analytical chemistry

Article

pubs.acs.org/ac

iMet: A Network-Based Computational Tool To Assist in the Annotation of Metabolites from Tandem Mass Spectra

Antoni Aguilar-Mogas,[†] Marta Sales-Pardo,[†] Miriam Navarro,[‡,¶] Roger Guimerà,*[,†,§] and Oscar Yanes*[,‡,¶]

[†]Departament d'Enginyeria Química, Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain

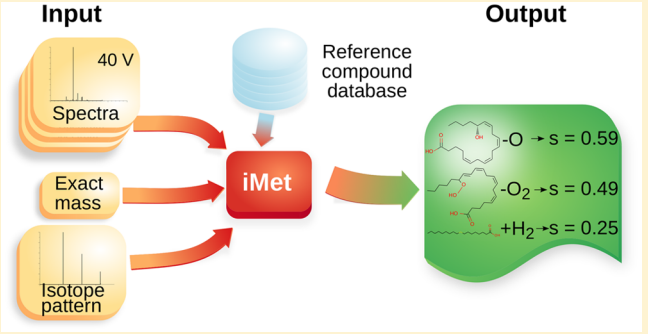
[‡]Metabolomics Platform, Department of Electronic Engineering (DEEEA), Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain

[¶]Biomedical Research Center in Diabetes and Associated Metabolic Disorders (CIBERDEM), Monforte de Lemos 35, 28029 Madrid, Spain

[§]Institució Catalana de Recerca i Estudis Avançats (ICREA), Lluís Companys 23, 08010 Barcelona, Catalonia, Spain

Supporting Information

ABSTRACT: Structural annotation of metabolites relies mainly on tandem mass spectrometry (MS/MS) analysis. However, approximately 90% of the known metabolites reported in metabolomic databases do not have annotated spectral data from standards. This situation has fostered the development of computational tools that predict fragmentation patterns *in silico* and compare these to experimental MS/MS spectra. However, because such methods require the molecular structure of the detected compound to be available for the algorithm, the identification of novel metabolites in organisms relevant for biotechnological and medical applications remains a challenge. Here, we present iMet, a computational tool that facilitates structural annotation of metabolites not described in databases. iMet uses MS/MS spectra and the exact mass of an unknown metabolite to identify metabolites in a reference database that are structurally similar to the unknown metabolite. The algorithm also suggests the chemical transformation that converts the known metabolites into the unknown one. As a proxy for the structural annotation of novel metabolites, we tested 148 metabolites following a leave-one-out cross-validation procedure or by using MS/MS spectra experimentally obtained in our laboratory. We show that for 89% of the 148 metabolites at least one of the top four matches identified by iMet enables the proper annotation of the unknown metabolites. To further validate iMet, we tested 31 metabolites proposed in the 2012−16 CASMI challenges. iMet is freely available at http://imet.seeslab.net.



The great success in the characterization of genes, transcripts, and proteins is a direct consequence of two factors. First, such molecules result from the concatenation of a small set of known monomers, namely, nucleotides and amino acids. Second, existing technologies and bioinformatic tools allow for the amplification and subsequent accurate characterization of the sequence of monomers. Metabolomics, in contrast, aims to identify and elucidate the structure of metabolites, which are not sequences of monomers and do not result from a residue-by-residue transfer of information. Instead, the large diversity of metabolites in living organisms results from a series of chemical transformations catalyzed mainly by enzymes.

The putative identification and structural annotation of metabolites in complex biological mixtures can be addressed by mass spectrometry (MS) and/or tandem mass spectrometry (MS/MS). The use of accurate mass of MS peaks for metabolite annotation[1−5] may entail high false positive rates because it only provides molecular formulas, despite recent implementations of network-based algorithms.[6,7] Therefore, as for the identification of proteins in proteomics, structural annotation of metabolites mostly relies on MS/MS analysis. However, predicting MS/MS spectra for metabolites is much more challenging than for peptides. As a result, annotating metabolites relies on their MS/MS spectra being present in reference databases.[8−11] In the simplest situation, the sample metabolite and its MS/MS spectra are already included in a reference library, so that the metabolite is annotated by matching both the intensities and the mass-to-charge (m/z) values of each fragment ion to values from pure standard metabolites in the spectral library. Unfortunately, only ∼10% of the known metabolites reported in databases such as HMDB[12]

Received: November 16, 2016
Accepted: February 21, 2017
Published: February 21, 2017

