

BioSM: Metabolomics Tool for Identifying Endogenous Mammalian Biochemical Structures in Chemical Structure Space

Mai A. Hamdalla,[†] Ion I. Mandoiu,[†] Dennis W. Hill,[‡] Sanguthevar Rajasekaran,[†] and David F. Grant^{*,‡}

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

Supporting Information

ABSTRACT: The structural identification of unknown biochemical compounds in complex biofluids continues to be a major challenge in metabolomics research. Using LC/MS, there are currently two major options for solving this problem: searching small biochemical databases, which often do not contain the unknown of interest or searching large chemical databases which include large numbers of nonbiochemical compounds. Searching larger chemical databases (larger chemical space) increases the odds of identifying an unknown biochemical compound, but only if nonbiochemical structures can be eliminated from consideration. In this paper we present BioSM; a cheminformatics tool that uses known endogenous mammalian biochemical compounds (as scaffolds) and graph matching methods to identify endogenous mammalian biochemical structures in chemical structure space. The results of a comprehensive set of empirical experiments suggest that BioSM identifies endogenous mammalian biochemical structures with high accuracy. In a leave-one-out cross validation experiment, BioSM correctly predicted 95% of 1388 Kyoto Encyclopedia of Genes and Genomes (KEGG) compounds as endogenous mammalian biochemicals using 1565 scaffolds. Analysis of two additional biological data sets containing 2330 human metabolites (HMDB) and 2416 plant secondary metabolites (KEGG) resulted in biochemical annotations of 89% and 72% of the compounds, respectively. When a data set of 3895 drugs (DrugBank and USAN) was tested, 48% of these structures were predicted to be biochemical. However, when a set of synthetic chemical compounds (Chembridge and Chemsynthesis databases) were examined, only 29% of the 458 207 structures were predicted to be biochemical. Moreover, BioSM predicted that 34% of 883 199 randomly selected compounds from PubChem were biochemical. We then expanded the scaffold list to 3927 biochemical compounds and reevaluated the above data sets to determine whether scaffold number influenced model performance. Although there were significant improvements in model sensitivity and specificity using the larger scaffold list, the data set comparison results were very similar. These results suggest that additional biochemical scaffolds will not further improve our representation of biochemical structure space and that the model is reasonably robust. BioSM provides a qualitative (yes/no) and quantitative (ranking) method for endogenous mammalian biochemical annotation of chemical space and, thus, will be useful in the identification of unknown biochemical structures in metabolomics. BioSM is freely available at <http://metabolomics.pharm.uconn.edu>.



■ INTRODUCTION

Metabolomics is a rapidly evolving discipline involving the study of small molecules or metabolites that characterize metabolic pathways of biological systems. It combines strategies to identify and quantify cellular metabolites using analytical techniques such as mass spectrometry (MS),¹ with the application of computational methods for information extraction and data interpretation.² Metabolomics has been labeled as one of the new “omics”, joining genomics, transcriptomics, and proteomics.³ It is of particular interest as endogenous metabolites represent the phenotype resulting from gene expression.⁴ Hence, changes in metabolic profiles can be used in a variety of applications, such as drug development,^{5–7} agriculture,^{8,9} and toxicology studies.¹⁰

MS coupled with chromatographic separation techniques such as liquid or gas chromatography and nuclear magnetic resonance (NMR) spectroscopy¹¹ are currently the major techniques used to simultaneously analyze large numbers of metabolites.² Regardless of the analytical method, a major challenge in metabolomics is the interpretation of the vast amount of data produced by these high-throughput techni-

ques.¹² The most common approach entails matching experimentally determined features, such as a mass spectrum or retention index, with computationally simulated features for a set of candidate compounds downloaded from a general chemical structure database.¹³ Various online chemical structure databases such as PubChem,¹⁴ Kyoto Encyclopedia of Genes and Genomes (KEGG),¹⁵ Human Metabolome Database (HMDB),¹⁶ and HumanCyc¹⁷ provide the fundamental support for molecular identification. The relative advantages or disadvantages of utilizing chemical structure databases vary depending on the size of the database. Small databases often will not contain the candidate compound of interest. On the other hand, searching large databases such as PubChem, often results in a large number of false positives, making identification of the “unknown” extremely difficult. Hence, cheminformatics methods are needed to more efficiently search large chemical databases in order to identify unknown endogenous biochemical compounds. Ideally, these

Received: October 24, 2012

methods would allow discrimination between candidate structures that are synthetic and candidate structures that are biochemical.^{18,19}

Nobeli et al.,²⁰ using two-dimensional (2D) molecular structures and cheminformatics tools, reported the first attempt to solve this problem. They visually examined the 2D molecular structures of 745 *E. coli* metabolites and manually derived a library of 57 structural fragments commonly found in those metabolites to reveal the main constituents of metabolites and to assist in the classification of the metabolome into biochemically relevant classes. Preliminary efforts correlating similarities between metabolites and protein structures, as well as with metabolic pathways were reported. In related work, Gupta and Aires-de-Sousa²¹ defined chemical space of endogenous biochemicals using the KEGG/LIGAND database. Any compound in KEGG that was involved in a metabolic reaction was included in the study. These included metabolites from different species as well as xenobiotics. The chemical space of nonmetabolites was represented by a random set of commercially available compounds from the ZINC²² chemical database. They compared both chemical spaces based on 2D and 3D structures and descriptors of global properties. They found that overlap between metabolites and nonmetabolites was smallest in the space defined by the global descriptors and suggested that the most discriminative features were the number of OH groups, the presence of aromatic systems, and molecular weight. Using a random forest (RF)²³ classifier and global molecular descriptors, they were able to correctly annotate 95% of the 1811 KEGG compounds used for training the model. A RF is a collection of unpruned classification trees created by using bootstrap samples of the training data and random subsets of variables to define the best split at each node.

Extending Gupta and Aires-de-Sousa's work, Peironcelly et al.²⁴ used 6954 molecular structures in HMDB to represent chemical space occupied by endogenous human metabolites and an updated collection of compounds from ZINC as nonbiological structures. Both data sets were clustered independently and 532 molecules (cluster centers) from each data set, selected to represent each cluster, were used for building the classification model. The remaining (6422) molecules were used for training the model. They showed that using MDL public keys²⁵ and RF resulted in the best accuracy for their classifier. The authors reported that 96% of 457 HMDB compounds not used for training the model, 54% of 6532 DrugBank compounds, and 22% of 6312 compounds from ChEMBL²⁶ were classified as endogenous metabolites.

Both Gupta²¹ and Peironcelly²⁴ employed fingerprints for classification. Molecular fingerprints represent the structure of a molecule as a list of binary values (0 or 1) that indicate the presence or absence of structural features in the molecule.²⁷ A structural feature may include properties (such as molecular weight), the presence/absence of an element, an unusual or important electronic configuration (such as triple-bonded nitrogen), rings and ring systems, and functional groups. An alternative approach is based on viewing a molecule as a graph and using graph-matching algorithms to find common substructures. Previous work²⁰ suggests that matching common substructures may describe structural similarity more accurately than fingerprint-based methods. Although this has been suggested, it has not been explored due to concerns related to computational efficiency. In addition, this approach of matching common substructures is consistent with how

endogenous biochemicals are produced enzymatically in vivo, i.e., from precursors with similar and/or overlapping structures.

Here we present BioSM, a molecular classifier that can identify endogenous mammalian biochemical structures contained within chemical structure space. BioSM uses the structures of known endogenous mammalian biochemical compounds as scaffolds to aid in the classification process, as opposed to other works that use fragments of known structures. The graph-based method implemented within BioSM can also be expanded to predict metabolic pathways since it links a set of annotated scaffold structures to each candidate structure.

