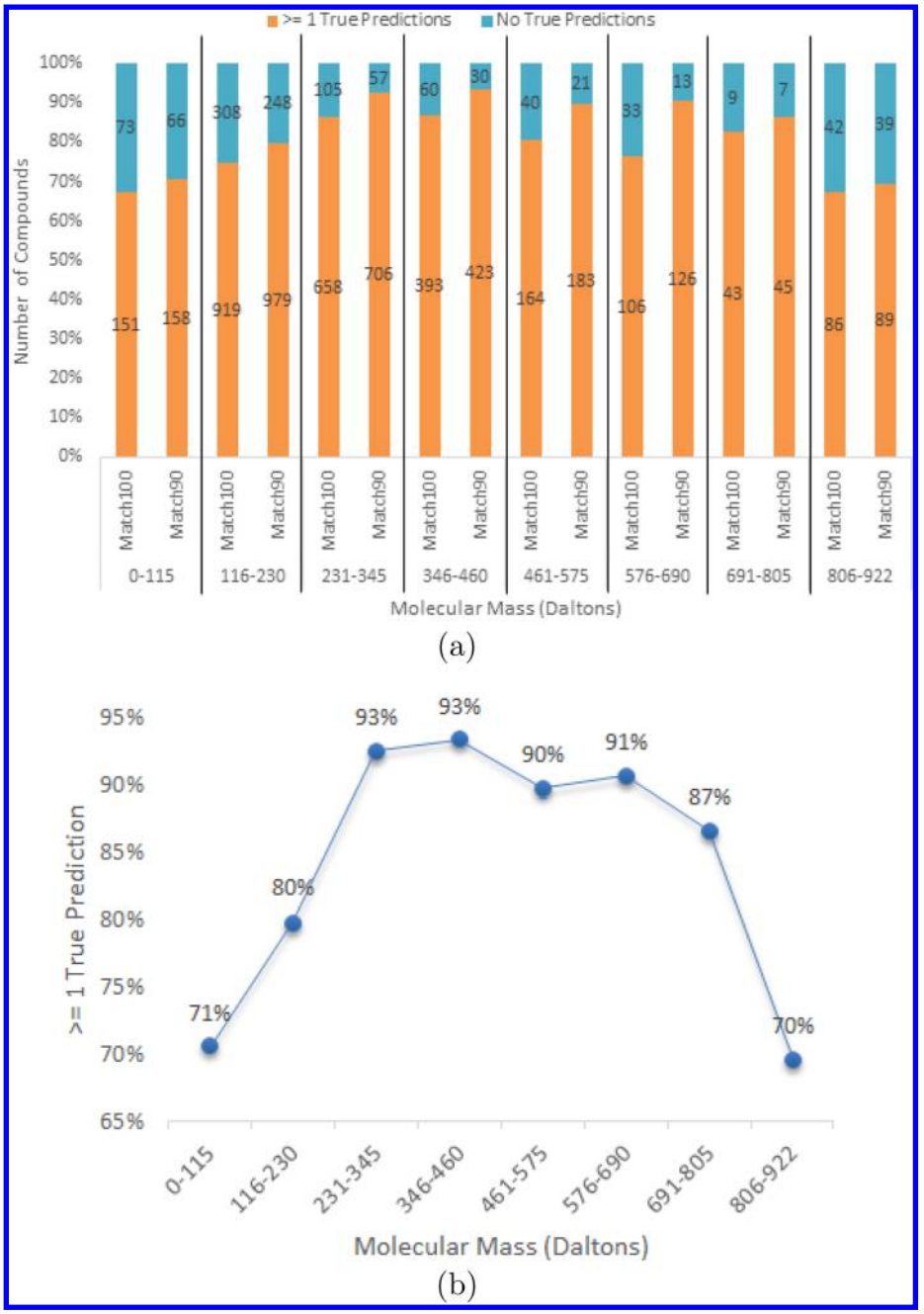

Figure 4. (a) Breakdown of the 1st order class predictions made by *Match100* versus those made by *Match90* for 3,190 compounds based on molecular mass from a set of LOOCV experiments. (b) SENS in each bin when Match90 was applied.

