Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

IEEE Int Conf Comput Adv Bio Med Sci. Author manuscript; available in PMC 2015 October 05.

Published in final edited form as:

IEEE Int Conf Comput Adv Bio Med Sci. 2012 February ; 2012: .

The use of graph matching algorithms to identify biochemical substructures in synthetic chemical compounds: Application to metabolomics

Mai Hamdalla[1], **David Grant**[2], **Ion Mandoiu**[1], **Dennis Hill**[2], **Sanguthevar Rajasekaran**[1], and **Reda Ammar**[1]

Mai Hamdalla: mai@engr.uconn.edu

[1]Computer Science and Engineering Department, University of Connecticut, Connecticut, USA

[2]Pharmaceutical Sciences Department, University of Connecticut, Connecticut, USA

Abstract

Metabolomics is a rapidly growing field studying the small-molecule metabolite profile of a biological organism. Studying metabolism has a potential to contribute to biomedical research as well as drug discovery. One of the current challenges in metabolomics is the identification of unknown metabolites as existing chemical databases are incomplete. We present a novel way of utilizing known mammalian metabolites in an effort to identify unknown ones. The system relies on a mammalian scaffolds database to aid the classification process. The results show that 96% of the mammalian compounds were identified as truly mammalian in a leave-one-out experiment. The system was also tested with a random set of synthetic compounds, downloaded from ChemBridge and ChemSynthesis databases. The system was able to eliminate 54% of the set, leaving 46% of the compounds as potentially unknown mammalian metabolites.

Keywords

metabolomics; metabolites; molecular similarity; structure matching; mass spectrometry; classification

I. Introduction

Metabolomics is a rapidly evolving discipline involving the systematic study of endogenous small molecules that characterize the metabolic pathways of biological systems [1]. It is closely related to the genomics, transcriptomics and the proteomics and plays an increasingly important role in current biomedical research [2, 3].

The goals of most metabolomics studies are to identify small-molecule metabolites in tissues and biofluids, and to correlate their levels with physiological and/or toxicological endpoints [4]. Although many challenges remain in this field, the metabolite identification process itself remains one of the most important.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Liquid chromatography coupled to electrospray ionization mass spectrometry (LC/MS) is becoming a method of choice for profiling metabolites in complex biological samples [4-6]. For any tissue or biofluid examined, there are hundreds to potentially thousands of compounds that can be “detected” using LC/MS. However, only a handful of these can be reliably associated with actual chemical structures even when searching large chemical databases. Additionally, screening multiple candidate compounds against the thousands of accessible compounds in databases does not seem to be a practical option [7]. Similarly, in drug discovery, the search for pharmaceutically active drugs can be considered a multi-objective optimization problem over an enormous search space of “possible” drugs [8, 9]. In such cases, chemoinformatic methods are used to constrain the compounds screened, in an attempt to narrow the search space of chemically diverse candidate compounds, such that they display ‘metabolite-likeness’ [10], ‘lead-likeness’ [11-13] or ‘drug-likeness’ [14, 15].

Nobeli *et al*.[16] presented a first attempt to examine the metabolome of an organism, using two-dimensional molecular structures and a variety of chemoinformatics tools. Based on the fact that similar molecules will tend to have similar biological properties [17] they used a library of 57 fragments to act as scaffolds. The fragments were manually derived by visual examination of metabolite 2D diagrams making them subjective.

In this paper, we establish a scaffolds database (1,400 compounds) including all currently known mammalian metabolites (to the best of our knowledge) and present a system capable of efficiently and accurately classifying unknown compounds as non-mammalian or mammalianlike. Our classification method is based on a novel scoring scheme that combines all matches of scaffolds to substructures of the unknown compounds, as well as matches of the unknown compound to substructures of the scaffolds.

II. Methods

Our classification process (summarized in Fig. 1) starts with a set of uncategorized candidate compounds. Each candidate compound is represented by its molecular structure in the form of a canonical SMILES string. SMILES (Simplified Molecular-Input Line-Entry System) is a way of presenting chemical molecular structures using short ASCII strings that are easily converted into two-dimensional models [18]. These sets of compounds first go through a filtration process where compounds containing at least one non-biological substructure are eliminated. Non-biological substructures (NBS) are substructures that are not commonly found in biological compounds. We empirically derived a list of non-biological substructures that were checked against our mammalian scaffolds database (scaffolds list). The scaffolds list is a list of structures known to exist in mammalian pathways. If a substructure was found amongst the scaffolds list, it was removed from the NBS list.