In our empirical evaluation of BioSM, we initially focused on a curated set of endogenous human biochemicals obtained from the KEGG/LIGAND database to represent the scaffolds list. The chemical space of nonbiological compounds was approximated by a randomly selected set of compounds from the Chembridge²⁸ and Chemsynthesis²⁹ chemical databases. Since structurally similar molecules tend to have similar properties,³⁰ we use a graph matching algorithm to identify compounds that are structurally similar to those in our scaffolds list. Our classification method is based on a novel scoring scheme that combines all matches of scaffolds to substructures of a candidate compound as well as matches of the candidate compound's structure to substructures of the scaffolds. We were also interested in determining whether increasing the number of scaffolds (i.e., increasing our representation of biochemical structure space) would improve model sensitivity and specificity. Therefore, we supplemented our initial KEGG scaffolds list with 2362 curated compounds from HMDB and HumanCyc and repeated the assessment experiments.

METHODS

Molecular Structure Matching. Marvin³¹ chemical structure processing software was used to generate canonical SMILES (Simplified Molecular-Input Line-Entry System)³² from structure data files (.sdf) for all compounds described in this work. The Small Molecule Subgraph Detector (SMSD) Toolkit³³ was used to carry out molecule similarity searches. SMSD is a Java based software library for finding the maximum common subgraph between small molecules using atom type matches and bond sensitivity information. In our work, two molecular structures match if and only if the smaller structure was an exact substructure (atom and bond types) of the larger structure being compared. A similarity score between two molecular structures was defined by

$$\text{similarity score} = \frac{N_{\text{SBS}}}{N_{\text{SPR}}} \quad (1)$$

where N_{SBS} represents the total number of atoms in the substructure and N_{SPR} represents the total number of atoms in the superstructure. Clearly, a candidate molecule may match more than one scaffold structure, resulting in several similarity scores computed for each candidate compound. Initially, the highest similarity score was selected to represent the degree of biochemical similarity between scaffold structures and the candidate compound's structure. However, we observed that multiple scaffolds could match different substructures of the candidate, significantly strengthening the evidence that the candidate compound is an endogenous mammalian biochemical. Thus, we developed a "union scaffold structure" approach that incorporates all scaffolds matching a candidate compound's structure and serves to reduce bias that might exist due to

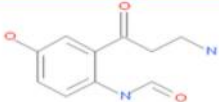
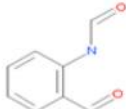
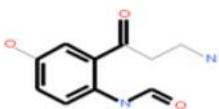
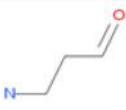
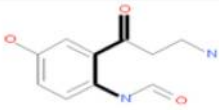
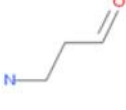
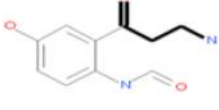
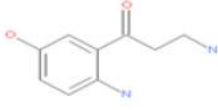
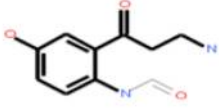

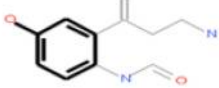
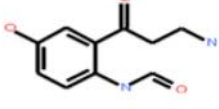
A	Candidate Structure		
	Scaffold Structure	Scaffold -> Candidate Mapping	Similarity Score
B1			11/15 = 0.73
B2			5/15 = 0.3
B3			5/15 = 0.3
B4			13/15 = 0.87
B5			7/15 = 0.47
C	Union Scaffold Structure		15/15 = 1.0

Figure 1. Matching a candidate structure (panel A) with four different scaffolds (panels B1– B5; note that scaffold B2 = scaffold B3) as substructures and the similarity score of each match. The union scaffold structure incorporating all scaffold matches is shown in panel C.

overlap among scaffolds. This representation provides a quantitative assessment of a candidate compound's overall "biochemical coverage". Figure 1 illustrates BioSM's scaffold matching process and shows how scaffolds are mapped onto the candidate structure to generate the union scaffold structure. When multiple matches exist, BioSM incorporates each one into the union scaffold structure being generated (Figure 1, matches B2 and B3). Please note that a disjoint union scaffold structure may be generated if matching substructure scaffolds do not overlap. Once a union scaffold structure is mapped to a candidate structure, a similarity score, known as the union-scaffold score (US), is computed using eq 1 with the candidate structure as the superstructure and the union scaffold structure as the substructure.

We considered using the number of scaffolds that match a candidate structure as an optional scoring parameter. We realized, however, that this approach would make BioSM's predictions biased depending on the over- or underabundance of any particular group of structures in the scaffolds list. Knowing that our scaffolds list is incomplete, since not all endogenous mammalian biochemical compounds are known, we decided to not include the number of scaffold matches in a candidate compound's score.

We also recognized that some candidate structures may be small and thus have very few scaffolds matching as substructures. Obviously, larger candidate compounds have a better chance of matching substructures in the scaffolds list. Accordingly, we modified our method to match and score scaffolds that are superstructures of a candidate structure as well as those that are substructures. This approach seems intuitive since many biochemical compounds are produced enzymatically (i.e., products) from larger precursor scaffolds (i.e., substrates) via biochemical pathways.³⁴ If a scaffold is found to be a superstructure of a candidate structure, a similarity score is computed using eq 1. In addition, a candidate compound may be a substructure of several scaffolds as shown in Figure 2. In that case, the scaffold with the highest similarity score is selected, and that score is used as the superstructure score.

Hence, a candidate compound can have a score of zero (when no matches are found), a union scaffold score, a superstructure score, or both. In order to have one value represent the structural match of a candidate compound to the biochemical scaffold structures, we combined the union scaffold and superstructure scores in two different ways. In the first approach, referred to as the sum of scores (SS), we obtained a candidate's overall score by adding the union scaffold score to

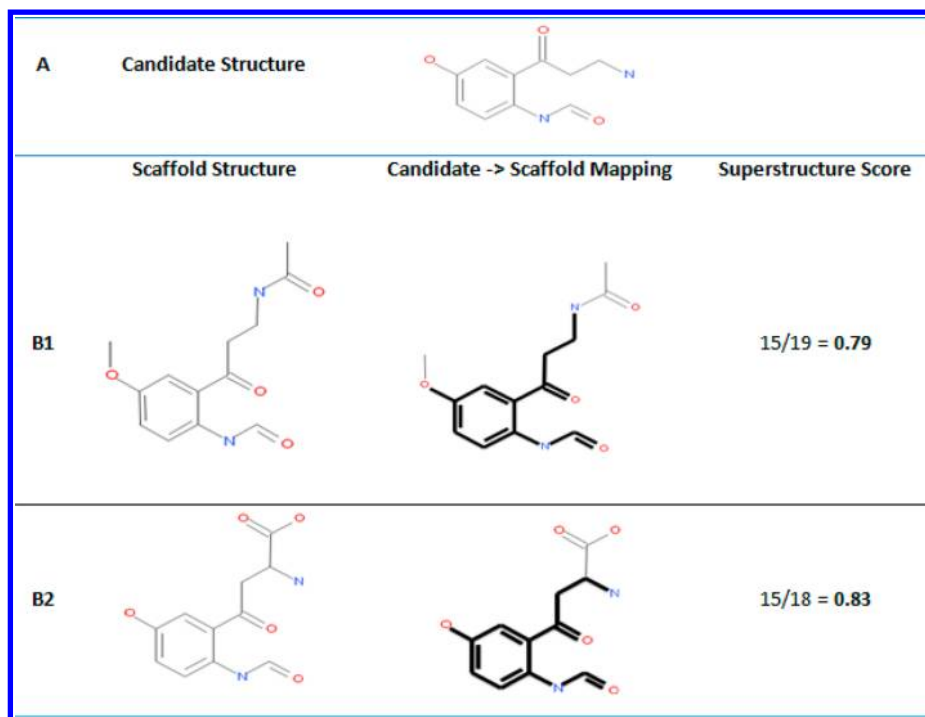


Figure 2. Matching a candidate structure (panel A) with two scaffolds (panels B1 and B2) as superstructures and the similarity score of each match. The scaffold structure with the highest similarity score (scaffold B2) is selected.

the superstructure score. In the second approach, referred to as the maximum score (MS), the candidate's score was the larger of the union scaffold score and superstructure score.