Pathway Classes Prediction Method. For a query compound c_q, TrackSM populates a vector V to represent the confidence that metabolic class M_i is the pathway class to which c_q belongs. Vector $V(c_q, M_i) = [Ss, Sc, Co]$, where Ss is a binary value representing the existence of at least one substructure compound *and* at least one superstructure compound that belongs to M_i, i.e.,

$$Ss(c_q, M_i) = \begin{cases} 1, & \text{if } \exists\ s_x \in S_b \text{ and } s_y \in S_p;\ s_x \rightarrow M_i,\ s_y \rightarrow M_i \\ 0, & \text{otherwise} \end{cases}$$

Such that S_b represents the set of scaffolds that are substructures of c_q and S_p represents the set of scaffolds found to be a superstructure of c_q. Hence, let $\overline{S}$ denote the set of scaffolds that structurally match c_q such that $\overline{S} = S_b \cup S_p$. Let Sc represent the highest *SimScore* as defined earlier by eq 1 found between c_q and all the matching scaffolds in $\overline{S}$ that belong to M_i; $Sc(c_q, M_i) = \max_{s_j \in S, s_j \rightarrow M_i} SimScore(c_q, s_j)$. Co is defined as the number of scaffolds in $\overline{S}$ that belong to pathway class M_i; $Co(c_q, M_i) = count(s_j \in \overline{S}, s_j \rightarrow M_i)$. Finally, all pathway classes in $CL(c_q)$ are ranked based on (Ss, Co, Sc) values and the class with the highest scores is predicted to be PC, the class to which c_q is associated.

Individual Pathways Prediction Method. In the second step, TrackSM predicts one or more individual pathways to which the query compound might belong to. List $PL(c_q)$ is populated and ranked via a method very similar to that used to populate $CL(c_q)$ with the exception of referencing individual

DOI: 10.1021/ci500517v
J. Chem. Inf. Model. 2015, 55, 709−718

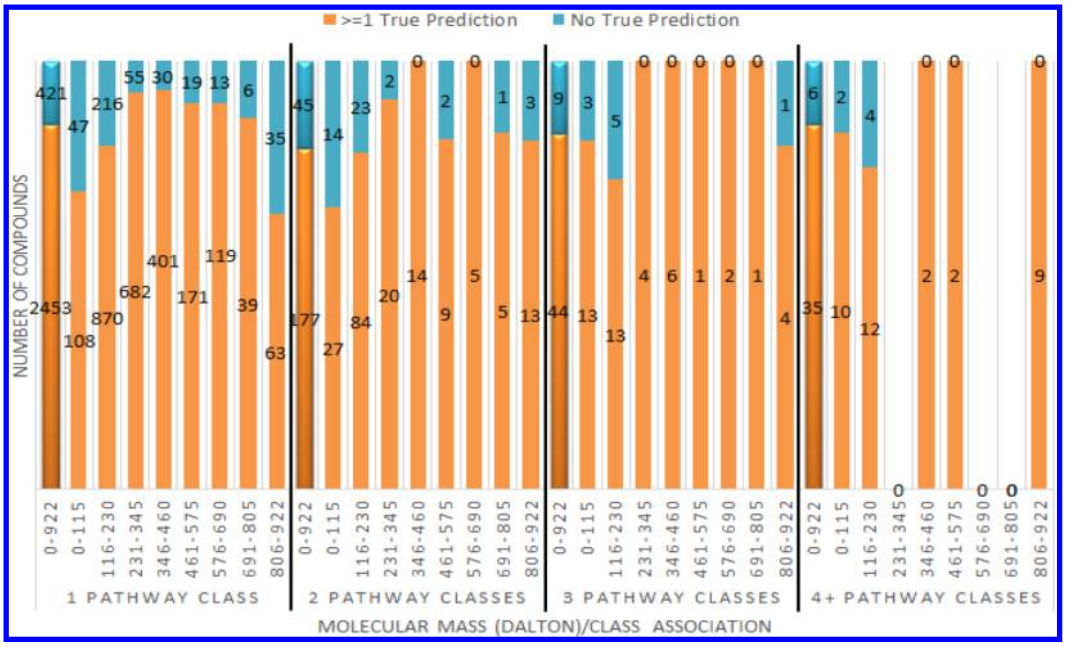

Figure 5. Accuracy of mass bins per number of class associations for TrackSM when using *Match90* to predict metabolic pathway classes.

