Journal of
CHEMICAL INFORMATION
AND MODELING

Article

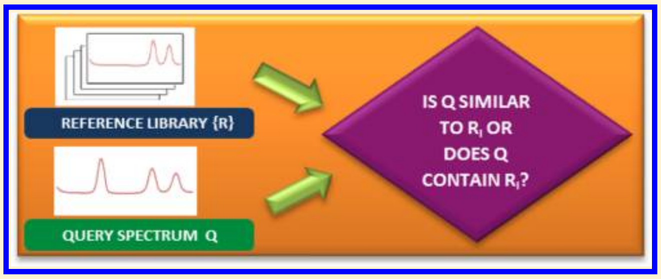
pubs.acs.org/jcim

Improving Spectral Library Search by Redefining Similarity Measures

Ankita Garg,* Catherine G. Enright,* and Michael G. Madden*

College of Engineering and Informatics, National University of Ireland, Galway, Ireland

Ⓢ *Supporting Information*



ABSTRACT: Similarity plays a central role in spectral library search. The goal of spectral library search is to identify those spectra in a reference library of known materials that most closely match an unknown query spectrum, on the assumption that this will allow us to identify the main constituent(s) of the query spectrum. The similarity measures used for this task in software and the academic literature are almost exclusively metrics, meaning that the measures obey the three axioms of metrics: (1) minimality; (2) symmetry; (3) triangle inequality. Consequently, they implicitly assume that the query spectrum is drawn from the same distribution as that of the reference library. In this paper, we demonstrate that this assumption is not necessary in practical spectral library search and that in fact it is often violated in practice. Although the reference library may be constructed carefully, it is generally impossible to guarantee that all future query spectra will be drawn from the same distribution as the reference library. Before evaluating different similarity measures, we need to understand how they define the relationship between spectra. In spectral library search, we often aim to find the constituent(s) of a mixture. We propose that, rather than asking which reference library spectra are *similar to* the mixture, we should ask which of the reference library spectra are *contained in* the given query mixture. This question is inherently asymmetric. Therefore, we should adopt a nonmetric measure. To evaluate our hypothesis, we apply a nonmetric measure formulated by Tversky [*Psychol. Rev.* **1977**, *84*, 327−352] known as the Contrast Model and compare its performance to the well-known Jaccard similarity index metric on spectroscopic data sets. Our results show that the Tversky similarity measure yields better results than the Jaccard index.

1. INTRODUCTION

Spectral library search is often used to identify the main constituent(s) of an unknown substance based on its spectrum. The aim is to find the closest matches (or hits) to the given query spectrum on the basis of its computed similarity measure to each entry in the reference library, sorted in order of decreasing similarity. To understand this process fully, it is important to understand how similarity defines the relationship between two objects and its consistency with human perceptions of similarity. The choice of similarity measure is fundamentally important, and it is one of the main challenges to determine which similarity measure should be selected for the task of interest.

The ideal similarity measure needs to be relevant to the given task of interest.[1] There are numerous similarity measures in the literature, most satisfying the metric axioms: minimality, symmetry, and triangle inequality[2] (which will be explained below in section 1.1). Existing similarity measures (even if they are not metrics) implicitly assume that the query/test spectra are drawn from the same distributional space as that of the reference library. These measures are best at finding exact matches (or hits). But in many practical spectroscopic applications, the objective is to find close or related matches to the query spectrum rather than just finding exact matches. The query spectrum may come from a mixture of substances which may be present in the reference library either as separate components in pure form or as a mixture with proportions that might not necessarily be the same as those in the query.[3] Therefore, the query spectra may not be drawn from the same distributional space as that of reference library. As a result, the metric axioms will be violated; they may also be unnecessary (as detailed in section 2).

In this paper, we propose that in such scenarios, what is needed is a *Contains* concept, as opposed to *similarity metrics*. Similarity metrics focus on finding the reference library spectra most or exactly *similar to* the given query spectrum. On the other hand, a Contains concept shifts the focus toward finding the reference library spectra that are *contained in* the given query spectrum. This is more directly relevant to the task of identifying constituent(s) reference spectra accurately.

The next subsection gives an overview of the metric axioms and explains how these metric axioms are violated in human similarity judgment. Section 2 discusses whether these metric axioms are needed in spectral library search and show examples of where they are violated in practice. Section 3 describes the proposed Contains concept. In section 3.1, to evaluate our hypothesis, we adopt Tversky's contrast model and compare its performance with that of the well-known Jaccard similarity index metric. We choose the Jaccard metric for this work because it is widely used, and we choose Tversky's contrast model because (as will be discussed in section 3.1), it is a direct

Received: February 10, 2015
Published: April 22, 2015

DOI: 10.1021/acs.jcim.5b00077

generalization of the Jaccard metric. The Tversky measure was proposed in the context of reflecting human intuitions of similarity. We apply it here in the new context of spectral library search using spectroscopic data sets.

Section 4 describes the data sets, experimental methodology, and performance measure used to evaluate the results. In addition, we also analyze the commonly used similarity measures to show that they obey the symmetry axiom. In section 5 the results are discussed. Finally, in section 6 we draw conclusions. It is concluded from the results that a Contains concept improves spectral library search performance in comparison to commonly used similarity metrics and, also, it captures similarity in the manner it is perceived.