Scaffolds and Synthetic Data Sets. The KEGG database served as the source of the first set of endogenous mammalian scaffolds used in this study. These scaffolds were selected based on their inclusion within at least one of 63 known KEGG mammalian pathways (scaffold pathway and metabolic class information is given in Supplemental Table 1 in the Supporting Information). However, some compounds were excluded from the final scaffold list. Compounds with elements other than C, H, N, O, P, and S are typically found only in marine organisms and extremely rare in mammals. Hence, we decided to treat these compounds as nonmammalian compounds and eliminated them (59 compounds). Molecules with a molecular mass less than 50 Da (12 compounds) were removed. Fifty-nine compounds with any atom type other than C, H, O, N, S, and P were eliminated as were compounds that had duplicate structures (174 compounds) or were polymers (223 compounds). Additionally, we eliminated compounds that did not have a formula associated (27 compounds) and all charged structures (11 compounds) except those in which the charge was due to quaternary amines or sulfonium ions. This curation resulted in a final list of 1565 mammalian scaffolds (KEGGscafs) for our initial representation of biochemical structure space.

The Chembridge and Chemsynthesis databases, comprising synthetic compounds for chemical synthesis and drug screening and design, were chosen to represent nonbiological chemical space. A set of 29 207 compounds was downloaded from the Chemsynthesis database on 7/18/11 and a set of 760 517 compounds was downloaded from the Chembridge database on 7/20/11. Because Chemsynthesis and Chembridge databases mainly contain compounds with low molecular weights, a value of 700 Da was set as the maximum molecular weight of

candidate compounds included in this study. Accordingly, 177 KEGG compounds (with masses greater than 700 Da) were eliminated from any testing set throughout this study and were only used for superstructure scaffold matching. This mass restriction was enforced to ensure that any compound with a mass range 50–700 Da was equally likely to be biological/nonbiological and thus discrimination would be based solely on structure. Similar to KEGGscafs, the combined synthetic set of compounds was curated by removing all compounds containing elements other than C, H, O, N, S, and P (297 721 structures), organic salts (3496 structures), charged compounds (39 170 structures), duplicate compounds (153 structures), and compounds with molecular mass less than 50 Da (8 structures). Additionally, we removed 127 compounds that were identical to compounds in KEGGscafs. This curation resulted in a final set of putative nonbiological compounds consisting of 483 615 structures.

In addition to these nonbiological compounds, we empirically derived a set of nonbiological substructures (NBS) which, to our knowledge, are not commonly found in mammalian biochemical compounds. The NBS list was checked against KEGGscafs. If an NBS was found to be part of a compound in KEGGscafs, the NBS was removed. This resulted in 35 substructures in the final NBS list (Supplemental Table 2). The NBS list was used as an initial filter in the identification process. If a candidate compound was found to contain at least one NBS, it was predicted to be nonbiological.

Accuracy Measures. To evaluate the performance of BioSM, several accuracy measures were employed. Sensitivity (SENS) refers to the proportion of biological compounds correctly predicted to be biological and is computed as

$$\text{SENS} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

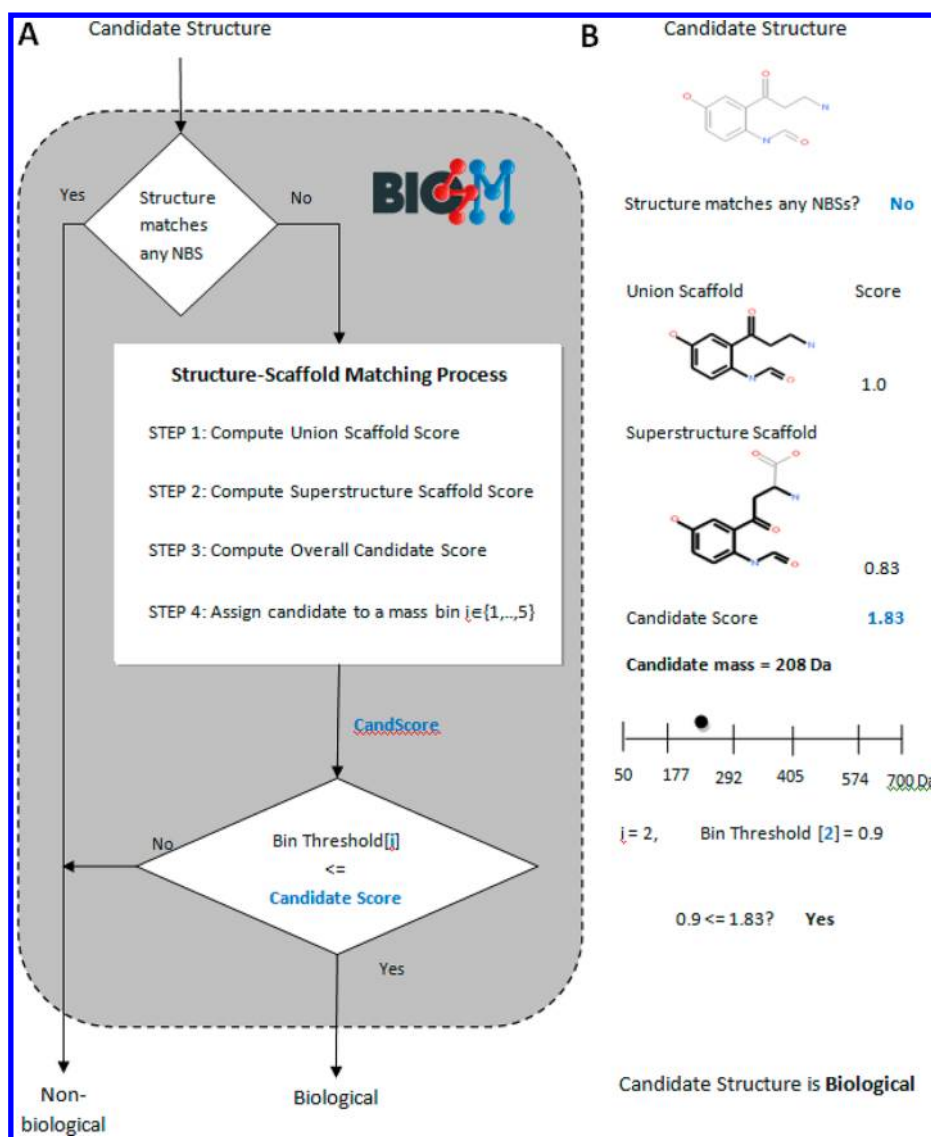


Figure 3. (A) General flow of BioSM and (B) an example showing how the union scaffold structure and superstructure scaffold are used in the prediction process based on SBSS.

where TP represents the number of true positives and FN represents the number of false negatives. Specificity (SPEC) refers to the proportion of nonbiological compounds correctly predicted to be nonbiological and is given by

$$\text{SPEC} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (3)$$

where TN is the number of true negatives and FP represents the number of false positives.³⁵ The positive predictive value (PPV) is the proportion of positive test results that are true positives and is defined by

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4)$$

The Matthews correlation coefficient (MCC),³⁶ defined by

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TN} + \text{FN})(\text{TN} + \text{FP})(\text{TP} + \text{FN})(\text{TP} + \text{FP})}} \quad (5)$$

is commonly used as a combined measure of the overall quality of two-class classifiers. MCC values range from 1 to -1, where MCC = 1 represents perfect prediction, MCC = 0 represents

essentially random prediction, and MCC = -1 represents perfectly inverse prediction. Finally, the *F*-score is the harmonic mean of SENS and PPV, i.e.,

$$F = 2 \frac{\text{SENS} \times \text{PPV}}{\text{SENS} + \text{PPV}} \quad (6)$$

Training Data. From the selected set of 1565 KEGGscaffs, there were 1388 compounds with molecular weights in the range 50–700 Da. These were used as the training set for our method. A set of 1388 synthetic compounds, selected from the synthetic compounds data set to match the mass distribution of the 1388 biological set, was used to represent nonbiological chemical space (the full list of compounds is available in the Supporting Information). Synthetic compounds containing one or more NBS were not used for training since BioSM applies the NBS filter before the scaffolds matching step.