DOI: 10.1021/acs.analchem.6b04512
Anal. Chem. 2017, 89, 3474−3482

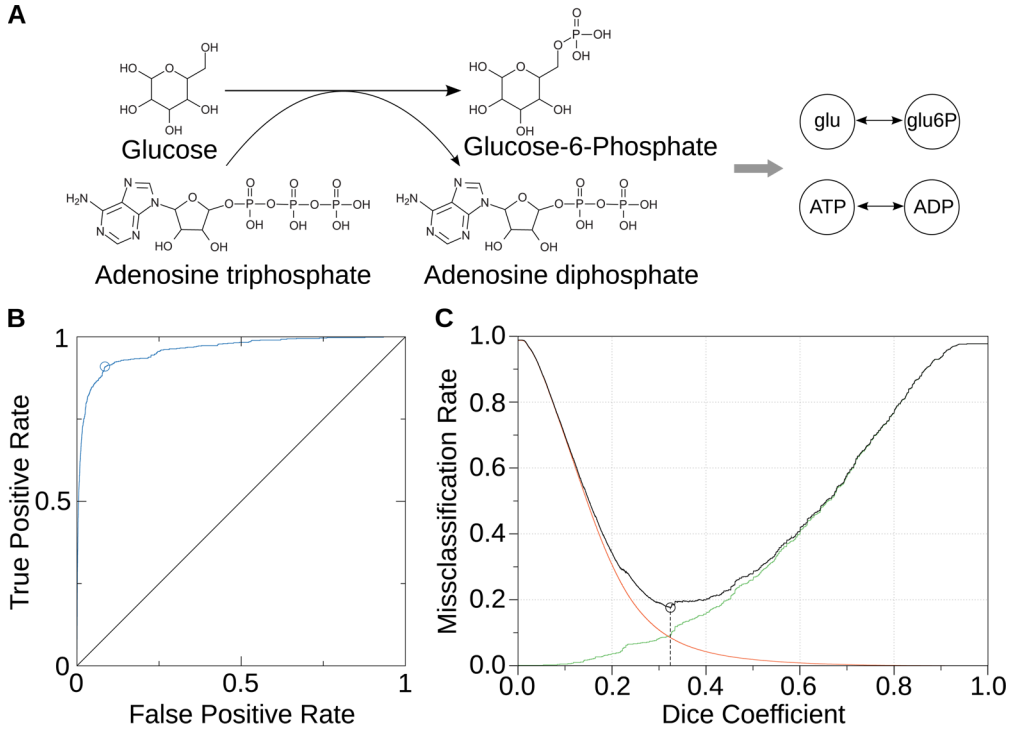


Figure 1. Neighbor metabolites. (A) An example of neighbor metabolites. Phosphorylation of glucose. Glucose (glu) is transformed into glucose-6-phosphate (glu-6P), while adenosine triphosphate (ATP) is dephosphorylated into adenosine diphosphate (ADP). Following the definition of RPs in KEGG, glu and glu-6P are one chemical transformation (phosphorylation) away from each other. The same applies to ATP and ADP. (B) ROC curve of the structural similarity of neighbor and non-neighbor metabolites on the basis of the D_c. The AUC is 0.96, indicating that neighbor metabolites have higher structural similarity than non-neighbor metabolites. The blue circle marks the maximum discrimination point. (C) False negative rate, i.e., ratio of RPs with a D_c below (green) a certain D_c value, and false positive rate, i.e., ratio of non-RPs with a D_c above (red) a certain value, as a function of that value. Black line corresponds to the sum of the other two curves. The black circle indicates the minimum of this curve, which corresponds to the maximum discrimination point.

and METLIN[13] have annotated spectral data.[14] Furthermore, chemical standards required to acquire MS/MS data are not available for other metabolites, and so, the size of MS/MS mass spectral libraries is not expected to grow significantly to solve the problem. To assist the structural annotation of this substantial percentage of known metabolites lacking MS/MS spectra in databases, efforts have emerged recently to heuristically predict MS/MS fragmentation patterns *in silico* and compare these to experimental MS/MS spectra.[15−23] Other methods use machine learning techniques[24,25] to predict the spectra. These methods require the structure of the detected compound to be available, from databases of known metabolites (e.g., HMDB, PubChem[26]), computationally predicted metabolite databases,[27] or uploaded structures by the user.

In the most challenging case, the sample metabolite is completely unknown; that is, the structure of the metabolite is not described. The description of novel metabolites in prokaryotic and eukaryotic organisms is precisely one of the next frontiers for metabolomics research.[28−33] Existing computational approaches to this problem only use neutral losses and characteristic fragment ions as signatures for unique chemical functional groups. These approaches have proved to be effective for classifying very specific lipid structures such as acyl-carnitines (e.g., fragments at m/z 85.0284 and 60.0808), glycerolipids, glycerophospholipids (e.g., fragment m/z 184.0730), and sphingolipids.[34−36] However, there is no tool that allows good structural annotation of metabolites when the correct molecular structure is not available.

To help in the annotation of metabolites that cannot be retrieved from databases of molecular structures, we have developed iMet. Its two only inputs are the electrospray ionization (ESI) quadrupole time-of-flight (Q-TOF) MS/MS spectra and the exact mass of an unknown metabolite. To increase the accuracy, the isotopic pattern of the intact unknown ion can be optionally supplied. Given these inputs, the algorithm identifies metabolites in a reference database that are likely to be structurally very similar to the unknown metabolite. Finally, iMet produces a list of candidates, ranked by their similarity to the unknown metabolite. The algorithm also suggests the chemical transformation that is most likely to separate each of the candidates from the unknown metabolite.

EXPERIMENTAL SECTION

Basic Principle of iMet. Metabolites can be represented as nodes in a network; two metabolites A and B are connected, that is, are neighbors, if one can obtain the chemical structure of B by a chemical transformation of A and vice versa (see Figure 1A). By a chemical transformation here, we mean the addition or removal of a moiety or a conformational change. By definition, neighbor metabolites are structurally more similar than a typical pair of non-neighbor metabolites, so that this structural similarity should be reflected in their MS/MS spectra because the fragmentation pattern of a metabolite highly depends on its chemical structure. Therefore, from the MS/MS spectrum of a metabolite that is not annotated in the network, a trained algorithm should be able to locate possible neighbors on the basis of spectral similarity.

DOI: 10.1021/acs.analchem.6b04512
Anal. Chem. 2017, 89, 3474−3482

To train one such classifier, we make use of the concept of reactant pair (RP) as defined in the KEGG database.[37] In KEGG, substrates and products of a known biochemical reaction are paired according to their chemical structure using graph theory[38] (as in Figure 1A). There are over 15 000 RPs defined in KEGG, including more than 9000 catalogued as being the "main" RP (the rest include generic reactions, symbolic reactions, and pairings with small molecules like oxygen, water, etc.). Because the metabolites forming a RP are structurally similar by definition, the two metabolites in a RP are neighbors in the network, so we use RPs as ground truth for the neighborhood. (Note, however, that not all neighbor metabolites are annotated as RPs in KEGG. This occurs, for example, when there is no described biochemical reaction that can transform one into the other, even though they are structurally very similar. Thus, the network of RPs in KEGG and our data set of neighbor metabolites is a subgraph of the full network of neighbor metabolites that potentially exist in nature.)

Construction of the Classifier. We used the R package "randomForest" for the classifier.[39] We trained the classifier with a data set that contains 825 RPs (see Table S3 for a complete listing of the RPs used). These are all the RPs for which we have MS/MS spectra. We also included in the training set an additional 49 175 randomly chosen pairs of metabolites. Therefore, we trained the algorithm with 50 000 pairs of metabolites, a set deliberately enriched with non-RPs. This enrichment reflects the reality of our database and of nature in general, that is, that any two random metabolites will not have, in general, a similar molecular structure. All of the spectra used in the training stage were taken from our database which includes spectra for 5060 metabolites (see "MS/MS database" section below). We need the training compounds to be in KEGG because we use KEGG RPs as our ground truth for the compound neighborhood.

The classifier uses the following features: (i) the cosine similarity between the MS/MS spectra of the two metabolites at all available collision energies; (ii) the mass difference between the two metabolites. The random forest[40] classifier has the advantage of automatically taking care of the nonmonotonic relationship between mass difference and probability of the neighborhood, as well as the complex nonlinear similarity patterns between MS/MS spectra at different collision energies. The classifier tries to predict whether the two metabolites are neighbors or not. In order to compute the cosine similarity, we discretize each spectrum in equal intervals of mass width δm. In this way, for each spectrum, we can construct an intensity vector $\mathbf{v}$ in which element v_i corresponds to the relative intensity of m/z values in the interval $[m_i, m_i + \delta m]$. (Note that we use δm = 0.01 Da, and we disregard relative intensity values below 1% of the highest m/z value.) Then, the cosine similarity c between spectra $\mathbf{v}$ and $\mathbf{u}$ is simply the dot product of the two vectors divided by the product of their norms.

$$c = \frac{\sum_i v_i u_i}{||\mathbf{v}||||\mathbf{u}||} \quad (1)$$

Our classification algorithm also takes into account experimental mass errors. Specifically, we introduced a shift in the exact mass of every metabolite of the training data set, changing its mass to a value randomly drawn from a Gaussian distribution centered around the exact mass of the metabolite and a standard deviation of 0.0025 Da. In this way, the algorithm correctly deals with the experimental error of the unknown target metabolite.