Candidate compounds surviving this elimination phase are then matched against the scaffolds list. Candidate compounds that contain one or more scaffold structure are scored and ranked. Candidates with a score higher than a predefined threshold were declared to be biological.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Compounds in our scaffolds list were compiled in the following manner: All compounds listed as components of one or more Metabolic Pathways in the KEGG database [19] were retrieved on 4/23/2011. Compounds in this list that were listed as participants in at least one of the following metabolic pathway groups were retrieved: Carbohydrate, Energy, Lipid, Nucleotide, Amino Acid, Other Amino Acid, Glycan, Cofactors, and Vitamins Metabolism. Each of these compounds were listed as participants in one or more of the following 91 KEGG numbered pathways: ko00010, ko00020, ko00030, ko00040, ko00051, ko00052, ko00053, ko00061, ko00062, ko00071, ko00072, ko00100, ko00120, ko00121, ko00130, ko00140, ko00190, ko00195, ko00196, ko00230, ko00240, ko00250, ko00260, ko00270, ko00280, ko00290, ko00300, ko00310, ko00330, ko00340, ko00350, ko00360, ko00380, ko00400, ko00410, ko00430, ko00440, ko00450, ko00460, ko00471, ko00472, ko00473, ko00480, ko00500, ko00510, ko00511, ko00512, ko00513, ko00514, ko00520, ko00531, ko00532, ko00533, ko00534, ko00540, ko00550, ko00561, ko00562, ko00563, ko00564, ko00565, ko00590, ko00591, ko00592, ko00600, ko00601, ko00603, ko00604, ko00620, ko00630, ko00640, ko00650, ko00660, ko00670, ko00680, ko00710, ko00720, ko00730, ko00740, ko00750, ko00760, ko00770, ko00780, ko00785, ko00790, ko00830, ko00860, ko00900, ko00910, ko00920, and ko01040.

Entries that were single elements, metals, inorganic, n polymers or had no elemental formula were removed from the list. For the remaining compounds, corresponding structures were downloaded from the PubChem database [20] in the form of canonical SMILES. Compounds that did not have an entry in the PubChem database were eliminated, resulting in a scaffolds list of 1,987 distinct structures in the mass range of 25 – 1000 daltons (da).

A. Structure-Scaffold Matching

In the structure matching step, the Small Molecule Subgraph Detector (SMSD) Toolkit [21] was used for molecule similarity searches. SMSD is a Java based software library for finding the Maximum Common Sub-graph (MCS) between small molecules. It uses atom type matches with bond sensitivity information to evaluate molecular similarity. In this study, SMSD has been restricted to consider a match only if the scaffold (smaller structure) is an exact substructure of the structure being compared (larger structure). This restriction has been enforced in both the structure elimination phase (using our list of NBSs) and the structure inclusion phase (matching candidates to the scaffolds). Since SMSD guarantees that a given compound is an exact substructure of another in terms of atoms, bonds, and structure, we found that the percentage of atoms discovered would be a sufficient similarity measure. Equation 1 is used to compute the similarity score between any two compounds (candidates and scaffolds)

$$Silmilarity\ Score = \frac{N_{SBS}}{N_{SPR}} \quad (1)$$

where N_{SBS} represents the number of atoms in the substructure and N_{SPR} represents the number of atoms in the superstructure. Table I shows an example of assigning similarity scores to candidates using (1). Obviously, a candidate compound may match more than one scaffold. Consequently, more than one score may be associated with it as seen in Table I.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Initially, we selected the score of the best match to represent a compound in the final ranking of all candidates. After examining a few compounds, it was clear that information was missed by following this approach. Hence the idea of creating a *union scaffold structure*, that incorporates all the scaffolds matching a candidate compound, was investigated.

B. Union Scaffold Construction

A union scaffold is formed by taking the union of all substructures of a candidate compound that are exact matches of scaffold structures. Table I illustrates the step by step construction of a union scaffold.

The union scaffold provides a quantitative assessment of a candidate compound's overall “biological coverage”. Equation 1 is then used to compute the similarity score between the candidate structure and the union scaffold. By constructing a union scaffold for each candidate compound, each candidate is assigned only one score and can be easily ranked.

In evaluating the effectiveness of the union scaffold concept, we noticed that some of the smaller mammalian compounds were not categorized as mammalian. The reason behind this was that larger candidate compounds have a higher chance of having scaffolds as substructures.

As the candidate structure gets smaller that chance decreases drastically. In next section we propose an approach for correcting this bias by matching candidate structures against substructures of scaffolds.