Cross-Validation Framework and Scoring Methods. Cross validation (CV) is one of the simplest and most widely used methods for tuning metaparameters and estimating the accuracy of classification algorithms while avoiding overfitting.³⁵ In this study, we used a nested CV framework,

whereby classification accuracy was empirically assessed using 2-fold CV, with parameter tuning performed by executing 5-fold CV on the training data (Supplemental Figure 1). Briefly, compounds in the scaffolds list and an equal number of mass-matched nonbiological compounds (selected as described above under "Training Data") were individually divided randomly into two halves; one-half for model training and the other half for model testing. The training half (694 biological and 694 nonbiological compounds) was further randomly split into $K = 5$ roughly equal parts, and then, each part was used to evaluate classification accuracy of models trained on the remaining ($K - 1$) parts. For the results of each training fold, the score where SENS = SPEC was recorded as the cutoff threshold of that fold. The average threshold of all five training sets was used as the cutoff score when evaluating the testing data.

Several methods for scoring a candidate compound were examined in this CV analysis. Specifically, the US reflects the value of eq 1 having the candidate compound as the superstructure and the union scaffolds as the substructure, SS reflects the sum of the union scaffold score and the superstructure score, and the MS reflects the larger of the union scaffold score and superstructure score. In preliminary experiments, we noted that the molecular weight of a compound had an impact on its final score. This is because smaller compounds are more likely to match larger scaffolds; larger compounds more likely to match smaller scaffolds; and compounds of intermediate size could match both smaller and larger scaffolds. Therefore, we chose to split the set of test compounds into five mass bins. Fivefold CV was used to determine bin boundaries ensuring that each bin had approximately the same number of compounds, as well as independent score threshold values for each bin. Both threshold scores and bin boundaries obtained from each of the 5 training folds were averaged before applying BioSM to the testing fold. Thus, the sum of threshold values obtained from each fold divided by the number of folds (5) would be the averaged threshold score applied by BioSM to the testing fold. We refer to classification obtained by applying the three scoring methods discussed above with independent threshold values for each of the five bins as *5-bin union-scaffold score* (SBUS), *5-bin sum of scores* (SBSS), and *5-bin maximum score* (SBMS), respectively. Figure 3 shows an overview of the general flow of BioSM and an illustrative example.

Prospective Validation Sets. To estimate the performance of our predictive model, five external validation sets were used; one set of drugs, two sets of putative human metabolites, one set of plant secondary metabolites, and one set of synthetic compounds. Supporting Information Figure 2 shows the mass distribution of the compounds in each validation data set. For each data set, any compound identical to any of KEGGscaffs was removed. Also, structures found in more than one data set were removed from all data sets except one, as explained below. The following is a description of the five data sets:

1. A data set which contained 7036 compounds obtained from DrugBank²⁹ version 3.0 downloaded on 01/18/2012, combined with a set of 5390 structures obtained from the 1989 USAN and the USP Dictionary of Drug Names,³⁷ was used as a drug data set. Salts, mixtures, compounds containing elements other than C, H, N, O, S, and P, duplicate structures, and compounds with

molecular weight outside the 50–700 Da range were removed resulting in a set of 3895 compounds.

2. We used compounds from HMDB version 2.5, downloaded on 7/15/2012, to represent human metabolites. Out of the 8534 molecules in that set, 174 compounds contained elements other than C, H, N, O, S, and P; 4209 molecules were outside the considered mass range (50–700 Da); and 133 compounds had duplicate structures. Additionally, 1138 molecules were eliminated because they were found in KEGGscaffs, and 132 were found in the drug data set. Finally, all charged structures except those in which the charge was due to quaternary amines or sulfonium ions were eliminated. This resulted in an independent data set of 2563 putative human metabolites.
3. We downloaded a set of 2396 compounds from HumanCyc version 16.0 on 5/24/2012 to represent another data set of putative human metabolites. A curated set of 158 compounds were available for testing after eliminating compounds containing elements other than C, H, N, O, S, and P (111 compounds), those not in the mass range 50–700 Da (289 compounds), compounds found in KEGGscaffs (198 compounds), charged compounds (792 compounds), duplicate structures (283 compounds), polymers (368 compounds), drugs (28 compounds), and HMDB compounds (169 compounds).
4. A data set of 2829 secondary plant metabolites,³⁸ as specified by KEGG, was downloaded on 6/25/2012 to represent plant structures. A total of 2416 compounds remained after removing compounds present in KEGGscaffs (75 compounds), drugs (54 compounds), compounds not in the mass range 50–700 Da (217 compounds), compounds containing elements other than C, H, N, O, S, and P (10 compounds), and compounds with charges (57 compounds).
5. A fifth data set of 458 207 compounds from the Chembridge and Chemsynthesis databases, not used in training the model, were used as a synthetic compound test set. The same curation steps described above were used for these compounds.

In addition to these five validation data sets, we classified a random set of compounds taken from the PubChem chemical database. On 12/15/2011, we downloaded 30 142 651 compounds from PubChem. We eliminated 1 003 580 compounds with molecular masses not in the range of 50–700 Da. We further eliminated 13 171 123 compounds that contained elements other than C, H, O, N, S, and P. Three replicate data sets, each containing approximately 320 000 compounds, were randomly chosen from the remaining 15 967 948 PubChem compounds resulting in a total of 959 420 molecules. Further curation resulted in the elimination of 7280 compounds with duplicate structures, 67 449 compounds with charges and 12 compounds that had disconnected structures. This resulted in three random samples totaling 883 199 test molecules. It should be noted that there was no attempt to remove compounds present in any of the other validation sets from the PubChem data set. The PubChem data set was intended to be a random sampling (other than curation requirements) of PubChem compounds.

KEGG, HMDB, and HumanCyc Scaffolds List. In order to determine whether BioSM's prediction accuracy would

improve if the number of scaffolds was increased, we compiled an updated scaffolds list of 3927 compounds (referred to as KHHscaffs) using our initial KEGGscaffs, plus additional compounds from the HMDB and HumanCyc databases. Only nonredundant compounds from HMDB and HumanCyc predicted to be endogenous mammalian biochemical compounds by BioSM using KEGGscaffs were included in KHHscaffs. This list consisted of the original 1565 KEGGscaffs, 2273 compounds from HMDB, and 89 compounds from HumanCyc. A set of compounds from the synthetic data set (randomly selected to match the KHHscaffs mass distribution) were chosen to represent nonbiological compounds. We then used the same cross-validation framework and scoring methods described earlier for KEGGscaffs. BioSM using KHHscaffs was used to analyze the following independent data sets:

1. the drug data set described above (3894 compounds),
2. the plant secondary metabolites data set (2354 compounds) after eliminating 62 compounds found in the KHHscaffs,
3. compounds from the synthetic data set (374 143 Chemsynthesist and Chembridge compounds) not used in training BioSM, and
4. one of the randomly generated Pubchem data sets (294 671 compounds).

RESULTS AND DISCUSSION

Comparison of Candidate Scoring Methods by CV.