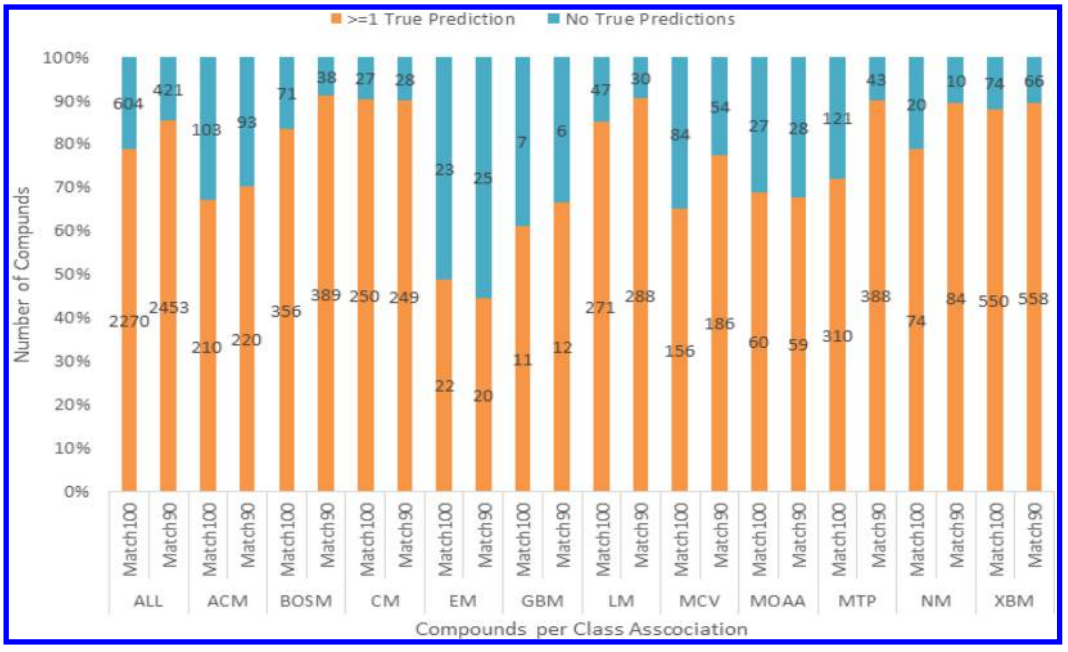

Figure 6. Distribution of class predictions when using *Match100* versus *Match90* based on the query compound's class association.

pathways instead of pathway classes. Similarly, for each candidate individual pathway P_r, associated with at least one compound in $\overline{S}$, a vector $V(c_q,P_r) = [Ss,Sc,Co]$ is populated. Vector $V(c_q,P_r)$ represents the confidence that pathway P_r is the predicted pathway to which c_q belongs. *Ss* is a binary value representing the existence of at least one substructure compound *and* at least one superstructure compound that belong to P_r, i.e.,

$$Ss(c_q, P_r) = \begin{cases} 1, & \text{if } \exists\ s_x \in S_b \text{ and } s_y \in S_p;\ s_x \rightarrow P_r,\ s_y \rightarrow P_r \\ 0, & \text{otherwise} \end{cases}$$

Let *Sc* represent the highest *SimScore* (defined by eq 1) found between c_q and all the matching compounds in $\overline{S}$ that belong to P_r; $Sc(c_q,P_r) = \max_{s_j \in \overline{S}, s_j \rightarrow P_r} SimScore(c_q,s_j)$. Finally, *Co* is defined as the number of scaffolds in $\overline{S}$ that belong to pathway class P_r; $Co(c_q,P_r) = count(s_j \in \overline{S} s_j \rightarrow P_r)$.