1.1. Metric Axioms.

When considering measurements of similarity and dissimilarity, the *metric axioms* (listed below) define whether a measurement can be considered to be a geometric distance metric. Geometric measurements of similarity are the most influential approaches used and have a long and successful history in matching data (see, for example, the work of Shepard[5,6] for an overview). These methods are based on the assumption that objects are represented as points in some coordinate space, so that the metric distance between the points reflects the similarity between the two objects.[7] Similarity can be derived from a distance function in this space, using an inverse relationship: the shorter the distance between two objects, the more similar they are. They are exemplified by multidimensional scaling (MDS) models.[8] The general form of an equation for geometric distance between two objects is[9]

$$D(i, j) = [\sum_{k=1}^{n} |X_{ik} - X_{jk}|^{r}]^{1/r} \tag{1}$$

where X_{ik} is the value of Object i on dimension k, X_{jk} is the value of Object j on dimension k, n is the number of dimensions, and the value of r defines the distance metric. When $r = 1$, $D(i, j) = \sum|X_{ik} - X_{jk}|$, it is called the City-block distance or the Manhattan distance. When $r = 2$, $D(i, j) = [\sum| X_{ik} - X_{jk}|^{2}]^{1/2}$, it is known as Euclidean distance, which is widely used in chemometrics.[2] The best value of r generally depends on the nature of particular objects and also on the objective in making the comparisons.[10,11]

The geometric methods of similarity are based on the following three metric axioms, given objects A, B, and C in the multidimensional space and a metric distance function D:

1. Minimality: The distance from Object A to Object B will be 0 if they are identical; otherwise, it will be positive, i.e. $D(A, B) \geq D(A, A) = 0$.
2. Symmetry: The distance from Object A to Object B is the same as the distance from Object B to Object A, i.e. $D(A, B) = D(B, A)$.
3. Triangle Inequality: The sum of distances from A to B and B to C will always be greater than or equal to distance from A to C, i.e. $D(A, B) + D(B, C) \geq D(A, C)$. As mentioned in above point, distances are symmetric so the triangle inequality definition can also be written as $D(B, A) + D(B, C) \geq D(A, C)$ which is used by Gower and Legendre.[12]

1.2. Limitations of the Metric Axioms.

Some limitations of the above axioms have been addressed by several critics, specifically in the psychology community where the goal is often to model human understanding of similarity. Nonetheless, measurements based on the metric axioms dominate and are widely used.[4,7,13] Attneave is one of the early researchers who noticed some weaknesses of the geometric methods of similarity.[13] Later, Tversky[4] provided empirical results that seem to be in conflict with each of the metric axioms, again in the context of modeling human intuition of similarity. He also suggested an alternative theoretical approach to similarity, based on feature matching known as the Contrast Model (explained in section 3.1). Shepard[8] also criticized the metric requirements of the geometric measures.

Minimality may be violated where an object is identified as another object more frequently than it is identified as itself. Tversky[4] cited that the probability of identifying two identical objects as the same is not constant for all objects, and if these identification probabilities are treated as a similarity measure, then it violates minimality.

The symmetry axiom states that the similarity value is the same in both directions, i.e., from Object A to B and from B to A. Tversky[4] cited various examples explaining how the direction matters in similarity and the symmetric assumption does not always hold in human intuition; for example, he stated that how in general it is said that "the son resembles the father" rather than "the father resembles the son" and "an ellipse is like a circle" rather than "circle is like an ellipse". In his study of similarity of countries with 21 pairs of countries as objects, he found that participants mostly chose the less prominent country as the subject and the more prominent country as the referent. Hence, the asymmetric similarity occurs when an object with more salient features is judged as less similar to an object with less salient features than vice versa.

The triangle inequality states that when Object A is similar to Object B, and Object B is similar to Object C; then Object A is similar to Object C. Tversky[4] has noted that the triangle inequality cannot be formulated in ordinal terms, and it cannot be readily refuted even with interval data. Its violation, however, can be exemplified using three objects A, B, and C; when A (e.g., lamp) and B (moon) share an identical feature (light), and B (moon), and C (ball) share an identical feature (shape), but A and C do not share any feature in common.[14]

Section 2 discusses the violation of the metric axioms and their necessity in spectral library search. It also proposes a Contains concept as an alternative to similarity metrics.

2. METRIC AXIOMS IN SPECTRAL LIBRARY SEARCH

The basic operation of library search is to compare an unknown query item q with known items in the reference library $L = \{l_1, l_2, ..., l_n\}$, to find the best matches (or hits) sorted in the order of decreasing similarity (or increasing distance). Almost always, a distance metric or similarity metric is used. It is computed for a query item q to each reference library item, $D(q, l_i)$ or $S(q, l_i)$ where i is from $1 \rightarrow |L|$, D is the distance metric and S is the similarity metric.