The accuracy measures explained above were used to compare results generated from 15 CV experiments for each of the scoring functions (US, MS, SS, SBUS, SBMS, and SBSS) as shown in Table 1. We carried out an analysis-of-variance

Table 1. Mean and Standard Deviation of Accuracy Measures Obtained for 15 Cross-Validation Experiments Using Six Different Scoring Methods and KEGGscaffs ($N = 1565$ Compounds)

		structure scoring methods					
		US	MS	SS	SBUS	SBMS	SBSS
SENS	mean	0.77	0.78	0.78	0.76	0.77	0.79
	stddev	0.02	0.02	0.02	0.03	0.03	0.02
SPEC	mean	0.71	0.71	0.72	0.71	0.71	0.73
	stddev	0.04	0.04	0.04	0.04	0.04	0.04
PPV	mean	0.73	0.74	0.74	0.73	0.73	0.75
	stddev	0.03	0.03	0.03	0.04	0.03	0.03
MCC	mean	0.49	0.5	0.5	0.47	0.48	0.51
	stddev	0.05	0.05	0.05	0.05	0.05	0.04
F score	mean	0.75	0.75	0.75	0.74	0.75	0.76
	stddev	0.02	0.02	0.02	0.02	0.02	0.02

(ANOVA)³⁵ to check for statistical significance between the six scoring methods. We used the Single Factor ANOVA function in Microsoft Excel 2007 to carry out all ANOVA analysis in this study. ANOVA results indicated no statistically significant difference between any of the six methods ($P > 0.05$). However, SBSS accuracy was consistently higher than the other methods on all measures and thus was selected as the scoring method for all remaining experiments.

It is noticeable (Table 1) that the sensitivity of the model in the CV experiments is relatively low. As explained in the Methods section, in each CV experiment only half of the KEGGscaffs were used for training the model and the other half

were used for testing. Thus, a candidate could be predicted to be nonbiological because there were no scaffolds in the randomly selected training set to match it in that specific experiment.

Leave-One-Out Cross-Validation Experiments. Using the averaged metaparameters determined by CV, we carried out a set of leave-one-out cross-validation (LOOCV) experiments on the $N = 1388$ structures (with masses between 50 and 700 Da) in our reference scaffolds database as an additional method of evaluating the accuracy of BioSM in predicting endogenous mammalian biochemical structures. N experiments were performed and for each experiment, $N - 1$ compounds (plus 177 KEGG compounds with masses 700–1200 Da) were used as scaffolds, and the remaining compound was treated as an unknown. This allowed the use of all but one scaffold in the prediction process. As a result, BioSM annotated 95% of the compounds as being biochemical (Supporting Information Table 3).

Prospective Validation. Five prospective data sets (drugs, plant secondary metabolites, two independent human metabolite data sets, and a synthetic molecule data set) were classified by BioSM using the SBSS method. The compounds in each data set were split into five bins (mass range/bin determined as described in the CV experiments), and the percentage of biochemical predictions per bin was computed (Figure 4). For the sake of comparison, the results from the LOOCV experiments with 1388 KEGG endogenous metabolites (described above) are also included in Figure 4. It is observed that the prediction accuracy for KEGG compounds (LOOCV results) is uniform across all mass bins. For the other data sets, compounds in the mass range 287–700 Da (bins 4 and 5) tended to have a higher probability of being predicted as endogenous mammalian biochemical structures. This was especially true for the HumanCyc compounds, plant metabolites, and drugs. The overall results (Table 2) show that out of the 2563 HMDB molecules, 89% were predicted to be biochemical structures. However, only 58% of HumanCyc compounds were predicted to be biological. Visual examination of the HumanCyc structures predicted to be nonbiological showed that many of them are indeed nonbiological. For example, anthrazene, triazene, and compounds with cyclopropane rings are included in the list (these nonbiochemical structures are given in the Supporting Information). Thus, the above results are consistent with the intent of the HMDB and HumanCyc databases to include compounds that are found in humans; however, these are not necessarily endogenous mammalian biochemical compounds.

For the 2416 plant compounds, 72% were predicted to be biochemical. Although this high percentage might seem initially surprising given that we are using mammalian scaffolds to represent biochemical space, this result is consistent with current biochemical and evolutionary data suggesting that plant secondary metabolites and mammalian biochemicals (i.e., our KEGGscaffs) share multiple conserved biochemical pathways and thus an overlapping biochemical phylogeny.³⁹ Interestingly, only 1% of the plant secondary metabolites matched one or more superstructure scaffolds; and those plant compounds were found to have relatively small molecular weights (116–299 Da). This suggests that plants have expanded upon conserved biochemical pathways to produce compounds containing unique combinations of common scaffolds; and these unique combinations are not substructures of known mammalian scaffolds.

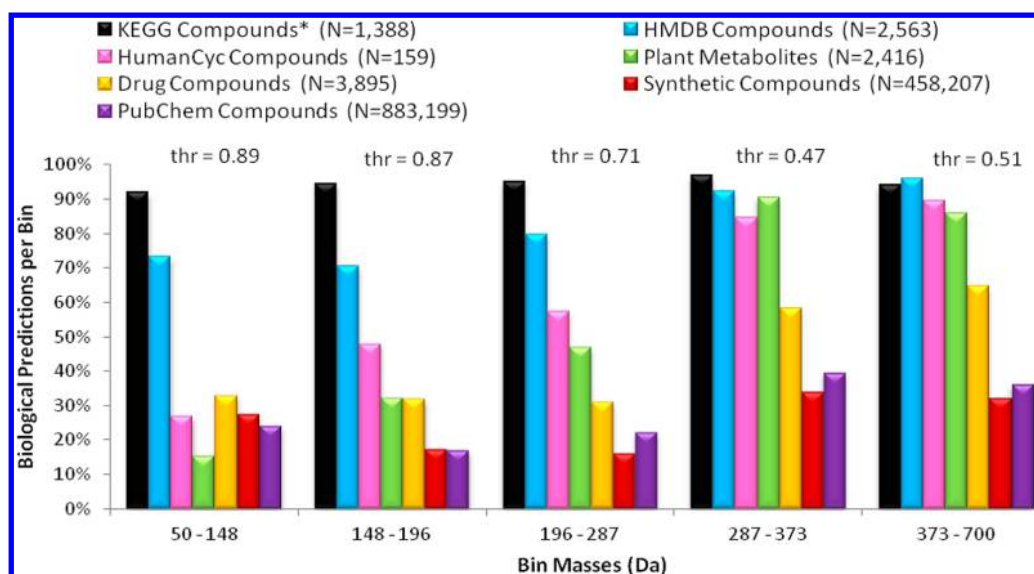


Figure 4. Biological predictions within each mass bin for each data set using KEGGscafs. SBSS bin threshold values (thr) are also displayed. *LOOCV results.

Table 2. Predictive Results Using the SBSS Classifier for Six Different Data Sets Using KEGGscafs

type	number of compounds	prediction		
		nonbiological (NBSS)	nonbiological (SBSS)	biological (SBSS)
HMDB	2563	1%	10%	89%
plant secondary metabolites	2416	0%	28%	72%
HumanCyc	158	7%	35%	58%
Drugs	3895	16%	36%	48%
Synthetics	458207	21%	50%	29%
PubChem	959420	22%	46%	32%

Of 3895 drug structures, 48% were predicted to be endogenous mammalian biochemical structures. These results are very similar to those found earlier by Peironcelly et al. using a similar drug data set.²⁶ It is perhaps not surprising that approximately half of the drugs were predicted to be endogenous biochemical structures since many are derived from natural products.⁴⁰ In contrast, only 29% of the synthetic compounds were predicted to be endogenous biochemical structures. By chance, synthetic compounds may be structurally similar to biochemical compounds. Indeed, as mentioned previously, we found 127 compounds that had to be removed from the synthetic data set prior to cross validation because they were identical to compounds in KEGGscafs.