C. Superstructure Matching

Using SMSD, each candidate compound is tested against larger scaffold compounds in the scaffolds list for sub-graph matches. If a scaffold is found to be a superstructure of a candidate, a similarity score is computed using (1). It is apparent that a candidate compound may be a substructure of several scaffolds, leading to the same issue of multiple scores. In this case, the highest similarity score, which represents the best match between this candidate and a scaffold, is used as the “superstructure score” (as shown in Table II).

D. Structure Scoring

At this point, a candidate compound can have a union scaffold score, a superstructure similarity score, or both. If we use the union scaffold score only, we might be excluding smaller structures from being classified as biological. If we use the superstructure score only, we may exclude larger candidate structures. We decided to use both scores, and to select among various methods of combining the two scores by cross validation.

Specifically, as discussed in the Results section, the best scoring scheme was selected by performing a 5-fold cross-validation on some training data. Two different ways of combining the union scaffold and superstructure matching scores were considered. The first approach computes a candidate compound's score by adding the union scaffold score and the superstructure score, while the other approach considers the candidate compound's score to be the maximum of the union scaffold score and the superstructure score.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

E. Synthetic Datasets

The Synthetic datasets used in the cross validation analysis (refer to Results section) and the independent testing experiments were randomly selected from ChemBridge [1] and ChemSynthesis [2] databases. About 400,000 compounds were retrieved from both databases. Synthetic compounds were restricted to the 6 biological elements C, H, N, O, P, and S. The mass distribution of molecules in ChemBridge was in the range of 150 – 700 da while that of ChemSynthesis was in the range of 50 – 300 da. By combining both databases we managed to have synthetic molecules with masses ranging from 50 – 700 da. Consequently, only the mammalian compounds that fell within this mass range were kept, reducing the mammalian scaffold list to 1,400 compounds. From the curated synthetic list 1,400 compounds were randomly selected to participate in the cross-validation analysis as a training set and 5,320 compounds were randomly chosen for independent testing experiments.

III. Results and Discussion

A. Comparison of Scoring Methods

For an initial evaluation of the scoring methods previously mentioned, we performed a cross validation analysis using our scaffolds list and a random set of 1,400 synthetic compounds (retrieved from ChemBridge and ChemSynthesis databases).

Cross Validation (CV) is one of the simplest and most widely used methods for estimating the accuracy of classification algorithms [22]. Briefly, both the synthetic and mammalian compounds were randomly split in half; one half for training the model and the other half for testing it. The training half was randomly split into K roughly equal parts, and then each part was used to evaluate classification accuracy of a model trained on the remaining (K – 1) parts. In our experiments we used K = 5, i.e., 5-fold cross-validation.

Several methods for scoring a candidate compound were examined in this CV analysis. Specifically, the Union-Scaffold Score (US) – reflects the value of (1); having the candidate compound as the superstructure and the union scaffolds as the substructure, the Sum of Scores (SS) – reflects the sum of the union scaffold score and the superstructure score, the Maximum Score (MS) – reflects the largest of the union scaffold score and the superstructure score. After some preliminary investigation, it was noticeable that the mass of a compound might have an impact on its final score. Therefore, we considered splitting test compounds into 5 mass bins and used CV to find cutoff thresholds for each bin. Bin boundaries were also found through CV. The same scoring methods, referred to as 5 Bin Union-Scaffold Score (5US), 5 Bin Sum of Scores (5SS) and 5 Bin Maximum Score (5MS), were applied to each of the 5 bins. Fifteen (5-fold) CV experiments were executed to evaluate the performance of our system regarding the scoring methods mentioned above.

Table III shows the average sensitivity (SENS), specificity (SPEC) and the Matthews correlation coefficient (MCC) over the 15 (5-fold) CV experiments for US, SS, MS, 5US,

[1] http://www.chembridge.com/index.php
[2] http://www.chemsynthesis.com/

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

5SS and 5MS. Sensitivity refers to the proportion of compounds that are biological and have been predicted by the system to be biological. Specificity refers to the proportion of compounds that are non-biological and have been predicted to be non-biological [22]. The MCC is used in machine learning as a measure of the quality of binary (two-class) classifications. It returns a value between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction and -1 an inverse prediction [22].

SS outperformed all other scoring techniques with a sensitivity of 88% while 5SS had the highest specificity of 78%. According to MCC, 5SS is the best classifier in all 6 methods. Accordingly, the “5 Bin Sum of Scores” method was used when testing the system on the independent datasets.