On the basis of the accurate m/z measurement of a protonated $(M + H)^+$ or deprotonated $(M - H)^-$ precursor ion of the unknown metabolite (mass error <0.005 Da), its MS/MS spectra, and its experimental isotopic distribution when available, the trained classifier yields a sorted list of candidate neighbors of the unknown metabolite, chosen from among the 5060 compounds included in our database. Moreover, iMet uses the most common chemical transformation between RPs to predict the unknown metabolite's chemical formula.

All in all, iMet outputs a sorted list of candidate neighbors of the unknown metabolite on the basis of mass difference and MS/MS spectral similarity. For every candidate, and given its mass difference with the unknown metabolite, iMet gives the chemical transformation (group of atoms) that converts the candidate into the unknown metabolite. The reliability of the prediction is given as a numerical score (s), whose value goes from 0 for the least reliable to 1 for the most reliable candidate.

iMet Step by Step. The general procedure followed by the algorithm is as follows (see the Supporting Information and Figure S1 for a detailed description of each step): (1) Obtain spectral similarities and mass differences between the unknown metabolite and each of the metabolites in the database. (2) Classify each of the metabolites in the database as candidate neighbors or non-neighbors of the unknown metabolite. (3) Determine the chemical transformation needed to transform each candidate neighbor into the unknown metabolite. (4) Prioritize those candidates that, aplying their assigned chemical transformation to their own chemical formula, yield the same final chemical formula for the unknown metabolite (formula consensus). (5) If provided, compare the isotope pattern of the unknown metabolite with the theoretical isotope pattern computed for each chemical formula proposed by each candidate. (6) Output the candidate neighbors ranked according to their score.

MS/MS Database. Our database is composed of 29 242 MS/MS spectra from 5060 different compounds obtained from the databases HMDB,[12] MassBank,[41] and METLIN[13] obtained with a Q-TOF instrument and at different collision energies and ionization modes. Although HMDB contains ∼45 000 compound entries, only ∼8% of those compounds have ESI MS/MS spectra. A similar percentage of the compounds in METLIN has this type of spectra.

RESULTS

Neighbor Metabolites Share Structural Similarities. A common way to compare chemical structures is by calculating the similarity between the fingerprints of two molecules. Fingerprints are representations of molecules that include in one object all the relevant molecular structure descriptors. The advantage of using molecular fingerprints is that they can be objectively compared by means of a similarity coefficient.[42] There exist different types of fingerprints and similarity coefficients, which makes it impossible to establish a universal criterion for structure comparison.[43] It is then necessary to find the combination of fingerprint type and similarity coefficient that best suits each particular problem. In our case, our aim is to discriminate RPs from non-RPs. After testing different combinations (see the Supporting Information and Table S1 for a detailed explanation), we found that the circular

DOI: 10.1021/acs.analchem.6b04512
Anal. Chem. 2017, 89, 3474−3482

fingerprint ECFP4[44] and the Dice coefficient[45] (D_c) had the largest discriminatory power.

To assess the structural similarity between neighbor metabolites, we used a subset of the database consisting of 3836 metabolites, including 825 RPs (due to computational constraints in terms of computation time and resources). We compared all the possible pairs of structures in this subset and computed the receiver operating characteristic (ROC) curve,[46] to check if structural similarity could be used to discriminate between RPs and non-RPs (Figure 1B). The area under the ROC curve (AUC statistic) is an overall measure of discriminatory power, which indicates how often RPs (and in general, neighbor metabolites) have a higher structural similarity than metabolites that are not RPs. The value of the AUC found in this case was 0.96, indicating that, in the vast majority of cases, two metabolites that are neighbors have a more similar structure than those that are not. We also found that 95% of the RPs have a D_c higher than 0.22. With the aim of establishing the best threshold in a scale of 0 to 1 to separate between RPs and non-RPs, we looked for the D_c value that minimizes the classification errors. To do so, we calculated the false positive rate (FPR) and the false negative rate (FNR) at each value of the D_c. In this context, the FPR corresponds to the proportion of non-RPs that have a higher D_c value than a given threshold, while the FNR represents the ratio of RPs that have a D_c value lower than that same threshold (Figure 1C). Combining these two curves, we obtain the missclassification ratio, that is, the ratio of pairs of metabolites that would be incorrectly classified if we used a certain value of their D_c to discriminate between RPs and non-RPs. We found that the missclassification ratio is minimal for a D_c value of 0.32, with a FPR of 0.09 and a FNR of 0.09, for a total missclassification ratio of 0.18 (highlighted in Figure 1B,C). By using this value as a threshold to separate between RPs and non-RPs, the classification error is minimum. This is also the point that maximizes the probability that a RP has a higher similarity than this value, while at the same time maximizing the probability that a non-RP has a lower value than this threshold (for a full probabilistic interpretation, see the Supporting Information).

Neighbor Metabolites Have Similar MS/MS Spectra. In order to numerically quantify the similarities between two MS/MS spectra, we used the cosine similarity, as this method is both time efficient and accurate (see the Supporting Information and Table S2 for a discussion). To validate the hypothesis that spectral similarity is indicative of the neighborhood in the network, we quantified to which extent metabolites that are neighbors have similar MS/MS spectra (Figure 2). To this end, we considered those metabolites in KEGG for which we had the experimental MS/MS spectra from public databases (1157 metabolites including 568 RPs for 10 V spectra, 1147 metabolites including 556 RPs for the 20 V spectra, and 1091 metabolites including 480 RPs for the 40 V spectra) and compared their spectra using the cosine similarity (Figure 2A,B).