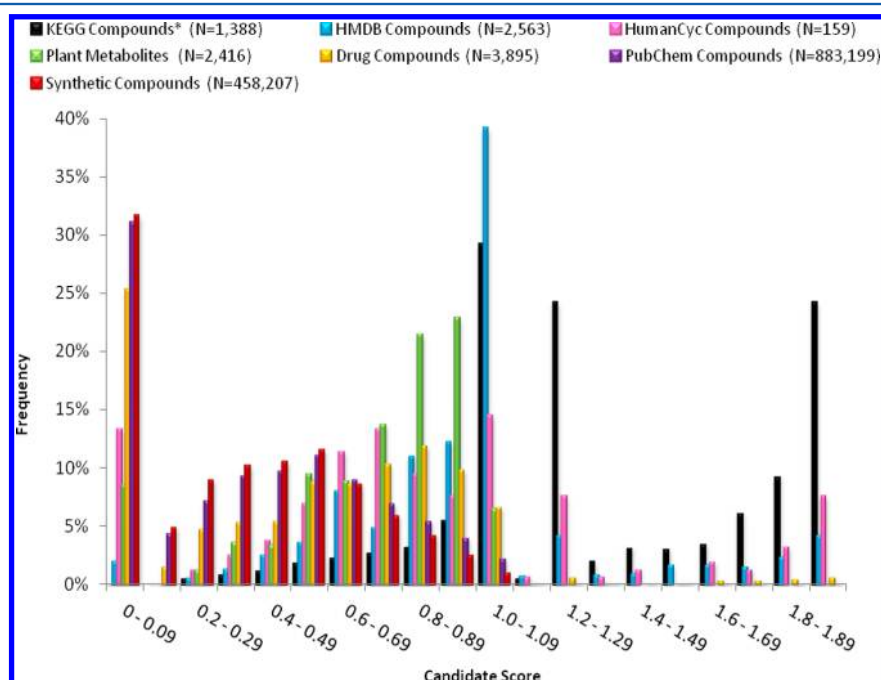


Figure 5. Frequency distribution of candidate scores for each data set. SBSS threshold values for each of the five bin masses are given in Figure 4. *LOOCV results.

In addition to these five prospective data sets, three random samples of approximately 294 000 compounds (883 199 total) from PubChem were tested. Of these, 34% ($\pm 0.02\%$) were predicted to be biochemical. This suggests that the Pubchem database contains mostly nonbiological compounds. Thus, for metabolomics studies where identification of unknown endogenous biochemicals is the primary goal, BioSM would facilitate more efficient use of large chemical databases such as PubChem by removing nonbiological candidate compounds from further consideration. For example, BioSM will be incorporated into MolFind,¹³ a recently described program that aids in the identification of unknown compounds detected in biological samples by LC/MS. Supporting Information Table 4 shows the detailed predictions results for each of the PubChem random samples as well as the average and standard deviation.

Next, we evaluated the distribution of candidate scores regardless of compound mass (Figure 5) for each prospective data set. PubChem compounds, synthetic compounds, and compounds in the drug data set have a large number of compounds (31%, 32%, and 25% respectively) with a candidate score of zero. After eliminating compounds with a zero score due to NBSs (Supporting Information Table 4) we found that 8% of Pubchem compounds, 10% of the synthetic compounds, and 9% of the drug compounds had no structural similarity with any of our scaffolds. It is also clear in Figure 5 that Pubchem compounds and synthetic compounds have very similar candidate score distributions.

A candidate score greater than 1.0 can only be achieved if the candidate compound has at least one matching substructure scaffold *and* at least one matching superstructure scaffold. Figure 5 shows that 82% of the KEGG endogenous compounds, 54% of the HMDB compounds, and 31% the HumanCyc compounds have a scores between 1 and 2. Only a few of the drug, plant, PubChem, and synthetic compound structures have candidate scores in that range (9%, 7%, 2%, and 1%, respectively). As mentioned earlier, only about 1% of the plant compounds matched one or more superstructure scaffolds. Thus, of the 7% of plant compounds with scores between 1 and 2, approximately 6% of these had a score of 1. Using KEGGscaffs, the largest threshold value over all 5 bins was 0.89. Therefore any compound, regardless of its mass, with a score of greater than 0.89 would be annotated as an endogenous mammalian biochemical compound.

KEGG, HMDB, and HumanCyc Scaffolds List. The analysis above was based on using BioSM and our curated set of 1565 KEGGscaffs. This assumes that these 1565 structures provide a complete (or nearly complete) representation of mammalian biochemical structure space. Thus, an important question is whether a larger scaffold list (larger biochemical structure space) would significantly change the results presented above. After updating the scaffolds list to 3927 compounds (KHHscaffs described above), we followed the same process for finding the best scoring method, cutoff values, and bin masses using 15 CV experiments with 3750 training scaffolds ($3927 - 177 = 3750$) in the 50–700 Da mass range. For the nonbiological set, we selected a random set of structures from the Chembridge and Chemsynthesis databases which matched the mass distribution of the 3750 training KHHscaffs. Note that since this nonbiological set was chosen at random from our curated data set of 483 615 synthetic compounds, it is not identical to the nonbiological set used for CV of KEGGscaffs. Table 3 shows the average accuracy

Table 3. Average and Standard Deviation of Accuracy Measures Obtained for 15 Cross-Validation Experiments Using 6 Different Scoring Methods and the KHHscaffs ($N = 3927$ Compounds)

		structure scoring methods					
		US	MS	SS	SBUS	SBMS	SBSS
SENS	mean	0.84	0.84	0.84	0.83	0.84	0.83
	stddev	0.01	0.01	0.01	0.02	0.02	0.02
SPEC	mean	0.72	0.72	0.72	0.73	0.73	0.75
	stddev	0.01	0.01	0.01	0.01	0.01	0.01
PPV	mean	0.81	0.81	0.81	0.82	0.82	0.83
	stddev	0.01	0.01	0.01	0.01	0.01	0.01
MCC	mean	0.56	0.56	0.57	0.56	0.57	0.58
	stddev	0.02	0.02	0.02	0.02	0.02	0.02
F score	mean	0.82	0.82	0.83	0.82	0.83	0.83
	stddev	0.01	0.01	0.01	0.01	0.01	0.01

measures of the 15 CV experiments for US, MS, SS, SBUS, SBMS, and SBSS methods. An ANOVA of the results in Table 3 indicated statistically significant ($P < 0.05$) differences between SPEC and PPV for one or more of the six scoring methods. Having the highest SPEC (0.75) and PPV (0.83), SBSS was selected as the scoring method for BioSM when using KHHscaffs to reanalyze the various data sets as described above. A further ANOVA of the SBSS CV results for KEGGscaffs and KHHscaffs showed a statistically significant ($P < 0.05$) difference between all measures (Supporting Information Table 5).

Figure 6 shows the results of LOOCV as well as the results of the prospective data sets per mass bin. Of the 3750 KHHscaffs, 96% were correctly predicted as biological using a LOOCV (Supporting Information Table 6). Even though this value is high, 4% of our scaffolds were still incorrectly annotated (these structures are found in the Supporting Information). In many cases, we noted that these false negatives were because BioSM requires an exact match between the scaffold and the candidate. This was particularly problematic for predicting specific classes of compounds. For example, lipids with a double bond in the middle of the structure were poorly predicted by BioSM since there may not be scaffolds that match either side of the double bond. We explored using scaffold matching without the requirement of exact bond matching; however, the specificity of the system was negatively affected. It is important to note that bin masses and cutoff thresholds changed after running CV with the updated KHHscaffs. This explains why some compounds predicted to be biological using KEGGscaffs might be predicted to be nonbiological using KHHscaffs or vice versa. Although the 96% sensitivity suggested by our LOOCV analysis is quite good, a possible approach to further improve BioSM would be to expand the set of scaffolds by using enzyme reaction information (oxidation and/or reduction reactions for example). In this case, not only would BioSM be searching for exact structure matches between scaffolds and candidate compounds but also among putative metabolites of those scaffolds. BioSM will apply a set of applicable enzyme reactions to a candidate compound; if any of the metabolites produced were found to be an endogenous mammalian biochemical compound by BioSM, then the candidate is also biochemical.