Distance and similarity metrics are strongly related and these terms are often used interchangeably, though strictly speaking they are reciprocal concepts i.e. when distance decreases, similarity increases and vice versa. A distance value is typically in the range $[0-\infty)$, with value 0 corresponding to two identical objects, while similarity values are typically bounded in a range such as $[0-1]$, with the maximum value corresponding to two identical objects. For example, a commonly used transform from a distance to a distance-based similarity metric is

$$S(a, b) = \frac{1}{(1 + D(a, b))} \tag{2}$$

Note in this paper we use the term similarity metrics, as it is commonly used, to refer to distance measures that satisfy metric axioms. By definition, metric distances must be symmetric and obey the triangle inequality, as defined above. Similarities are also symmetric but do not obey the triangle inequality: it can easily be shown that if a distance metric obeys the triangle inequality, the corresponding similarity metric from eq 2 will not.

Metrics are widely used for searches even when queries are not drawn from the same population as the reference library. If there is no distributional difference between the reference library and the queries, it is equally likely that an item could be in the library set or in the query set, so a symmetric measure makes sense. In practice, however, this assumption is often violated. Objects to be included in reference libraries are often carefully selected and may not be drawn from the same distribution as the query set. For example, consider a reference library with four pure spectra p_1, p_2, p_3, and p_4 and a ternary query mixture $p_1 + p_2 + p_3$. When the similarity of the query mixture to pure spectrum p_1 is measured, the similarity coefficient reduces because of the presence of extra features (due to p_2, p_3). We want these extra features to make a small, or no, contribution to the similarity value as the user is interested in finding the reference library spectra contained in the query spectrum. This is an asymmetric relationship. In such scenarios, we propose to use a Contains concept; defined in section 3.

While the triangle inequality has a clear mathematical definition that is easily verified, it does not always map well to real-world intuitions of similarity. According to Tversky[4] the triangle inequality axiom implies that if Object A is quite similar to Object B, and Object B is quite similar to Object C, then Objects A and C cannot be much dissimilar to each other. For example, consider three mixtures $M_1(p_1 + p_2)$, $M_2(p_2 + p_3)$, and $M_3(p_3 + p_4)$. The mixture M_1 is similar to mixture M_2 (because of the presence of p_2); mixture M_2 is similar to mixture M_3 (because of the presence of p_3); but mixture M_3 is not similar to mixture M_1. This indicates that the triangle inequality is violated.

From the discussion above, it is concluded that the symmetry and the triangle inequality can be violated in spectral library search. We also note that the minimality axiom is applicable as we would expect the maximum similarity coefficient for identical spectra.

3. CONTAINS CONCEPT

Our proposed Contains concept modifies the focus of library matching. Instead of seeking to identify the reference library spectra that are similar to the given query spectrum, it focuses on finding the reference library spectra contained in the given query mixture. Therefore, it will take into account the inherent asymmetry. This will be useful in spectral library search in real applications. Consider the example mentioned above, the Contains concept will seek to find which of the reference library spectra p_1, p_2, p_3, and p_4 are contained in a given mixture spectrum $p_1 + p_2 + p_3$.

On the other hand, when we use similarity metrics, the question is whether the query item is similar to a reference library item. The validity of this question comes into doubt in spectral library search; we seek to determine the presence of a particular substance in a mixture, as a mixture will contain additional features because of the presence of the other constituents. By using a Contains concept, we focus more on the features present in the reference library item than on the additional features present in the query item. This way, we get a better similarity value.

The relationship between the Contains concept and the metric axioms is as follows:

1. It satisfies the minimality axiom. Two identical objects will, obviously contain each other.
2. It is directional, as it depends on whether we ask if a query item contains any of the reference library items or if any reference library item contains the query item. As a result, the similarity measure from the query item to the reference library item will not be same as the similarity measure from the reference library item to the query item. Specifically, we propose to give more importance to the features present in reference library. For example, when we are searching a spectral library for a given query item, which can be a pure or a mixture of two or more substances, we will be more concerned with the features present in a spectral library reference item.
3. The Contains concept considers asymmetric similarity, therefore, direction must be considered when assessing triangle inequality. A directional triangle inequality holds, for example, if Object A contains Object B, and Object B contains Object C, then Object A contains Object C, i.e., $S(A \rightarrow C) \leq S(A \rightarrow B) + S(B \rightarrow C)$. However, when direction is not considered triangle inequality can be violated easily. Taking the example mentioned above, $S(A \rightarrow C)$ is not less than or equal to $S(B \rightarrow A) + S(C \rightarrow B)$. The reason for this inequality is that Object B does not contain Object A and Object C does not contain Object B; rather, Object B is contained by Object A and Object C is contained by Object B.

Some researchers[3,15−18] have identified the presence of inherent asymmetry in chemical similarity measures. Khan and Madden developed a new distance measure *Modified Euclidean* for spectral library search, which computes the distance between reference and query spectrum using asymmetric weights.[3] In Modified Euclidean, they assigned weights according to the amplitude of intensity of query spectrum as compared to reference spectrum and to presence/absence of peaks at a given wavenumber in the query spectrum as compared to the reference spectrum. Despite being called Modified Euclidean Distance (because it modifies standard Euclidean distance), it itself is not a distance metric as it is asymmetric. Bradshaw[15] used the Tversky model of similarity in Daylight Chemical Information Systems on structural descriptor data set. Bradshaw[16] discussed that in chemical information systems it is fundamental to look for a super structure. Mestres and Maggiora[17] discussed and presented a generalized asymmetric form of field-based molecular similarity indices. An empirical study by Chen and Brown[18] presented the evidence of asymmetry in chemical similarity measures and used Tversky similarity coefficients for experimenting and suggested that using asymmetric coefficients is beneficial for database searching. However, these papers do not discuss whether the metrics are really needed and how they are violated in practical applications.