We used the ROC curve to quantify the power of spectral similarity to distinguish pairs of metabolites that are RPs in KEGG (and, therefore, neighbors) from those that are not (Figure 2C–E). In this case, the area under the ROC curve indicates how often RP metabolites have a higher spectral similarity than metabolites that are not RPs. For the three collision energies 10, 20, and 40 V in negative ionization mode, we found $AUC_{10\ V} = 0.81$, $AUC_{20\ V} = 0.83$, and $AUC_{40\ V} = 0.80$, which indicates that the similarity between MS/MS spectra is

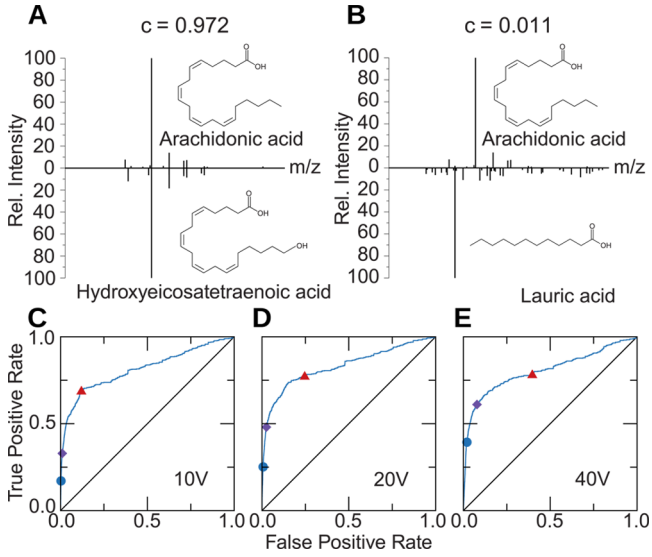


Figure 2. Similarity of MS/MS spectra discriminates between neighbor and non-neighbor metabolites. (A) MS/MS spectrum similarity for two neighbor metabolites (spectral similarity 0.972) and (B) for two non-neighbor metabolites (spectral similarity 0.011). (C–E) Classification power of the cosine similarity. We show the ROC curve for the cosine similarity when discriminating between RPs and non-RPs in KEGG for different collision energies (10, 20, and 40 V) in negative ionization mode, with a total area under the curve of (C) 0.81, (D) 0.83, and (E) 0.80, respectively. The three highlighted symbols correspond to cosine similarities of 0.5 (blue dot), 0.1 (purple diamond), and the first nonzero value (red triangle).

useful to identify neighbor metabolites. Comparing MS/MS spectra in positive ionization mode gave similar AUC values ($AUC_{10\ V} = 0.81$, $AUC_{20\ V} = 0.84$, and $AUC_{40\ V} = 0.81$). Note that these metrics quantify the discriminatory power of the spectral similarity when comparing only two MS/MS spectra with the same collision energy and ionization mode. Comparing points with the same value of spectral similarity in the three ROC curves, we observed that MS/MS spectra acquired at high collision energies have higher sensitivity but lower specificity and conversely for low collision energies. For example, the point at which the spectral similarity is 0.5 has a sensitivity of 0.1708 in 10 V, that increases to 0.2518 in 20 V and to 0.3938 in 40 V. The same point has a specificity of 0.9972 in 10 V, but it decreases to 0.9931 in 20 V and to 0.9797 in 40 V. We can find the same situation for the point with a spectral similarity of 0.1, with sensitivities of 0.3292, 0.4802, and 0.6104 and specificities of 0.9899, 0.9744, and 0.9214, respectively, for 10, 20, and 40 V (Figure 2C–E). This implies that trying to classify two metabolites as RPs or non-RPs by comparing their spectra obtained at high collision energies results in a highly conservative classification, discarding pairs that are actually neighbor metabolites (low specificity or TNR) but assuring that most of the metabolites classified as neighbor metabolites are real neighbors (high sensitivity or TPR). In contrast, spectral similarity becomes a poorer classification method at low collision energies: while most of the neighbor metabolites are correctly classified as such, some non-neighbor metabolites are also labeled as neighbors. Finally, the analysis reveals that information is usually nonredundant: some pairs of metabolites have high spectral similarity at high collision energies and low similarity at low energies, whereas for other pairs the opposite is true.

DOI: 10.1021/acs.analchem.6b04512
Anal. Chem. 2017, 89, 3474–3482

These results indicate that, indeed, spectral similarity at a fixed collision energy is predictive to some extent. As we show next, however, the predictive power of spectral similarity can be increased by considering spectra at different collision energies simultaneously and combining them with mass difference and, optionally, the isotopic pattern of the unknown metabolite (i.e., precursor ion).

Neighbor Metabolites Have Well-Defined Mass Differences and Chemical Transformations. To complement the information obtained from the spectral similarity, we study the differences in exact mass between neighbor metabolites (Figure 3). The mass difference between two metabolites corresponds to the mass of the group of atoms added to (or removed from) one of the metabolites to convert it into the other.

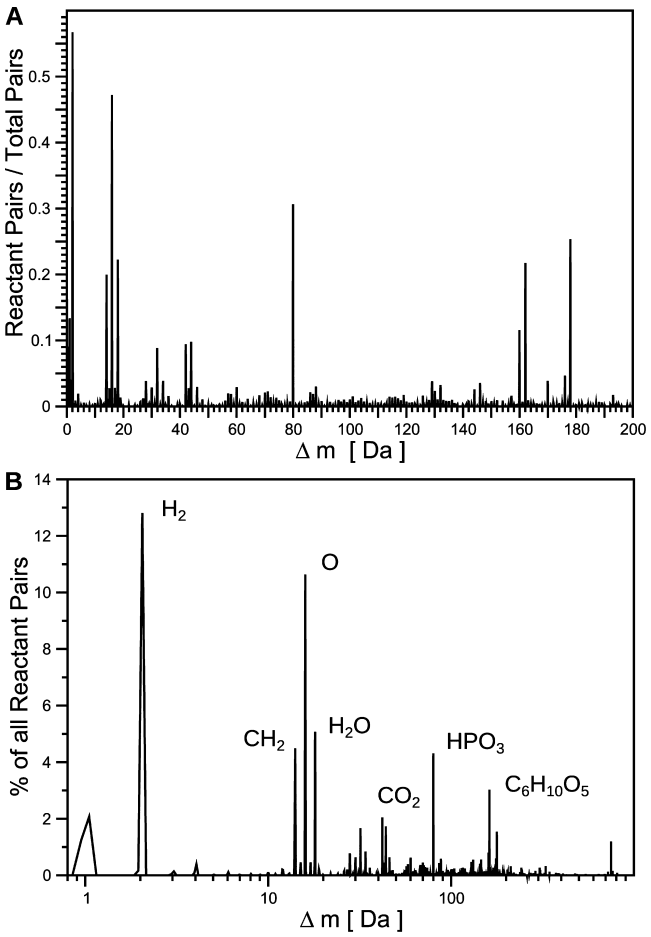


Figure 3. Ratio of neighborhood and of chemical transformation. (A) Fraction of RPs that have associated a mass difference within a specific interval. We constructed the figure using all compounds listed in KEGG, with bins of 0.01 Da. (B) Percentage of all RPs in KEGG with a certain mass difference. The most frequent mass differences correspond to well-defined moieties. The seven most frequent moieties (highlighted in the figure) account for 46% of all RPs.