Using KHHscaffs, BioSM predicted 74% of the 2354 plant compounds, 42% of the 3894 drug compounds, 26% of the 374 143 synthetic compounds, and 25% of the 294 671 random

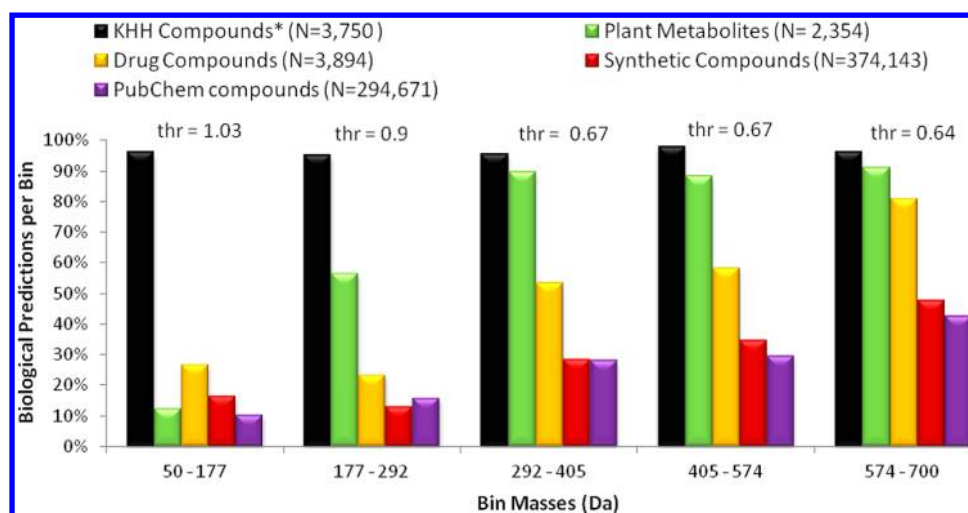


Figure 6. Biological predictions within each mass bin for each data set using KHHscafs. SBSS bin threshold values (thr) are also displayed. *LOOCV results.

Pubchem compounds as biological. It is important to point out that this 25% value for PubChem does not include compounds that were eliminated during the initial curation steps (mass range requirement, compounds with elements other than C, H, N, O, P, and S, stereoisomers, salts, and disconnected structures). Thus, starting with approximately 29 000 000 PubChem compounds with MIMW between 50 and 700 Da, we estimate that approximately 3 680 000 (13%) of these would be annotated as mammalian biochemical compounds using our curation steps and BioSM. Supporting Information Figure 3 shows the distribution of candidate scores from each data set regardless of compound mass.

Figure 7 illustrates the percentage of molecules predicted to be biological by BioSM using KHHscafs versus KEGGscafs in

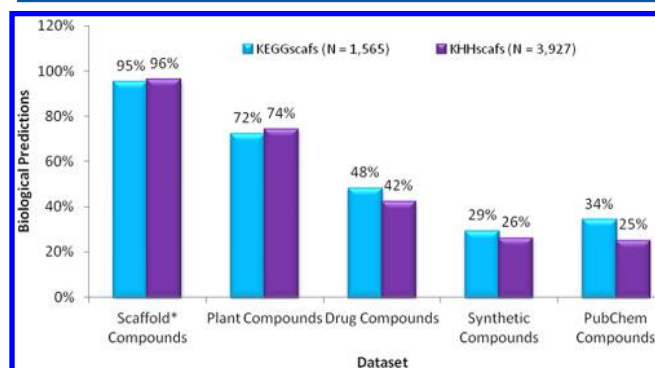


Figure 7. Percentage of biological predictions in each data set using KEGGscafs versus using KHHscafs. *Refer to LOOCV results when using the KEGGscafs data set (turquoise bar) and the KHHscafs (purple bar) as defined in the Methods section above.

each of the prospective data sets. Although sensitivity, specificity, MCC, PPV, and *F* score are all significantly higher when using KHHscafs (Supporting Information Table 5), overall, the percentages predicted to be biological are very similar using the two sets of scaffolds. Thus, it is unlikely that the use of additional scaffolds will significantly improve our representation of biochemical structure space as defined here and that the model is reasonably robust. One could argue that the 2362 added scaffolds may not have contributed appropriate biochemical structure diversity since they were predicted to be

biological using KEGGscafs. However, this seems unlikely due to the large number of nonredundant structures added and the fact that all CV model parameters were significantly improved compared to KEGGscafs. Further slight improvements may still be possible by iteratively expanding the scaffold list; notably, out of the 275 HMDB compounds classified as nonbiological using KEGGscafs, 91 of these were classified as biological using KHHscafs.

It is difficult to measure the accuracy of BioSM based on the results displayed in Figure 7 as there is no definite answer as to whether or not each compound in these data sets is actually an endogenous mammalian biochemical. Yet it is still interesting to see how BioSM classifies compounds from each data set.

Due to the unavailability of sufficient nonbiochemical structures for CV training, BioSM is not currently able to classify compounds with masses above 700 Da. However, regardless of the mass, a quantitative score can be calculated for any compound. Thus, a simple ranking based on candidate scores might still be useful for compounds with masses greater than 700 Da.

Currently, BioSM does not allow annotation of candidate compounds with halogens (i.e., F, Cl, Br) since the current scaffolds list is based upon endogenous human biochemical compounds. However, BioSM can be easily tailored to specific application domains. For example, if one is interested in identifying unknown chemical structures in plant samples, the current scaffolds list can be supplemented with known plant biochemical structures and the NBS list could be appropriately modified (e.g., C#N would be allowed).

CONCLUSIONS

In this work, we describe the development and validation of BioSM, a novel supervised classifier that uses endogenous mammalian biochemical scaffolds to predict whether a candidate chemical structure is biochemical or synthetic. BioSM was able to correctly classify 96% of 3750 biochemical compounds in a leave-one-out cross validation experiment. In addition, our results suggest that approximately 13% of PubChem compounds are mammalian biochemicals. Thus BioSM may be useful for searching large chemical databases in metabolomics applications where the number of potential false positives is very large. Additionally, BioSM can place molecules

in the context of metabolic pathways since it can link potentially unknown biochemicals to matched substructure and superstructure scaffolds for which metabolic pathways are known.

Availability and Requirements

Project name: BioSM (Biological Structure Matcher)
Project home page: <http://metabolomics.pharm.uconn.edu/>
Operating system(s): platform independent (Windows, MAC, Linux/Unix)
Programming language: Java
Other requirements: Java 1.6 or higher
License: Creative Commons <http://creativecommons.org/licenses/by/3.0/>
Any restrictions to use by nonacademics: none

■ ASSOCIATED CONTENT

Supporting Information

KEGGscafs; KHHscafs; LOOCV results (KEGGscafs); LOOCV results (KHHscafs); prediction results using KEGGscafs (HMDB, HumanCyc, plants, and drugs); prediction results using KHHscafs (plants and drugs); list of nonbiological substructures; list of synthetic compounds used in cross validation (KEGGscafs and KHHscafs). This material is available free of charge via the Internet at <http://pubs.acs.org>. Prediction results using KEGGscafs (synthetic compounds and three sets of PubChem compounds) and prediction results using KHHscafs (synthetic compounds and PubChem compounds) is available free of charge via the Internet at <http://metabolomics.pharm.uconn.edu>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: david.grant@uconn.edu

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

M.A.H. would like to thank Dr. Syed A. Rahman for his prompt support with the SMSD toolkit and Dr. Sahar Al Seesi for her continuous support and guidance throughout the project. 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. M.A.H. was responsible for designing the algorithm, software development, and manuscript preparation. She was also responsible for testing and benchmarking BioSM. D.F.G. and D.W.H. were involved in testing the chemical relevance of the tool. I.I.M. was involved in the supervision of the algorithm development. D.F.G., I.I.M., D.W.H., and S.R. were involved in the overall supervision of the project, manuscript preparation, intellectual input, and guidance. All authors have given approval to the final version of the manuscript.