B. Validation on Independent Test Data

Traditionally, one would use unseen data to validate the performance of a system. In this case, we had already used all the mammalian scaffolds available (to our knowledge) in the training phase. We are not aware of other true mammalian compounds to use in the validation step. To overcome this limitation in data availability, we carried out a set of leave-one-out experiments on our mammalian scaffolds list using the bin boundaries and the similarity score thresholds obtained by the 5-fold Cross Validation experiments.

For a dataset with N compounds, N experiments were performed. For each experiment, $N-1$ compounds were used as scaffolds and the remaining compound was used for testing. As a result, our system was able to identify 96% of the scaffolds as mammalian compounds. Table IV shows the results broken down into 13 (50 da) bins. Each row represents a bin with the number of compounds classified as mammalian/non-mammalian and the percentage of each.

Table V shows the performance of the system when a set of 5,320 randomly selected synthetic compounds were tested. Similar to Table IV, the results are shown in the form of 13 50 da bins. Our system classified 46% of the compounds as mammalian compounds. In other words, it was able to filter out 54% of the compounds as being non-mammalian. That being said, a potential use of our system is to look for compounds classified as mammalian among synthetic lists because they are more likely to have biological activity.

IV. Conclusion

In this study, we presented a novel supervised classification method with the capability of eliminating compounds that are non-mammalian by efficiently using known mammalian metabolites. To this aim we developed a scaffolds database (1,400 compounds) that incorporates all known mammalian metabolites (to the best of our knowledge). We also introduced new ways of handling multiple scaffold matches by constructing a union scaffold structure and incorporating superstructure matches.

Leave one out experiments results show that 96% of the mammalian compounds are correctly identified by the proposed classification scheme with detection thresholds selected by cross-validation on the training data. In validation experiments conducted on an

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

independent set of synthetic compounds, 54% of the compounds were eliminated as being non-mammalian. These encouraging results suggest that the proposed method can be a useful aid in the difficult processes of identification of unknown metabolites and drug discovery. In ongoing work we are exploring further improvements in classification accuracy by using known biological pathway information.

Acknowledgments

This project was supported in part by grants R01-GM087714 and R01-LM010101 from NIH, the Agriculture and Food Research Initiative Competitive Grant no. 2011-67016-30331 from the USDA National Institute of Food and Agriculture, and awards IIS-0546457 and IIS-0916948 from NSF. MH would like to thank Syed A. Rahman for his prompt support with SMSD. MH would also like to thank Michael Zuba for his critical review and Sahar Al Seesi for her continuous support and advice.

References

1. Kouskoumvekaki, Irene; Panagiotou, Gianni. Navigating the human metabolome for biomarker identification and design of pharmaceutical molecules. Journal of biomedicine & biotechnology. 2011
2. Lewis-Stanislaus, Avalyn E.; Li, Liang. A method for comprehensive analysis of urinary acylglycines by using ultra performance liquid chromatography quadrupole linear ion trap mass spectrometry. 2010; 21:2105–2116.
3. Fiehn, Oliver. Metabolomics - the link between genotypes and phenotypes. Plant Molecular Biology. 2002; 48:155–171. [PubMed: 11860207]
4. Kind, Tobias; Fiehn, Oliver. Advances in structure elucidation of small molecules using mass spectrometry. Bioanalytical Reviews. 2010; 2:23–60. [PubMed: 21289855]
5. Kertesz, Tzipporah M.; Hill, Dennis W.; Albaugh, Daniel R.; Hall, Lowell H.; Hall, L Mark; Grant, David F. Database searching for structural identification of metabolites in complex biofluids for mass spectrometry based metabonomics. Bioanalysis. 2009; 1:1627–1643. [PubMed: 21083108]
6. Bowen, Benjamin P.; Northen, Trent R. Dealing with the unknown: Metabolomics and metabolite atlases. Journal of The American Society for Mass Spectrometry. 2010; 21:1471–1476. [PubMed: 20452782]
7. Wendt, Bernd; Uhrig, Ulrike; BoIs, Fabian. Capturing Structure-Activity Relationships from Chemogenomic Spaces. Journal of Chemical Information and Modeling. 2011
8. Lipinski C, Hopkins A. Navigating chemical space for biology and medicine. Nature. 2004; 432:855–861. [PubMed: 15602551]
9. Paolini GV, et al. Global mapping of pharmacological space. Nat Biotechnol. 2006; 24:805–815. [PubMed: 16841068]
10. Dobson, Christopher M. Chemical space and biology. Nature. 2004; 432:824–828. [PubMed: 15602547]
11. Rishton GM. Nonleadlikeness and leadlikeness in biochemical screening. Drug Discov Today. 2003; 8:86–96. [PubMed: 12565011]
12. Hann MM, Oprea TI. Pursuing the leadlikeness concept in pharmaceutical research. Curr Opin Chem Biol. 2004; 8:255–263. [PubMed: 15183323]
13. Wunberg T, et al. Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits. Drug Discov Today. 2006; 11:175–180. [PubMed: 16533716]
14. Gillet VJ, et al. Identification of biological activity profiles using substructural analysis and genetic algorithms. J Chem Inf Comput Sci. 1998; 38:165–179. [PubMed: 9538517]
15. Wagener M, van Geerestein VJ. Potential drugs and nondrugs: prediction and identification of important structural features. J Chem Inform Comput Sci. 2000; 40:280–292.
16. Nobeli, Irene; Ponstingl, Hannes; Krissinel, Eugene B.; Thornton, Janet M. A structure-based anatomy of the e.coli metabolome. Journal of Molecular Biology. 2003; 334:697–719. [PubMed: 14636597]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