Specific to predicting individual pathways, we have developed an additional method referred to as *Match90ClassBased*. In this method, TrackSM uses the predicted pathway class *PC* in the first step to further guide its search for individual pathway associations for c_q. Hence, $PL(c_q)$ is only populated with individual pathways that have associations with scaffolds that structurally match c_q *and* belong to the predicted pathway class *PC*. Hence, any scaffold in $\overline{S}$ must belong to the predicted class *PC*. All the calculations following this step are similar to that of the previously explained method.

DOI: 10.1021/ci500517v
J. Chem. Inf. Model. 2015, 55, 709−718

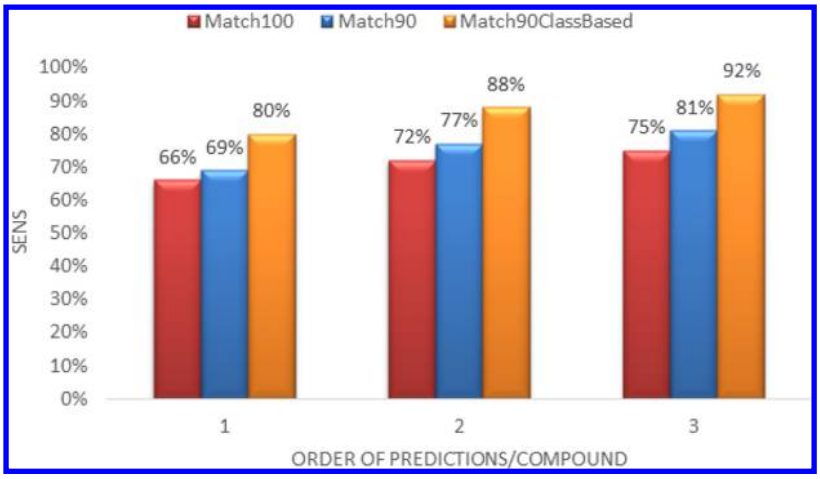

Figure 7. Prediction accuracy of the first, second, and third orders of predictions made by TrackSM with *Match100*, *Match90*, and *Match90ClassBased* when predicting individual metabolic pathways.

Data set Pathway information concerning 3190 small molecules of the data set used by Gao et al.,[20] as well as their molecular structures, were downloaded (January 2013) from the KEGG database. The distribution of those compounds among the pathway classes and the number of individual pathways associated with each class are shown in Table 1. The data show that some compounds are associated with more than one pathway class since the total number of compounds belonging to all classes (4404) is greater than the actual number of compounds (3190). Figure 2a shows that 90% of the 3190 scaffolds used in this study are associated with only one pathway class, while 7% are associated with two classes and 3% are associated with 3 or more pathway classes. Figure 2b demonstrates the distribution of scaffolds based on their association to individual pathways. Of the 3,190 scaffolds, 85% are associated with one individual pathway. This means that 5% (90%−85%) of the compounds associated with one pathway class belong to more than one individual pathway within that given class. Nine percent are associated with 2 individual pathways, and 6% are associated with 3 or more individual pathways. The mass distribution of the molecules in the scaffolds list used in this work (Figure 2) shows that the majority of the molecules (76%) fall in the mass range of 116−460 Da.

Leave-One-Out Cross Validation Test. In this study, a set of leave-one-out cross validation (LOOCV) experiments were carried out on the N = 3,190 small molecules in our reference scaffolds database as a method for evaluating the accuracy of TrackSM. N experiments were performed, and for each experiment, $N-1$ compounds were used as scaffolds and the remaining compound was treated as the query compound. This allowed the use of all but one scaffold in the prediction process.

Accuracy Measures. To evaluate the performance of TrackSM, we used two measures, sensitivity (SENS) and positive predictive value (PPV). *SENS* is a measure representing the percentage of query compounds with at least one correctly predicted pathway reproduced by TrackSM and is computed as

$$SENS = \frac{TP}{TP + FN} \tag{2}$$

where TP represents the number of compounds with at least one correct prediction and FN represents the number of compounds with all false predictions. *PPV* is a measure representing the percentage of correct pathway prediction assignments made by TrackSM and is defined by