To evaluate our hypothesis, we have adopted Tversky's contrast model described in the next section. This model is appropriate to represent the proposed Contains concept and to test our hypothesis. It uses a feature based binary representation and computes similarity as a linear combination of common and distinctive features by allowing different weights on each of them. It is also appealing because it is a direct generalization of the Jaccard metric that is widely used in spectroscopic library search.

3.1. Contrast Model. In the experimental psychology literature, there is a large amount of research concerned with similarity.[19] It is one of the most central theoretical concepts in psychology.[4,20] It is strongly linked to knowledge and behavior. It serves as an organizing principal in human perception.[4,7,9] Shepard inferred that similarity is a measure of the degree to which one stimuli generalizes to other.[6] It "refers to the outcome of a comparison among entities, usually a comparison based on many of the entities properties. Objects are similar to the degree that they have features in common and do not have distinctive features" (ref 21, p 4). And, "for similarity to be a useful construct, one must be able to specify the ways or respects in which two things are similar" (refs 20, p 254). The concept of similarity is strongly attached to knowledge. In psychology, the work and generalization on similarity is hugely influenced by Shepard's work on similarity and categorization.[6,8,22]

Tversky,[4] a psychologist, provided empirical evidence from different domains on the weakness of the metric axioms and suggested an alternative approach to geometric methods of similarity known as the contrast model. Unlike geometric methods of similarity, which represent objects as points in some coordinate space, Tversky's contrast model represents objects by a set of features and similarity is defined as a linear combination, or a contrast, of the counts of their common and distinctive features with different weights on each of them. The object representation as a set of features is viewed as the result of a prior process of extraction and compilation. Tversky defines that features may correspond to components of the object or they may be either concrete or abstract properties. The contrast model of similarity measures two objects to be more similar if they have more common features and fewer distinctive features. They are less similar if they have more distinctive features and less common features. The term feature usually denotes the value of a binary variable. However, it is also applicable to ordinal or cardinal variables.

Let a and b be two objects represented by feature set A and B, respectively. Then, the similarity of a to b, denoted by $S(a, b)$, is a linear combination of common and distinctive features,[4,7] i.e.

$$S(a, b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A) \tag{3}$$

where

- $f(A \cap B)$ represents set of common features of A and B
- $f(A - \mathrm{B})$ represents distinctive features of A that B does not have
- $f(B - \mathrm{A})$ represents distinctive features of B that A does not have
- f satisfies feature additivity, i.e. $f(A \cup B) = f(A) + f(B)$, whenever A and B are disjoint
- f and S are interval scales

The parameters θ, α, and β reflect weights given to common and distinctive features and $\theta, \alpha, \beta \geq 0$.

The contrast model does not define a single similarity scale but rather a family of scales characterized by different values of parameter θ, α, and β. It expresses the similarity between objects as a weighted difference of the count of their common and distinctive features. Another form of the contrast model is the ratio model, given by

$$S(a, b) = \frac{f(A \cap B)}{(f(A \cap B) + \alpha f(A - B) + \beta f(B - A))},$$
$$\alpha, \beta \geq 0 \tag{4}$$

where similarity is normalized so that S lies between 0 to 1. It is a generalized form of the contrast model.[4]

The contrast model is appropriate to represent our Contains concept as it is a weighted combination of common and distinctive feature counts. By weighing reference spectra features more heavily than a query spectra features, i.e., $\alpha > \beta$, we focus more on referent features contained in query spectra rather than on the additional features present in query spectra. As a result, the presence of those additional query features will have less of an effect on the measuring coefficient.

3.2. Applications for Contains. Our proposed Contains concept is of interest in many practical applications. In section 5 we evaluate the necessity and appropriateness of Contains as opposed to similarity metrics on data sets from Raman Spectroscopy and Mass Spectrometry experimentally. Chen and Brown[18] evaluated the performance of the Tversky measure (which we identify as being suitable to represent Contains) in computer-assisted drug discovery using two large pharmaceutical databases, the NCI anti-AIDS database and the J&J corporate database. They concluded that the relative weights of the training and the query object can be adjusted to more effectively achieve different purposes of similarity. We expect to see more applications of our proposed concept in scenarios where there is a unique fingerprint or representation of training objects and the user aims to identify which of them are contained in the test object. For example, it could prove to be useful in the task of identification of tandem mass spectra in data sets generated by SWATH.[23] In a general experimental setup, each tandem mass spectrum used for database searching contains fragment ions from a single peptide.[24] However, SWATH produces highly complex and composite fragment-ion spectra where our proposed Contains concept might be of great interest and benefit as the user is interested in finding all the tandem mass spectra contained in complex and composite spectra.