As before, we take the mass difference, Δm, between KEGG RPs as ground truth. We considered 15 366 different metabolites including 7862 RPs. Calculating the mass difference for every pair of metabolites in our database and plotting the proportion of RPs for each value of the mass difference yields a curve that reflects the probability of two metabolites being neighbors given their mass difference (although specific systems may deviate from this general average pattern[47]).

As we show in Figure 3A, this curve displays well-defined maxima at specific Δm values. Therefore, it is much more likely that two metabolites are neighbors if their Δm corresponds to one of the maxima of the distribution in Figure 3A.

To understand what these Δm represent, we extracted 717 distinct chemical transformations (see Table S4 for a summary of the 200 most common). Since each chemical transformation implies a well-defined mass difference, the distribution of mass differences among RPs is localized around certain values that correspond to the most common interconversions of atoms (Figure 3B). For example, 10.6% of all RPs correspond to the net addition of an oxygen atom (Δm = 15.995 Da), 12.8% to the net addition of H_2 (Δm = 2.016 Da), and 4.3% to the addition of a phosphate group (Δm = 79.966 Da). In summary, a relatively small number of transformations account for a large number of the observed RPs.

Cross-Validation of iMet Using 148 Test Metabolites. To validate iMet, we run two cross-validation experiments, for a total of 148 different metabolites tested. The first cross-validation experiment consisted of 48 different metabolites whose spectra were taken experimentally in our lab. The second experiment consisted of a leave-one-out cross validation of 100 metabolites, taking their spectra directly from our database. We excluded the spectra of all tested metabolites from the training set and manually removed their entries from our database, effectively turning them into unknown compounds for the purpose of validation. We manually evaluated the output of the algorithm in terms of the distance in chemical transformations from each of the candidates to the test metabolite. We considered a candidate metabolite to be a neighbor of the test metabolite if they were 2 or fewer chemical transformations away. We also evaluated the performance of iMet by comparing the similarity of the chemical structure of each candidate proposed by iMet to the test metabolite. To do so, we used the ECFP4 fingerprints and the D_c as described in previous sections. We considered two metabolites (the unknown metabolite and the candidate output by iMet) to be structurally similar if the D_c of their molecular fingerprints was above 0.32 (see above). Although iMet is capable of finding an arbitrary number of candidate neighbors, we restricted our analysis to the top 4 candidates output by the algorithm (as ranked by their score). Additionally, we evaluated the quality of the first candidate output by iMet.

In the first validation experiment, we obtained in our laboratory MS/MS spectra of 48 metabolites in different conditions, for a total of 52 different tests as some metabolites are tested in both positive and negative ionization modes separately (all the test spectra can be found in Supporting File 1; see also Table S5 for cross-references in different databases of the test metabolites used). To ensure structural and biochemical diversity of tests, these include nucleotides and nucleosides, both natural and unnatural amino acids, vitamins, sphingolipids, polyamines and fatty acids, among others (see Table S6 for a complete listing of pathways covered by these tests). For these 48 metabolites, we ran iMet against our reference database (see "MS/MS database" section above) and evaluated the quality of each prediction. A total of 65% of the top 4 candidates proposed by iMet were neighbors of the test metabolite, meaning that they are 2 or less chemical transformations away from it. Out of these top 4 four candidates, 70% of the first candidates proposed by iMet were neighbors of the test metabolite. In 85% of the cases, iMet located at least one neighbor among the top four proposed