$$PPV = \frac{TP}{TP + FP} \tag{3}$$

RESULTS AND DISCUSSION

We formalized six possible ways for ranking the candidate classes in $CL(c_q)$ and candidate pathways in $PL(c_q)$ referred to as *SsScCo*, *ScSsCo*, *ScCoSs*, *CoScSs*, *CoSsSc*, and *SsCoSc*. *SsScCo* indicates sorting the candidate pathway classes in CL (c_q) by the *Ss* value (in descending order) then breaking ties with the pathway with the highest *Sc* followed by the highest *Co*. Table 2 shows the sensitivity acquired when a set of LOOCV experiments predicting metabolic pathway classes were carried out using the 3,190 KEGG compounds. The results from *SsScCo*, *ScCoSs*, and *ScSsCo* seemed comparable and better than those obtained by *SsCoSc*, *CoSsSc*, and *CoScSs*. We carried out an ANOVA to check for statistical significance between the top 3 ranking methods. ANOVA results indicated no statistical significance ($P > 0.05$). However, *SsScCo* accuracy was consistently higher than the other methods and thus was selected as the ranking method for TrackSM.

Metabolic Pathway Class Predictions. In this study we evaluated the predictive method by a set of LOOCV experiments using a data set of 3,190 KEGG compounds. The 1st and 2nd order of predictions made by Gao et al.[20] as well as those of TrackSM using both *Match100* and *Match90* with the *SsScCo* ranking method are shown in Figure 3. TrackSM was able to predict at least one correct pathway class for 85% of the compounds using *Match90* versus 79% when using *Match100*. Both methods reflect an improvement over the results reported by Gao et al. (77%).

Actually, using TrackSM *Match90* to predict only one class per query compound had a 4% improvement in SENS over Gao et al.'s method when they predicted 2 classes per compound. This also indicates that TrackSM has a better PPV than that of Gao et al. When TrackSM using *Match90* made two class predictions per candidate compound, 93% of the 3,190 compounds had at least one correct class prediction.

In order to determine whether predictions made by TrackSM were equally accurate among compounds having different masses we distributed the 3,190 compounds into 8 bins according to