Section 4 details the experiments we have conducted to evaluate the Tversky similarity measure on spectroscopic data and its results. Previous papers[15,18] have implemented Tversky's contrast model for assessing chemical similarity using a structural descriptor data set, but without explicitly considering Contains, of course. In our experiments, we reduce the spectrum into a binary feature based representation using peaks presence/absence over a specified range as its features, i.e. 0 if a peak (feature) is absent or 1 if peak is present over a specified wavenumber range.

4. MATERIALS AND METHODS

This section presents details about the data sets, the preprocessing steps taken, the procedure we adopted to perform spectral library search, and the methodology adopted to evaluate the performance of the similarity measure.

4.1. Description of the Data Set. *4.1.1. Extended Chlorinated.* The Extended Chlorinated data set was one that was first described by Conroy.[25] The data set contains 433 spectra, each having 2473 channels/data-points. The data set contains both pure spectra of solvents, spectra of various mixtures, and spectra doped with varying degrees of fluorescence. The data set is made using 29 solvents specified in Table 1. The reference library consists of the pure solvents. For a few solvents, there are multiple spectra. The total number of spectra in the reference library is 34. The remaining 399 spectra are used as a test set. On visual inspection of them, we observed 22 spectra with large fluorescence responses that dominated the entire range because of the contamination of mixtures with high concentration of rhodamine B. Since we did not include any preprocessing to attempt to correct such effects, these were removed as outliers from the data set prior to any experiments. This results in a testing data set containing 377 spectra of various mixtures.

Table 1. List of Solvents Used along with Different Grades[25]

chlorinated	nonchlorinated
111-trichloroethane	ethanol, cyclopentane, acetophenol
chloroform	*n*-pentane, xylene, nitromethane
dichloromethane	dimethylformamide, tetrahydrofuran
chlorobenzene	diethyl ether, petroleum ether
carbontetrachloride	cyclohexane, acetone, toluene
	acetonitrile, 2-propanol, 14-dioxane
	hexane, methyl alcohol, ethyl acetate
	1-butanol, cyclohexane, 1-propanol
	iso-pentane, benzene

4.1.2. Micromass. The Micromass data set is taken from UCI[26] and first appeared in the work of Mahé et al.[27] The aim is to identify microorganisms from the mass spectrometry data. Each spectrum is represented by a vector $x \in \mathcal{R}^p$; p is the number of channels or bins involved in the peak-list representation. Each entry in the vector x represents the intensity of the peaks found in that bin. We note this can be framed as a Contains query where the task is to predict which of the K reference bacterial species are contained in the query spectrum. The main objective is similar to the spectral library search, i.e. to identify the components of the bacterial mixture. A prespecified reference and the test set is provided. The reference data set consist of 571 pure mass spectra of 20 Gram positive and negative bacterial species covering nine genera. The test data set contains 360 mass spectra. The test spectra were represented by two pairs of strains, which are mixed according to nine different concentration ratios. We have use these prespecified sets for our experiments.

4.2. Data Preprocessing. The spectra in the data set are normalized using min−max normalization. Given a spectrum $S = [s_1, s_2, ..., s_N]$, the normalized spectrum S' defined as

$$s_i' = \frac{s_i - \min}{\max - \min} \tag{5}$$

where min and max are the minimum and maximum values of the original spectrum. Each normalized spectrum is scaled between 0 and 1.

4.3. Data Transformation. Encoding as a pattern of bits is one of the common strategies to increase the efficiency of database searching.[28] Grotch was the first to report the use of peak/no peak (or binary) encoding of mass spectra for library searches.[29] If a particular peak is present at a particular position over a specified range, then its corresponding bit is set to 1 in the binary feature vector, otherwise it is set to 0. Very minor peaks are excluded by using an intensity threshold, i.e. the presence of peak above a given intensity threshold is coded as 1; otherwise, it is coded as 0. In this representation, it is only the presence/absence of particular peaks which sets/unsets the bit and the peak height is not considered. We have used this approach to transform spectra into binary vectors. As the Micromass data set considers the peak-list representation for each mass spectrum, the spectrum has been normalized and then transformed into binary-representation. However, the Extended Chlorinated data set consists of raw spectrum in which we need to find the peaks. Therefore, we developed software in Java to implement the procedure. For detecting the peaks, the MATLAB function *peakfinder*[30] is used, and to interact with MATLAB, a Java API called matlabcontrol[31] is used. This peak detection function uses the alternating nature of derivatives along with amplitude threshold to find the local maxima or minima. For our experiments, we save the peak list for each of the pure spectra present in our reference library. In theory, measures for our Contains will not differ in response to the selection of different intensity thresholds in cases when the same threshold settings are applied. But it needs to kept in mind that in practice the peak intensity may reduce in a query spectrum due to the presence of its other constituents.

4.4. Similarity Measures. This section details the similarity measures used, i.e., the Jaccard coefficient and the Tversky coefficient, to conduct our experiments. We also show that the commonly used similarity measures obey the symmetry axiom.

Given two objects A and B, we define b to be the number of features present in Object A but not in Object B, c to be the number of features present in Object B but not in Object A, a to be the number of features present in both objects, and d to be the number of features absent in both objects. In addition, the total number of features measured are $n = a + b + c + d$. The total number of features present in Object A is $(a + b)$, and the total number of features present in Object B is $(a + c)$. The binary association coefficients or similarity measures are based on a, b, c, and d. These are described by the presence or absence of features illustrated in Table 2.