■ ABBREVIATIONS

MS, mass spectrometry; NMR, nuclear magnetic resonance; KEGG, kyoto encyclopedia of genes and genomes; HMDB, human metabolite database; RF, random forest; SMILES, simplified molecular-input line-entry system; US, union-

scaffolds score; SS, sum of scores; MS, maximum score; NBS, nonbiological substructures; CV, cross validation; SBUS, 5-bin union-scaffold score; SBSS, 5-bin sum of scores; SBMS, 5-bin maximum score; LOOCV, leave-one-out cross validation

■ REFERENCES

- (1) Dettmer, K.; Aronov, P. A.; Hammock, B. D. Mass Spectrometry-based Metabolomics. *Mass Spectrom. Rev.* **2007**, *26*, 51–78.
- (2) Roessner, U.; Bowne, J. What is metabolomics all about? *BioTechniques* **2009**, *46*, 363–365.
- (3) Rochfort, S. Metabolomics Reviewed: A New “Omics” Platform Technology for Systems Biology and Implications for Natural Products Research. *J. Nat. Prod.* **2005**, *68*, 1813–1820.
- (4) Hall, R.; Beale, M.; Fiehn, O.; Hardy, N.; Sumner, L.; Bino, R. Plant metabolomics: the missing link in functional genomics strategies. *The Plant Cell* **2002**, *14*, 1437–1440.
- (5) Adams, J. C.; Keiser, M. J.; Basuino, L.; Chambers, H. F.; Lee, D.-S.; Wiest, O. G.; Babbitt, P. C. A mapping of drug space from the viewpoint of small molecule metabolism. *PLoS Comput. Biol.* **2009**, *5*.
- (6) Harvey, A. L. Natural products in drug discovery. *Drug Discovery Today* **2008**, *13*, 894–901.
- (7) Khanna, V.; Ranganathan, S. Physiochemical property space distribution among human metabolites, drugs and toxins. *BMC Bioinf.* **2009**, *10* (Suppl 1), S10.
- (8) Dixon, R. A.; Gang, D. R.; Charlton, A. J.; Fiehn, O.; Kuiper, H. A.; Reynolds, T. L.; Tjeerdema, R. S.; Jeffery, E. H.; German, J. B.; Ridley, W. P.; Seiber, J. N. Applications of metabolomics in agriculture. *J. Agric. Food Chem.* **2006**, *54*, 8984–8994.
- (9) Nadella, K. D.; Marla, S. S.; Kumar, P. A. Metabolomics in agriculture. *OMICs* **2012**, *16*, 149–159.
- (10) Heux, S.; Fuchs, T. J.; Buhmann, J.; Zamboni, N.; Sauer, U. A high-throughput metabolomics method to predict high concentration cytotoxicity of drugs from low concentration profiles. *Metabolomics* **2012**, *8*, 433–443.
- (11) Reo, N. V. NMR-based metabolomics. *Drug Chem. Toxicol.* **2002**, *25*, 375–382.
- (12) Kertesz, T.; Hill, D. W.; Albaugh, D.; Hall, L.; Hall, L.; Grant, D. F. Database searching for structural identification of metabolites in complex biofluids for mass spectrometry-based metabolomics. *Bioanalysis* **2009**, *1*, 1627–1643.
- (13) Menikarachchi, L. C.; Cawley, S.; Hill, D. W.; Hall, L. M.; Hall, L.; Lai, S.; Wilder, J.; Grant, D. F. MolFind: a software package enabling HPLC/MS-based identification of unknown chemical structures. *Anal. Chem.* **2012**, *84*, 9388–9394.
- (14) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623–W633.
- (15) Kanehisa, M.; Goto, S.; Kawashima, S.; Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **2002**, *30*, 42–46.
- (16) Wishart, D. S.; Knox, C.; Guo, A. C.; Eisner, R.; Young, N.; Gautam, B.; Hau, D. D.; Psychogios, N.; Dong, E.; Bouatra, S.; Mandal, R.; Sinelnikov, I.; Xia, J.; Jia, L.; Cruz, J. a; Lim, E.; Sobsey, C. a; Shrivastava, S.; Huang, P.; Liu, P.; Fang, L.; Peng, J.; Fradette, R.; Cheng, D.; Tzur, D.; Clements, M.; Lewis, A.; De Souza, A.; Zuniga, A.; Dawe, M.; Xiong, Y.; Clive, D.; Greiner, R.; Nazzyrova, A.; Shaykhtudinov, R.; Li, L.; Vogel, H. J.; Forsythe, I. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.* **2009**, *37*, D603–D610.
- (17) Romero, P.; Wagg, J.; Green, M. L.; Kaiser, D.; Krummenacker, M.; Karp, P. D. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.* **2004**, *6*, R2.1–R2.17.
- (18) Schymanski, E. L.; Meringer, M.; Brack, W. Automated Strategies To Identify Compounds on the Basis of GC/EI-MS and Calculated Properties. *Anal. Chem.* **2011**, *83*, 903–912.
- (19) Hamdalla, M.; Grant, D.; Mandoiu, I.; Hill, D.; Rajasekaran, S.; Ammar, R. The use of graph matching algorithms to identify biochemical substructures in synthetic chemical compounds: Applica-

tion to metabolomics. In *2012 IEEE 2nd International Conference on Computational Advances in Bio and medical Sciences (ICCBMS)*, Las Vegas, NV, Feb 23–25, 2012.

(20) Nobeli, I.; Ponstingl, H.; Krissinel, E. B.; Thornton, J. M. A structure-based anatomy of the E.coli metabolome. *J. Mol. Biol.* **2003**, *334*, 697–719.

(21) Gupta, S.; Aires-de-Sousa, J. Comparing the chemical spaces of metabolites and available chemicals: models of metabolite-likeness. *Mol. Diversity* **2007**, *11*, 23–36.

(22) Irwin, J. J.; Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

(23) Breiman, L. Random forests. In *Machine Learning*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2001; Vol. 45, pp 5–32.

(24) Peironcelly, J. E.; Reijmers, T.; Coulier, L.; Bender, A.; Hankemeier, T. Understanding and classifying metabolite space and metabolite-likeness. *PLoS One* **2011**, *6*.

(25) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

(26) Warr, W. A. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). *J. Comput.-Aided Mol. Des.* **2009**, *23*, 195–198.

(27) James, C. A.; Weininger, D.; Delany, J. Fingerprints - Screening and Similarity. In *Daylight Theory Manual*; Daylight Chemical Information Systems, Inc.: Irvine, CA and Santa Fe, NM, 2000.

(28) Chembridge. www.chembridge.com/ (accessed July 20, 2012).

(29) Chemsynthesis. www.chemsynthesis.com/ (accessed July 18, 2012).

(30) Maggiora, G. M. Shanmugasundaram, V. Molecular Similarity Measures. In *Cheminformatics and Computational Chemical Biology*; Humana Press: New York, 2011; Vol. 672, pp 39–100.

(31) Marvin, version 5.10; ChemAxon, Hungary, Budapest, 2012.

(32) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(33) Rahman, S. A.; Bashton, M.; Holliday, G. L.; Schrader, R.; Thornton, J. M. Small Molecule Subgraph Detector (SMSD) toolkit. *J. Cheminf.* **2009**, DOI: 10.1186/1758-2946-1-12.

(34) Macchiarulo, A.; Thornton, J. M.; Nobeli, I. Mapping human metabolic pathways in the small molecule chemical space. *J. Chem. Inf. Model.* **2009**, *49*, 2272–2289.

(35) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer Series in Statistics; Springer: New York, 2009.

(36) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442–451.

(37) *USAN and the USP Dictionary of Drug Names*; United States Pharmacopeia: Rockville, MD, 1989.

(38) KEGG Phytochemical Compounds. www.genome.jp/kegg-bin/get_htext?org_name=br08003&query=&htext=br08003.keg&filedir=&highlight=&option=-&extend=C1-162B19&uploadfile=&format=&wrap=&length=&open=&close=&hier=0 (accessed June 25, 2012).

(39) Weng, J.-K.; Philippe, R. N.; Noel, J. P. The rise of chemodiversity in plants. *Science* **2012**, *336*, 1667–1670.

(40) Mishra, B. B.; Tiwari, V. K. Natural products in drug discovery: Clinical evaluations and investigations. In *Opportunity, Challenge and Scope of Natural Products in Medicinal Chemistry*; Research Signpost: Ontario, Canada, 2011; Vol. 661, pp 1–62.