DOI: 10.1021/ci500517v
J. Chem. Inf. Model. 2015, 55, 709−718

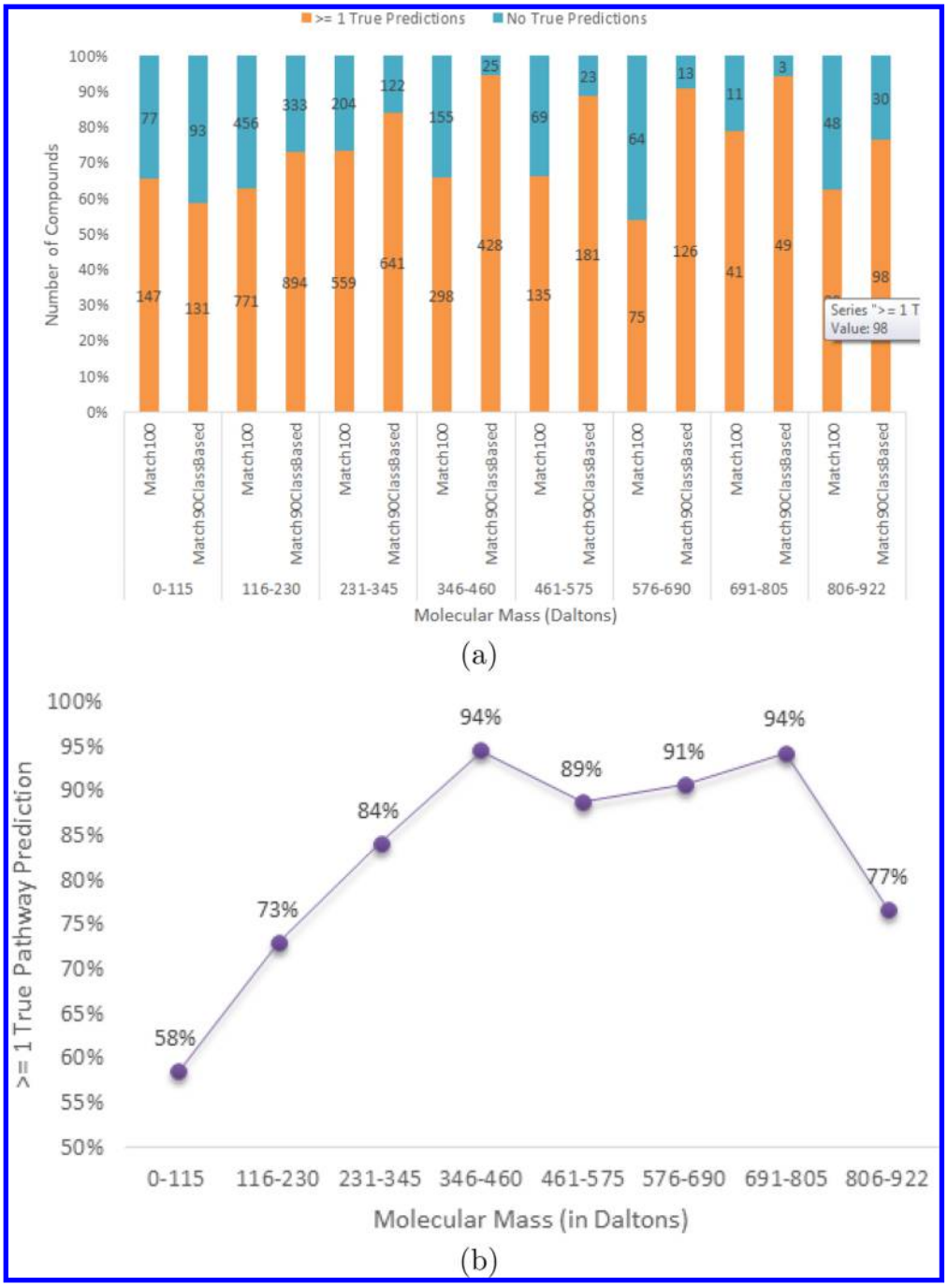

Figure 8. (a) Percentage of compounds with at least one correct individual pathway prediction when compounds are distributed by mass among 8 mass bins. (b) SENS in each bin when *Match90ClassBased* was applied to predict one individual pathway per query compound.

compound masses. Results from both prediction methods (*Match100* and *Match90*) were compared in each bin. Figure 4a displays the first order of class predictions made by TrackSM using *Match100* versus *Match90*. The results suggest that *Match90* outperforms *Match100* across all mass bins. Figure 4b plots the SENS in each bin when *Match90* was applied. The plot shows that TrackSM is capable of predicting the metabolic class of a given compound in the mass range 231−460 Da with 93% accuracy. It also shows that bins 3 through 7 acquire an average SENS of 90%, while the average SENS in bin 1 and bin 8 is approximately 70%. We suggest that predictions at both ends of the mass range are poorer because as compounds become very large or very small, there is a higher chance for them to match with many scaffolds as substructures only or superstructures only, respectively. Thus, nonspecific matches cause a decrease in sensitivity.

Figure 5 shows the distribution of the 1st order of *Match90* class predictions based on a compound's molecular mass and the number of pathway classes it is associated with. The first bar in each subsection shows the overall predictions per number of class associations. *Match90* can predict at least one class to which a compound might belong with a SENS of 85%, 80%, 83%, and 85% for compounds associated with one, two, three, and four or more classes, respectively. This suggests that the number of class associations for any given compound does not significantly affect the prediction quality of TrackSM. It is clear from Figure 5 that the predictions for compounds that belong

DOI: 10.1021/ci500517v
J. Chem. Inf. Model. 2015, 55, 709−718

to more than one metabolic class do not follow the same distribution (based on mass) as that of those associated with only one class. However, it is difficult to draw firm conclusions since only 10% of the compounds used in this analysis are associated with more than one class.

To further analyze our results, we next explored TrackSM prediction accuracy for each metabolic class. Figure 6 shows the distribution of compounds among the 11 KEGG metabolic classes based on the 1st order prediction produced by *Match100* versus *Match90*. In this analysis, only 2,874 of the compounds were included as they are associated with only one class. *Match100* had a 4% improvement over *Match90* when associating compounds to class EM. Both methods performed equivalently when associating compounds to classes CM, MOAA, and XBM. *Match90* outperformed *Match100* in the other 7 classes with a highest improvement of 18% in class MTP. *Match90* correctly associated 90% of the compounds belonging to six metabolic classes (BOSM, CM, LM, MTP, NM, and XBM). It was also noted that *Match90* could only correctly associate 44% of the compounds belonging to class EM. This is likely due to the small size of class EM, with only 45 scaffold associations.