Table 2. Four Binary Cases for Comparing Object *A* and Object *B*[a]

Object *A*	Object *B*	
1	1	*a*
1	0	*b*
0	1	*c*
0	0	*d*

[a] 1 denotes a feature present, 0 denotes a feature absence.[2]

Gower[32] noted that when the absence of a feature in both objects is deemed to convey no information, then d should not occur in the similarity measure. The Jaccard similarity coefficient, also known as the Tanimoto similarity measure,[33] is the most widely used association coefficient.[34] It defines similarity as

$$\mathrm{Ta}(A, B) = \frac{a}{a + b + c} \tag{6}$$

The Tversky contrast ratio model defines similarity coefficient as

$$\mathrm{Tv}(A, B) = \frac{a}{\alpha b + \beta c + a} \tag{7}$$

The Tversky contrast ratio model can be regarded as a generalization of the Jaccard similarity measure: if $\alpha = \beta = 1$, we get the Jaccard coefficient, which is a metric. In our

DOI: 10.1021/acs.jcim.5b00077

experiments, we have adjusted the parameter α from 0 to 1. The parameter β is calculated as $1 - \alpha$.

For performance comparisons, the Jaccard similarity index measure is chosen on the basis of following studies. Nikolova and Jaworska,[1] Monev,[35] and Willett et al.[36] reviewed binary association coefficients in chemical information system. Willett[37] evaluated 13 similarity measures for binary fingerprint code. Chen and Reynolds[38] evaluated the effectiveness of Euclidean distance and the Jaccard coefficient on 2D fragment based descriptors. The results from these studies indicated that the Jaccard coefficient gives the best performance. The Jaccard similarity index metric is a distance metric that satisfies all three metric axioms. Therefore, it is used for comparing the performance of the Tversky measure.

In the next section, we check the symmetry axiom on the commonly used binary similarity measures.

4.4.1. Symmetry Axiom.

For completeness Table 3 shows an analysis of the commonly used binary similarity indexes (mentioned by Brereton;[2] p 78). It lists the directional similarity coefficients i.e. $S(A, B)$, the similarity from A to B and $S(B, A)$, the similarity from B to A. Consider the two objects A and B with a (feature count present in both objects), b (feature count present in Object A but absent in B), c (feature count present in Object B but absent in A), and d (feature count absent in both objects). When computing $S(A, B)$ and $S(B, A)$, the number of common features present/absent, i.e., a and d remains the same; while the values of distinct feature counts i.e., b and c, are swapped. The table below shows the resulting formula for each. It can be observed that for all the commonly used measures listed below $S(A, B)$ is equal to $S(B, A)$. Therefore, these measures are symmetric; however, as we have discussed symmetry is not needed in spectral similarity search.

Table 3. Symmetry Axiom Check on Binary Similarity Measures[a]

	similarity		
similarity index S	$S(A, B)$	$S(B, A)$	obeys symmetry axiom? (Y/N)
Jaccard similarity index	$a/(a + b + c)$	$a/(a + c + b)$	Y
Simple matching	$(a + d)/(a + b + c + d)$	$(a + d)/(a + c + b + d)$	Y
Sokal and Sneath (a)	$a/[a + 2(b + c)]$	$a/[a + 2(c + b)]$	Y
Rogers and Tanimoto	$(a + d)/[a + 2(b + c) + d]$	$(a + d)/[a + 2(c + b) + d]$	Y
Sorensen	$2a/(2a + b + c)$	$2a/(2a + c + b)$	Y
Gower and Legendre	$[a - (b + c) + d]/(a + b + c + d)$	$[a - (c + b) + d]/(a + c + b + d)$	Y
Ochiai	$a/[(a + b)(a + c)]^{1/2}$	$a/[(a + c)(a + b)]^{1/2}$	Y
Snockal and Sneath (b)	$ad/[(a + b)(a + c)(d + b)(d + c)]^{1/2}$	$ad/[(a + c)(a + b)(d + c)(d + b)]^{1/2}$	Y
Pearson's φ	$(ad - bc)/[(a + b)(a + c)(d + b)(d + c)]^{1/2}$	$(ad - cb)/[(a + c)(a + b)(d + c)(d + b)]^{1/2}$	Y
Russell and Rao	$a/(a + b + c + d)$	$a/(a + c + b + d)$	Y

[a] From ref 2, p 78.