DOI: 10.1021/acs.analchem.6b04512
Anal. Chem. 2017, 89, 3474−3482

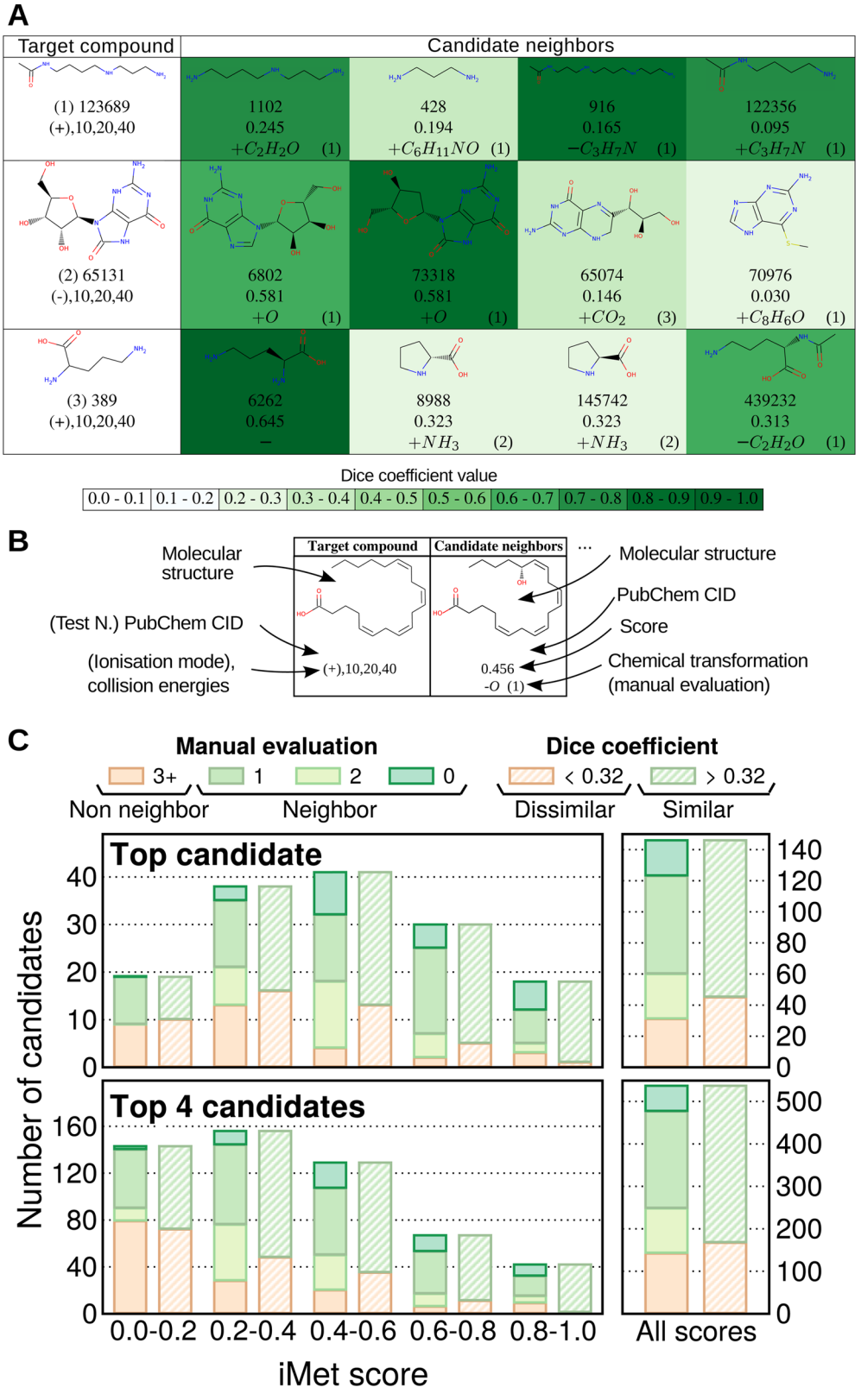


Figure 4. Cross-validation of iMet. (A) Results for three naturally occurring metabolites. The test metabolites are shown in the first column along with their PubChem CID (123689, *N*(8)-acetylspermidine; 65131, 8-hydroxyguanosine; 389, ornithine) and the ionization mode as well as the collision energies used to obtain the MS/MS spectra. The other columns contain the top four candidate neighbors ranked from highest to lowest score, along with their PubChem CID, the iMet score, the proposed chemical transformation, and its evaluation. Each candidate neighbor metabolite is colored according to the value of the D_c with the target metabolite, as shown in the colorbar. (B) Legend for the columns in (A). (C) Summary of validation for 148 test metabolites (see Figures S4–S6). We display the number of top candidate neighbors that are neighbors as well as candidates that are structurally similar to the test metabolite, as a function of the iMet score.

DOI: 10.1021/acs.analchem.6b04512
Anal. Chem. 2017, 89, 3474–3482

candidates. Evaluating the results in terms of structural similarity, we found that 78% of the top candidates identified by iMet were, indeed, structurally similar to the target, with the correct chemical transformations (Figure S4, see also Table S7 for complete results). For 91% of the cases, at least one of the top four candidates suggested by iMet was structurally similar to the target, and the proposed chemical transformation was also correct.

For the second cross-validation experiment, we used 100 randomly selected metabolites (see Table S5 for cross-references and Table S8 for the pathway coverage of these tests) whose MS/MS spectra we took directly from our reference MS/MS database. We followed a leave-one-out cross-validation procedure, so that each metabolite was tested individually by removing it from the database. In this validation, 78% of the top four candidates proposed by iMet were neighbors of the test metabolite. Out of the top 4 candidates, 83% of the first candidates proposed by iMet were neighbors of the test metabolite. In 92% of the cases, iMet was able to locate at least one neighbor of the test metabolite among the top four candidates. Taking into account the structural similarity between candidates and test metabolites, we found that 67% of candidate neighbors among the top 4 are structurally similar to the test metabolite ($D_c > 0.32$), and 65% of the first candidates were structurally similar to the unknown test metabolite. In 88% of the cases, the algorithm correctly predicted one of the top four candidates as being structurally similar to the test metabolite (see Figure S5 and Table S9 for the complete table of results).

Overall, combining both cross-validation experiments (the experimental validation and the leave-one-out validation), iMet was able to correctly identify a neighbor metabolite as the top candidate in 79% of the cases (Figure 4). In 89% of the cases, iMet proposed at least one neighbor among the top four candidates. Using the structure similarity criterion, iMet proposed a structurally similar metabolite as the top candidate in 69% of the cases. In 89% of the cases, at least one of the top four candidates was structurally similar to the test metabolite. Using the Tanimoto coefficient[48,49] to assess structural similarity yielded very similar results (see Figure S6). In 88% of the cases, the top formula proposed by iMet was the correct formula of the test metabolite.

CASMI Challenge. To simulate another scenario of metabolites not present in a database, we tested iMet using metabolites proposed in the Critical Assessment of Small Molecule Identification (CASMI) challenges from years 2012–2016. We downloaded the spectra of 31 different metabolites obtained using an ESI-QTOF mass spectrometer and that had PubChem Compound ID. Since iMet is designed to allow structural annotation of novel metabolites not present in databases, we tested the 31 metabolites in CASMI against our reference database of 5060 metabolites, which does not contain any of these CASMI metabolites. These tests were conducted without human intervention beyond downloading the spectra and using them as inputs for iMet. For these 31 metabolites, 26% of the top candidates suggested by iMet were neighbors of the test metabolite, and in 32% of the cases, iMet was able to locate at least one neighbor of the test metabolite. In terms of structural similarity, 42% of the top candidates were structurally similar to the test metabolite, and in 48% of the results, iMet was able to locate at least one structurally similar metabolite. In 48% of the cases, the top formula proposed by iMet was the correct formula of the test metabolite.

It should be noted that 26 out of the 31 CASMI metabolites were obtained by using other collision energies than those used in the training set (10, 20, or 40 V). We tested them nevertheless to evaluate the performance of iMet when confronted with spectra obtained using inaccurate experimental data (for example, we introduced spectra obtained at 25 V as if they were obtained at 20 V or 35 V spectra as if they were 40 V). For these 26 metabolites, 27% of the top candidates suggested by iMet were neighbors of the test metabolite (42% of them were structurally similar), and in 35% of the results, iMet was able to locate at least one neighbor among the top four candidates (in 50% of the cases, at least one of the top four candidates was structurally similar to the test metabolite). These results (depicted in Table S10) suggest that iMet does not substantially decrease its accuracy when using slightly different collision energies as inputs.

DISCUSSION

The structural annotation of novel metabolites in relevant organisms is one of the next frontiers in metabolomics research. We have built iMet on the grounds that many of the metabolites that remain to be discovered and characterized are probably chemically related and therefore structurally similar to the existing ones in databases such as HMDB or genome-based metabolic reconstructions,[50,51] typically by the addition, removal, or modification of a single group. iMet is intended to provide key information, such as the molecular formula of the novel compound, neighbor (structurally similar) compounds, and the moiety (if necessary) to transform the known neighbors into the unknown. iMet does not propose structures *de novo*, but rather, it provides chemical information for organic chemists to propose candidate chemical structures based on chemical knowledge.[28]