Individual Metabolic Pathway Predictions. We next evaluated TrackSM to predict individual pathways to which a candidate compound might belong. We show results from applying *Match100* and *Match90* as well as a method exclusive to pathway predictions referred to as *Match90ClassBased* (described in Materials and Methods). Figure 7 shows the SENS of the 1st, 2nd, and 3rd orders of prediction when using *Match100*, *Match90*, and *Match90ClassBased*. *Match90ClassBased* outperformed the other two methods. Specifically, with the 1st order of individual pathway prediction, *Match90ClassBased* had 80% accuracy, while *Match100* had only 66% and *Match90* had 69%. When making 2 predictions per query compound, *Match90ClassBased* was able to predict at least one individual pathway for 88% of the compounds. An inherent limitation of our approach is scaffold inclusion; as the number of scaffolds included increases, the accuracy of TrackSM would likely increase. However, there is perhaps an upper limit to the scaffold number, above which the results are not significantly improved. This, in fact, was recently shown in our previous work when predicting whether an unknown compound was biological or synthetic using an approach similar to the one described here.[11] Although there were improvements in model sensitivity and specificity using a much larger scaffold list (double in size), the data set comparison results were very similar. These results suggest that additional scaffolds will not greatly improve our representation of biochemical pathway structure space.

Finally, we distributed the 3,190 compounds into 8 bin masses and compared individual pathway results from each prediction method (*Match100* and *Match90ClassBased*) for each bin. Figure 8a displays the first order of individual pathway predictions made by TrackSM using *Match100* versus *Match90ClassBased*. It is obvious that *Match90ClassBased* outperforms *Match100* across all bins except bin 1. Bin 1 consists of very small compounds with masses up to 115 Da. Hence, allowing any flexibility in the structure matching process (*Match90ClassBased*) is very likely to identify significantly different structures as a match to the unknown structure resulting in poorer pathway predictions. Figure 8b plots the SENS in each bin when *Match90ClassBased* was applied to predict one individual pathway per query compound. The plot shows that TrackSM is capable of predicting the individual pathway of a given compound in the mass range 346−460 Da and 691−805 Da with 94% accuracy. It also shows that bins 4 through 7 acquire an average SENS of 92%. Similar to pathway class predictions, individual pathway predictions for compounds on both ends of the mass spectrum are of noticeably lower sensitivity than the rest of the bins.

CONCLUSIONS

In this paper we developed and evaluated TrackSM, a bioinformatics tool designed to predict previously known metabolic pathway classes as well as the individual pathways to which small, previously unknown molecules might be associated, based only on their molecular structures. TrackSM can place molecules in the context of metabolic pathways since it can link newly identified biochemicals to matched substructure and superstructure scaffolds for which metabolic pathways are known. Validation experiments show that TrackSM is capable of associating 93% of structures to their correct pathway classes as defined by KEGG and 88% to the correct individual KEGG pathway. These results suggest that TrackSM may be a valuable tool to aid in recognizing the biochemical functions of small molecules and of broad use in annotating metabolomics data for systems biology applications.

AUTHOR INFORMATION

Corresponding Authors

*(D.F.G.) E-mail: david.grant@uconn.edu.

*(I.I.M.) E-mail: ion@engr.uconn.edu.

Author Contributions

M.A.H. was responsible for designing the algorithm, software development, and manuscript preparation. She was also responsible for testing and benchmarking BioSM. I.I.M., D.F.G., and S.R. were involved in the overall supervision of the project, manuscript preparation, intellectual input, and guidance.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was funded in part by NIH Grant 1R01GM087714, the Agriculture and Food Research Initiative Competitive Grant No. 2011-67016-30331 from the USDA National Institute of Food and Agriculture, Award IIS-0916948 from NSF, and the Booth Engineering Center for Advance Technology (BECAT) at the University of Connecticut.