4.5. Methodology for Measuring Search Performance.

The performance of spectral library search is evaluated using retrieval accuracy. In our experiments, we compute the similarity coefficient (using both eqs 6 and 7) between a query spectra and each of the reference library pure spectra. Pure spectra with a similarity coefficient greater than the specified threshold are included in the hit list for that given query. This hit list is then compared to the true list of constituents of the query spectra to find the number of relevant retrieved items. Then, the value of retrieval accuracy for the ith query spectra is computed using

$$\text{retrieval accuracy}_i = \frac{\text{no. (relevant retrieved items)}_i}{\text{no. (total retrieved items)}_i} \tag{8}$$

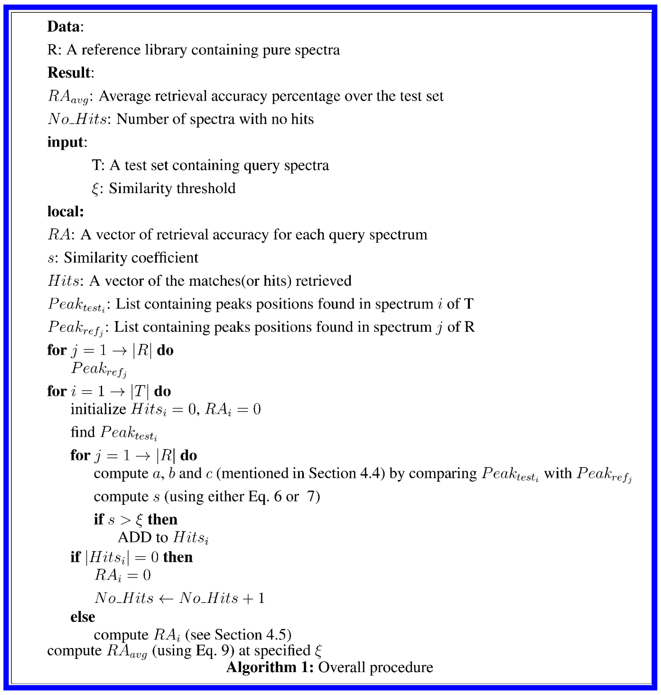
This value will be within the range [0, 1]. In cases where the hit list is empty, a retrieval accuracy of zero is reported. Next, an average value of retrieval accuracy across the full test set is computed. The average retrieval accuracy over the test set is calculated using

$$\text{average retrieval accuracy} = \frac{\sum_{i=1}^{n=\text{number of test spectra}} \text{retrieval accuracy}_i}{\text{number of test spectra}} \tag{9}$$

We report the average percentage retrieval accuracy in the results. In addition to this, we also report the number of query spectra for which we do not retrieve any hits.

4.6. Overall Procedure.

The procedure followed to evaluate the performance is described in Algorithm 1.

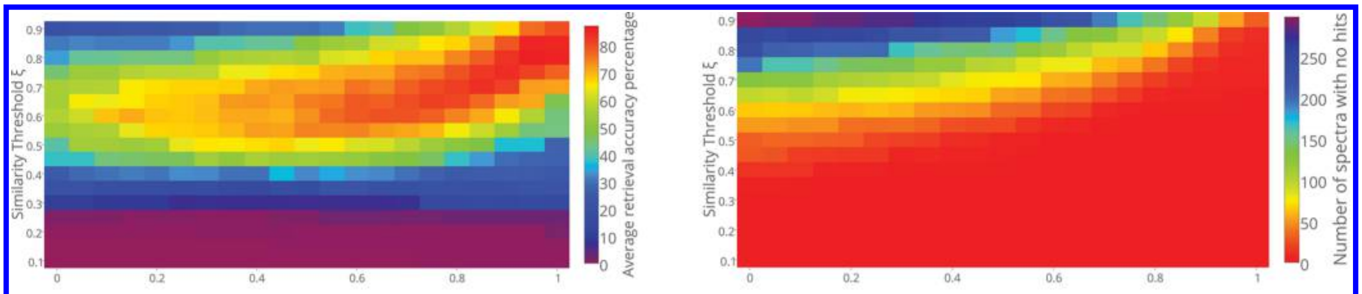


```
Data:
R: A reference library containing pure spectra
Result:
$RA_{avg}$: Average retrieval accuracy percentage over the test set
$No\_Hits$: Number of spectra with no hits
input:
    T: A test set containing query spectra
    $\xi$: Similarity threshold
local:
$RA$: A vector of retrieval accuracy for each query spectrum
$s$: Similarity coefficient
$Hits$: A vector of the matches(or hits) retrieved
$Peak_{test_i}$: List containing peaks positions found in spectrum $i$ of T
$Peak_{ref_j}$: List containing peaks positions found in spectrum $j$ of R
for $j = 1 \rightarrow |R|$ do
    $Peak_{ref_j}$
for $i = 1 \rightarrow |T|$ do
    initialize $Hits_i = 0$, $RA_i = 0$
    find $Peak_{test_i}$
    for $j = 1 \rightarrow |R|$ do
        compute $a$, $b$ and $c$ (mentioned in Section 4.4) by comparing $Peak_{test_i}$ with $Peak_{ref_j}$
        compute $s$ (using either Eq. 6 or  7)
        if $s > \xi$ then
            ADD to $Hits_i$
    if $|Hits_i| = 0$ then
        $RA_i = 0$
        $No\_Hits \leftarrow No\_Hits + 1$
    else
        compute $RA_i$ (see Section 4.5)
compute $RA_{avg}$ (using Eq. 9) at specified $\xi$
```