17. Pattenon D, Cramer Richard D, Ferguson Allan M, Clark Robert D, Weinberger Laurence E. Neighborhood behavior: A useful concept for validation of molecular diversity descriptors. Journal of Medicinal Chemistry. 1996; 39:3049–3059. [PubMed: 8759626]
18. Weinberger, David. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences. 1988; 28:31–36.
19. Kanehisa, Minoru; Goto, Susumu; Kawashima, Shuichi; Nakaya, Akihiro. The kegg databases at genomenet. Nucleic Acids Research. 2002; 30:42–46. [PubMed: 11752249]
20. Wang, Yanli, et al. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic acids research. Jul; 2009 37(Web Server issue):W623–W633. [PubMed: 19498078]
21. Rahman, Syed; Bashton, Matthew; Holliday, Gemma; Schrader, Rainer; Thornton, Janet. Small molecule subgraph detector (smsd) toolkit. Journal of Cheminformatics. 2009; 1:12–13. [PubMed: 20298518]
22. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome H. The Elements of Statistical Learning. 2008

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

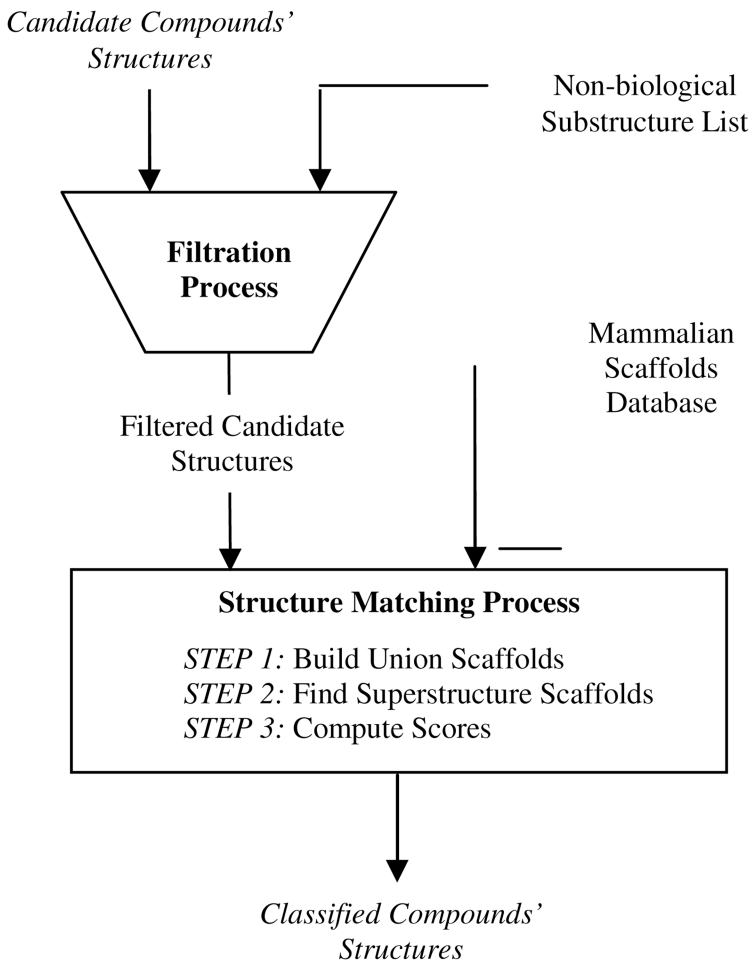
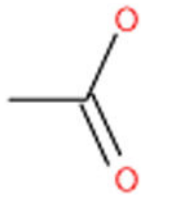
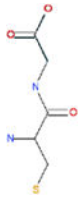
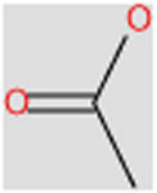
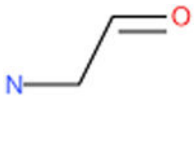
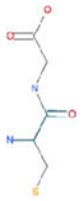
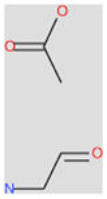
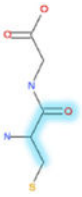
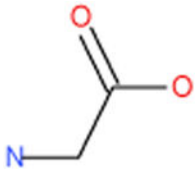
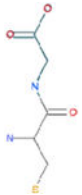
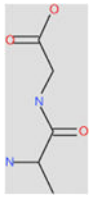