REFERENCES

(1) Pireddu, L.; Szafron, D.; Lu, P.; Greiner, R. *Nucleic Acids Res.* **2006**, *34*, W714−9.

(2) Chen, N.; del Val, I. J.; Kyriakopoulos, S.; Polizzi, K. M.; Kontoravdi, C. *Curr. Opin. Biotechnol.* **2012**, *23*, 77−82.

(3) Dale, J. M.; Popescu, L.; Karp, P. D. *BMC Bioinform.* **2010**, *11*, 15.

(4) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. *Nucleic Acids Res.* **2006**, *34*, D354−7.

(5) Mithani, A.; Preston, G. M.; Hein, J. *Bioinformatics (Oxford, England)* **2009**, *25*, 1831−2.

(6) Karp, P. D.; Paley, S. M.; Krummenacker, M.; Latendresse, M.; Dale, J. M.; Lee, T. J.; Kaipa, P.; Gilham, F.; Spaulding, A.; Popescu, L.; Altman, T.; Paulsen, I.; Keseler, I. M.; Caspi, R. *Briefings Bioinf.* **2010**, *11*, 40−79.

DOI: 10.1021/ci500517v
J. Chem. Inf. Model. 2015, 55, 709−718

(7) Gao, J.; Ellis, L. B. M.; Wackett, L. P. *Nucleic Acids Res.* **2010**, *38*, D488−91.
(8) Moriya, Y.; Shigemizu, D.; Hattori, M.; Tokimatsu, T.; Kotera, M.; Goto, S.; Kanehisa, M. *Nucleic Acids Res.* **2010**, *38*, W138−43.
(9) Kanehisa, M.; Goto, S.; Kawashima, S.; Nakaya, A. *Nucleic Acids Res.* **2002**, *30*, 42−46.
(10) Hu, L.-L.; Chen, C.; Huang, T.; Cai, Y.-D.; Chou, K.-C. *PloS One* **2011**, *6*, e29491.
(11) Hamdalla, M.; Mandoiu, I.; Hill, D.; Rajasekaran, S.; Grant, D. *J. Chem. Inf. Model.* **2013**, *53*, 601−612.
(12) Menikarachchi, L. C.; Cawley, S.; Hill, D. W.; Hall, L. M.; Hall, L.; Lai, S.; Wilder, J.; Grant, D. F. *Anal. Chem.* **2012**, *84*, 9388−9394.
(13) Menikarachchi, L. C.; Hamdalla, M. A.; Hill, D. W.; Grant, D. F. *Comput. Struct. Biotechnol. J.* **2013**, *5*, e201302005.
(14) Menikarachchi, L.; Hill, D.; Hamdalla, M.; Mandoiu, I.; Grant, D. *J. Chem. Inf. Model.* **2013**, *53*, 2483−2492.
(15) Macchiarulo, A.; Thornton, J. M.; Nobeli, I. *J. Chem. Inf. Model.* **2009**, *49*, 2272−2289.
(16) Nobeli, I.; Thornton, J. M. *BioEssays* **2006**, *28*, 534−45.
(17) Cai, Y.-D.; Qian, Z.; Lu, L.; Feng, K.-Y.; Meng, X.; Niu, B.; Zhao, G.-D.; Lu, W.-C. *Mol. Diversity* **2008**, *12*, 131−7.
(18) Breiman, L. *Machine Learning*, 45th ed.; Kluwer Academic Publishers: 2001; Vol. *45*, pp 5−32.
(19) STITCH. http://stitch.embl.de/.
(20) Gao, Y.-F.; Chen, L.; Cai, Y.-D.; Feng, K.-Y.; Huang, T.; Jiang, Y. *PloS One* **2012**, *7*, e45944.
(21) STRING. http://string.embl.de/.
(22) Rahman, S. A.; Bashton, M.; Holliday, G. L.; Schrader, R.; Thornton, J. M. *J. Cheminform.* **2009**, DOI: 10.1186/1758-2946-1-12.

DOI: 10.1021/ci500517v
J. Chem. Inf. Model. 2015, 55, 709−718