We have systematically demonstrated that neighbor metabolites have similar MS/MS spectra, and furthermore, the MS/MS spectral similarity has enough discriminatory power to distinguish between neighbor metabolites from non-neighbor metabolites. Our cross-validation and the CASMI challenges demonstrate that iMet is only limited by the experimental MS/MS library against which the unknown metabolite is compared and to a lesser extent the space of chemical transformations. In particular, iMet will fail to annotate correctly the neighbor metabolite when no structurally similar metabolites are present in the reference database. With databases having MS/MS spectra for only 8–10% of their compounds,[14] this may seem to be a serious limitation. However, the reasonably high percentage of well-annotated compounds in our results and the structural properties of metabolic networks suggests otherwise, anticipating that as the number of MS/MS spectra from known metabolites rises in public databases so will the predicting power of iMet. Regarding the space of chemical transformations, our network of RPs is restricted to those biochemical reactions described in the KEGG database, which does not account for all chemical transformations occurring at any biological system. The KEGG is, however, to our knowledge the only database that systematically shows paired substrates and products according to their structure transformations using graph theory. The significance of using this information is that we can compute the probability of two metabolites being neighbors on the basis of the mass difference between them, without taking into consideration other attributes such as their chemical structures,[52] functional groups, chemical reactivity, or metabolic pathways.[53] Rather, iMet uses

DOI: 10.1021/acs.analchem.6b04512
Anal. Chem. 2017, 89, 3474–3482

a large set of chemical transformations described in biological systems to propose chemical formulas and structures by adding (or removing) a group of atoms to a known chemical structure. This concept of chemical transformation is similar to that used previously by other computational approaches aiming to reduce the ambiguity in metabolite annotation.[1−3,53,54] However, these previous approaches, that do not make use of MS/MS data, require that the interrogated metabolite falls into a known metabolic pathway, and its chemical formula and structure must be known and described in a database.[1−3,54]

CONCLUSIONS

Despite the existence of different computational tools that predict fragmentation patterns *in silico*, none of these algorithms are designed to annotate metabolites for which there are no chemical structures available. iMet has the potential to fill this gap. Our algorithm has proven itself to be a valuable tool as a stand-alone application. In terms of MS/MS information, we have systematically demonstrated that MS/MS spectral similarity has enough discriminatory power to distinguish metabolites that are one chemical transformation away from each other. The cross-validation and CASMI challenge results demonstrate that iMet is mainly limited by the experimental MS/MS library against which the unknown metabolite is compared. The predicting power of the algorithm will increase as the number of MS/MS spectra from known metabolites in the iMet reference database increases. This version of the algorithm is intended to provide the molecular formula of the unknown compound, structurally similar compounds, and the moiety to transform the known neighbors into the unknown, for organic chemists to propose candidate chemical structures based on chemical knowledge. Future work will head toward generating candidate chemical structures of the unknown metabolite.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b04512.

Spectra of the test metabolites used in the experimental cross-validation (ZIP)

iMet step by step, comparison of simple cosine and spectral alignment methods, elucidation of the theoretical isotope pattern, general procedure followed by the iMet algorithm, fingerprints and similarity coefficients, AUCs for fingerprints and similarity coefficients, similarity threshold, cumulative distribution functions for the Tanimoto coefficient, comparison of cosine and spectral alignment methods, probability of discrimination of similarity coefficients, AUCs for spectral alignment and cosine methods, scores of the experimental validation, scores of the leave-one-out cross-validation, score comparison between Dice and Tanimoto coefficients, RPs in the training set, net atomic additions, cross-references of test metabolites, pathways represented in the experimental tests, experimental validation, pathways represented in the leave-one-out cross-validation tests, and leave-one-out cross-validation and CASMI challenges 2012−2016 approximate collision energies (PDF)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: roger.guimera@urv.cat.

*E-mail: oscar.yanes@urv.cat.

ORCID

Antoni Aguilar-Mogas: 0000-0002-5952-3629

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Spanish Ministerio de Economía y Comptetitividad [SAF2011-30578 and BFU2014-57466 to O.Y. and FIS-2013-47532-C3 to A.A.-M., M.S.-P., and R.G.], the James S. McDonnell Foundation [220020228], and the European Union [PIRG-GA-2010-277166 to R.G., PIRG-GA-2010-268342 to M.S.-P., and FET-317532-MULTIPLEX to M.S.-P. and R.G.]. The authors thank Mr. Manuel Miranda for his help in the web implementation of the tool.

REFERENCES

(1) Breitling, R.; Pitt, A. R.; Barrett, M. P. *Trends Biotechnol.* **2006**, *24*, 543−548.

(2) Rogers, S.; Scheltema, R. A.; Girolami, M.; Breitling, R. *Bioinformatics* **2009**, *25*, 512−518.

(3) Weber, R. J. M.; Viant, M. R. *Chemom. Intell. Lab. Syst.* **2010**, *104*, 75−82.

(4) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84*, 283−289.

(5) Brown, M.; Wedge, D. C.; Goodacre, R.; Kell, D. B.; Baker, P. N.; Kenny, L. C.; Mamas, M. A.; Neyses, L.; Dunn, W. B. *Bioinformatics* **2011**, *27*, 1108.

(6) Pirhaji, L.; Milani, P.; Leidl, M.; Curran, J.; Avila-Pacheco, T.; Clish, C. B.; White, F. M.; Saghatelian, A.; Fraenkel, E. *Nat. Methods* **2016**, *13*, 770−776.

(7) Li, S.; Park, Y.; Duraisingham, S.; Strobel, F. H.; Khan, N.; Soltow, Q. A.; Jones, D. P.; Pulendran, B. *PLoS Comput. Biol.* **2013**, *9*, e1003123.

(8) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263−269.

(9) Rojas-Cherto, M.; Peironcely, J. E.; Kasper, P. T.; van der Hooft, J. J. J.; de Vos, R. C. H.; Vreeken, R.; et al. *Anal. Chem.* **2012**, *84*, 5524−5534.

(10) Nikolskiy, I.; Mahieu, N. G.; Chen, Y. J.; Tautenhahn, R.; Patti, G. J. *Anal. Chem.* **2013**, *85*, 7713−7719.

(11) Tautenhahn, R.; Cho, K.; Uritboonthai, W.; Zhu, Z.; Patti, G. J.; Siuzdak, G. *Nat. Biotechnol.* **2012**, *30*, 826−828.

(12) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; et al. *Nucleic Acids Res.* **2013**, *41*, D801−D807.

(13) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; et al. *Ther. Drug Monit.* **2005**, *27*, 747−751.

(14) Vinaixa, M.; Schymanski, E. L.; Neumann, S.; Navarro, M.; Salek, R. M.; Yanes, O. *TrAC, Trends Anal. Chem.* **2016**, *78*, 23−35.

(15) Heinonen, M.; Shen, H.; Zamboni, N.; Rousu, J. *Bioinformatics* **2012**, *28*, 2333−2341.