Figure 1.
The Classification Process.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table I

Union Scaffold Construction

Matched Scaffold	Candidate Compound	Similarity Score	Union Scaffold Structure	Union Scaffold Score
		4/11 = 0.36		4/11 = 0.36
		4/11 = 0.36		8/11 = 0.73
		4/11 = 0.36		9/11 = 0.82
		5/11 = 0.45		10/11 = 0.91

Four scaffolds are found to be substructures of this candidate compound. The highest similarity score is 0.45. Instead of ignoring the matches that have a lower similarity score, we build a union scaffold. The union scaffold similarity score of this candidate is 0.91.

Author Manuscript

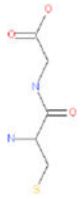
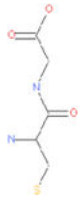
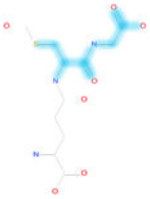
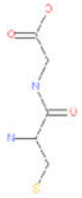
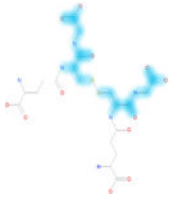
Author Manuscript

Author Manuscript

Author Manuscript

Table II

Superstructure Matching

Candidate Compound	Matched Superstructure Scaffold	Superstructure Score
		11/20 = 0.55
		11/22 = 0.5
		22/40 = 0.55

This candidate compound is a substructure of 3 scaffolds. The highest similarity score of the 3 matches is selected to be the superstructure score of the candidate compound. In this case the superstructure score is 0.55

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table III

5-Fold CV Average Accuracy Results

	Structure Scoring Methods					
	US	MS	SS	5US	5MS	5SS
SENS	70%	59%	88%	83%	84%	86%
SPEC	65%	71%	57%	75%	76%	78%
MCC	0.36	0.3	0.47	0.57	0.60	0.64

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table IV

Leave One Out Analysis

Bin #	Bin masses (da)	Mam Count	Non-Mam Count	Mam%	Non-Mam %
1	50-100	55	10	85%	15%
2	100-150	228	7	97%	3%
3	150-200	284	5	98%	2%
4	200-250	151	5	97%	3%
5	250-300	127	7	95%	5%
6	300-350	165	7	96%	4%
7	350-400	99	3	97%	3%
8	400-450	91	5	95%	5%
9	450-500	42	1	98%	2%
10	500-550	35	4	90%	10%
11	550-600	34	3	92%	8%
12	600-650	14	0	100%	0%
13	650-700	18	0	100%	0%
Average				**96%**	**4%**

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table V

Synthetic Dataset Test Results

Bin #	Bin masses (da)	Mam Count	Non-Mam Count	Mam%	Non-Mam%
1	50-100	100	147	40%	60%
2	100-150	399	494	45%	55%
3	150-200	405	693	37%	63%
4	200-250	187	406	32%	68%
5	250-300	187	322	37%	63%
6	300-350	305	349	47%	53%
7	350-400	207	181	53%	47%
8	400-450	240	125	66%	34%
9	450-500	119	44	73%	27%
10	500-550	110	38	74%	26%
11	550-600	110	31	78%	22%
12	600-650	34	19	64%	36%
13	650-700	43	25	63%	36%
Average				**46%**	**54%**

IEEE Int Conf Comput Adv Bio Med Sci. Author manuscript; available in PMC 2015 October 05.