(16) Menikarachchi, L. C.; Cawley, S.; Hill, D. W.; Hall, L. M.; Hall, L.; Lai, S.; et al. *Anal. Chem.* **2012**, *84*, 9388−9394.

(17) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. *J. Cheminf.* **2016**, *8*, 3.

(18) Gerlich, M.; Neumann, S. *J. Mass Spectrom.* **2013**, *48*, 291−298.

(19) Li, L.; Li, R.; Zhou, J.; Zuniga, A.; Stanislaus, A.; Wu, Y.; et al. *Anal. Chem.* **2013**, *85*, 3401−3408.

(20) Ridder, L.; van der Hooft, J. J. J.; Verhoeven, S.; de Vos, R. C. H.; Vervoort, J.; Bino, R. J. *Anal. Chem.* **2014**, *86*, 4767−4774.

(21) Tsugawa, H.; Kind, T.; Nakabayashi, R.; Yukihira, D.; Tanaka, W.; Cajka, T.; Saito, K.; Fiehn, O.; Arita, M. *Anal. Chem.* **2016**, *88*, 7946−7958.

DOI: 10.1021/acs.analchem.6b04512
Anal. Chem. 2017, 89, 3474−3482

(22) Verdegem, D.; Lambrechts, D.; Carmeliet, P.; Ghesquière, B. *Metabolomics* **2016**, *12*, 98.
(23) Zhou, J.; Weber, R. J. M.; Allwood, J. W.; Mistrik, R.; Zhu, Z.; Ji, Z.; Chen, S.; Dunn, W. B.; He, S.; Viant, M. R. *Bioinformatics* **2014**, *30*, 581.
(24) Dührkop, K.; Shen, H.; Meusel, M.; Rousu, S.; Böcker, M. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 12580−12585.
(25) Allen, F.; Pon, A.; Wilson, M.; Greiner, R.; Wishart, D. *Nucleic Acids Res.* **2014**, *42*, W94−W99.
(26) Bolton, E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. *Annu. Rep. Comput. Chem.* **2008**, *4*, 217−240.
(27) Jeffryes, J. G.; Colastani, R. L.; Elbadawi-Sidhu, M.; Kind, T.; Niehaus, T. D.; Broadbelt, L. J.; Hanson, A. D.; Fiehn, O.; Tyo, K. E. J.; Henry, C. S. *J. Cheminf.* **2015**, *7*, 44.
(28) Kalisiak, J.; Trauger, S. A.; Kalisiak, E.; Morita, H.; Fokin, V. V.; Adams, M. W. W.; et al. *J. Am. Chem. Soc.* **2009**, *131*, 378−386.
(29) Jansen, R. S.; Addie, R.; Merkx, R.; Fish, A.; Mahakena, S.; Bleijerveld, O. B.; Altelaar, M.; IJlst, L.; Wanders, R. J.; Borst, P.; et al. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 6601−6606.
(30) Serhan, C. N.; Clish, C. B.; Brannon, J.; Colgan, S. P.; Chiang, N.; Gronert, K. *J. Exp. Med.* **2000**, *192*, 1197−1204.
(31) Ariyanayagam, M. R.; Oza, S. L.; Mehlert, A.; Fairlamb, A. H. *J. Biol. Chem.* **2003**, *278*, 27612−27619.
(32) Cohen, L. J.; Kang, H.-S.; Chu, J.; Huang, Y.-H.; Gordon, E. A.; Reddy, B. V. B.; Ternei, M. A.; Craig, J. W.; Brady, S. F. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, E4825−E4834.
(33) Kell, D. B.; Oliver, S. G. *Metabolomics* **2016**, *12*, 148.
(34) Kind, T.; Liu, K.; Lee, D.; DeFelice, B.; Meissen, J.; Fiehn, O. *Nat. Methods* **2013**, *10*, 755−758.
(35) Ma, Y.; Kind, T.; Yang, D.; Leon, C.; Fiehn, O. *Anal. Chem.* **2014**, *86*, 10724−10731.
(36) Lynn, K.; Cheng, M.; Chen, Y.; Hsu, C.; Chen, A.; Lih.; et al. *Anal. Chem.* **2015**, *87*, 2143−2151.
(37) Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. *Nucleic Acids Res.* **2004**, *32*, D277−D280.
(38) Hattori, M.; Okuno, Y.; Goto, S.; Kanehisa, M. *J. Am. Chem. Soc.* **2003**, *125*, 11853−11865.
(39) Liaw, A.; Wiener, M. *R. News* **2002**, *2*, 18−22.
(40) Breiman, L. *Machine Learning* **2001**, *45*, 5−32.
(41) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; et al. *J. Mass Spectrom.* **2010**, *45*, 703−714.
(42) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. *J. Med. Chem.* **2014**, *57*, 3186−3204.
(43) Stumpfe, D.; Bajorath, J. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 260−282.
(44) Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(45) Dice, L. R. *Ecology* **1945**, *26*, 297−302.
(46) Hanley, J. A.; McNeil, B. J. *Radiology* **1982**, *143*, 29−36.
(47) Morreel, K.; Saeys, Y.; Dima, O.; Lu, F.; Van de Peer, Y.; Vanholme, R.; Ralph, J.; Vanholme, B.; Boerjan, W. *Plant Cell* **2014**, *26*, 929−945.
(48) Tanimoto, T. T. *An elementary mathematical theory of classification and prediction*; IBM Corp.: NY, 1958.
(49) Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(50) Caspi, R.; Altman, T.; Billington, R.; Dreher, K.; Foerster, H.; Fulcher, C. A.; Holland, T. A.; Keseler, I. M.; Kothari, A.; Kubo, A.; et al. *Nucleic Acids Res.* **2014**, *42*, D459−D471.
(51) Thiele, I.; Swainston, N.; Fleming, R. M. T.; Hoppe, A.; Sahoo, S.; Aurich, M. K.; Haraldsdottir, H.; Mo, M. L.; Rolfsson, O.; Stobbe, M. D.; et al. *Nat. Biotechnol.* **2013**, *31*, 419−425.
(52) Hamdalla, M. A.; Rajasekaran, S.; Grant, D.; Măndoiu, I. I. *J. Chem. Inf. Model.* **2015**, *55*, 709−718.
(53) Li, S.; Park, Y.; Duraisingham, S.; Strobel, F. H.; Khan, N.; Soltow, Q. A.; et al. *PLoS Comput. Biol.* **2013**, *9*, e1003123.
(54) Gipson, G. T.; Tatsuoka, K. S.; Sokhansanj, B. A.; Ball, R. J.; Connor, S. C. *Metabolomics* **2008**, *4*, 94.

DOI: 10.1021/acs.analchem.6b04512
Anal. Chem. 2017, 89, 3474−3482