Algorithm 1: Overall procedure

5. RESULTS

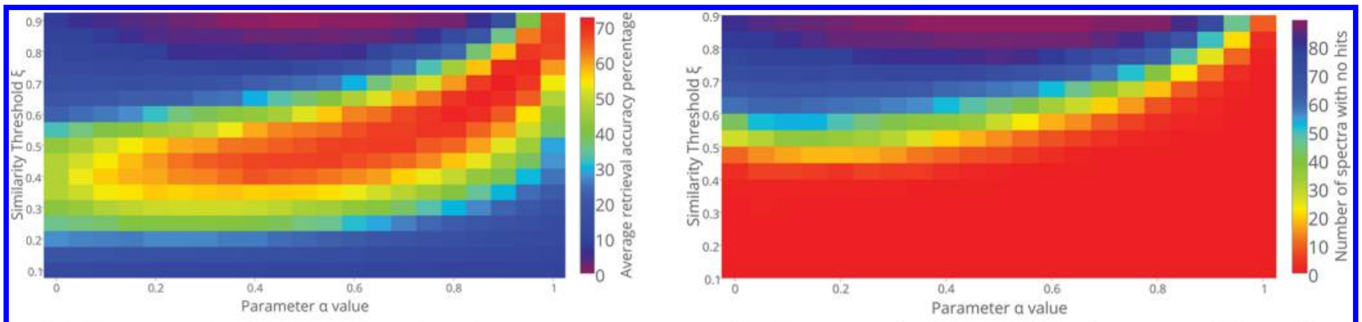
Figures 1 and 2 report the results for Extended Chlorinated and Micromass data sets, respectively. For the Tversky coefficient, the experiments are conducted with varying values for α and β within the range of [0, 1]. The performances of the different similarity measures are not compared at the same threshold value; because the formulas are different, the effects of thresholds will be different. Therefore, we choose the threshold with the best retrieval accuracy for each approach and compare these best-case results. It should be noted that when α is set to



(a) Heat map for the average retrieval accuracy over the test set. The red color indicates a better average retrieval accuracy.

(b) Heat map for the number of spectra with no hits over the test set. The red color indicates the lower number of spectra with no hits.

Figure 1. Extended Chlorinated data set results at specified threshold levels with varying parameters values.



(a) Heat map for the average retrieval accuracy over the test set. The red color indicates a better average retrieval accuracy.

(b) Heat map for the number of spectra with no hits over the test set. The red color indicates the lower number of spectra with no hits.

Figure 2. Micromass data set results at specified threshold levels with varying parameters values.

0.5, then β becomes 0.5, the resulting coefficient becomes symmetric. The threshold values tested ranged from 0.1 to 0.9 with an increment of 0.05. Table 4 provides a summary of the best performance for each of the data sets for both symmetric and asymmetric measures.

Table 4. Summary of the Best Performance Results for the Extended Chlorinated and Micromass Data Sets

data set	measure	RA_{avg}%	*No_Hits*	threshold ξ	parameter α
Extended Chlorinated	Jaccard	76.65	47	0.45	NA[a]
	Tversky	87.00	11	0.80	0.95
Micromass	Jaccard	68.51	2	0.30	NA[a]
	Tversky	72.40	4	0.70	0.90

[a] NA is not applicable.

Figure 1 reports the average retrieval accuracy percentage over the test set and the number of query spectra with no hits (or zero retrieval accuracy) at the specified similarity threshold levels for the Extended Chlorinated data set with varying values for α. The color scale shown in Figure 1b is inverted to the color scale shown in Figure 1a so that the same color corresponds to the best performance (i.e., red). The Jaccard measure results at each threshold are provided in the Supporting Information. The best average retrieval accuracy for the Jaccard coefficient is 76.65% at a threshold of 0.45. The average retrieval accuracy shows that the symmetric similarity coefficient performs badly when the threshold is high and there are lot of query mixtures for which nothing is detected. The reason for this is that a mixture will have additional features in comparison to the pure substance, and the Jaccard coefficient reduces due to these distinguishing features. However, if the threshold is lowered then the performance improves. This illustrates that the Jaccard similarity index is affected by mixtures not being drawn from the same distribution as the training/reference library (which contains only pure solvents).

The best average retrieval accuracy for the Tversky measure is 87.00% with the parameter setting ($\alpha = 0.95, \beta = 0.05$) at the threshold of 0.8. The parameters α and β do influence the retrieval performance. The improvement in the performance is because the Tversky measure considers the asymmetry. The peaks present in the reference spectra are weighed more than the peaks present in query spectra. This is evident from Figure 1a where we clearly see an asymmetric pattern where the better performance is where $\alpha > \beta$. We also observe that we obtain better performance with high similarity threshold levels. This is because of the lower weights on the distinctive features present in query spectrum which consequently make less contribution comparatively to the Jaccard measure. On decreasing the threshold the retrieval accuracy is reduced. This is because, by reducing the similarity threshold, the possibility of false positives increases. On the other hand, the retrieval accuracy for the Jaccard index improves until the threshold is reduced to 0.45. The reason for this is that the similarity value reduces due to distinguishing features (or peaks). As soon as the threshold is lowered, the Jaccard index is able to detect hits. However, even with the threshold = 0.45, there are 47 query spectra with no hits (or matches). The number of test spectra with no hits for the Tversky measure is 11, which is comparatively low when compared to the best performance of the Jaccard index. When these spectra are further examined, it is found that for eight spectra the peak detection algorithm did not detect some peaks in the query spectrum that are present in the pure substance spectrum, at the specified peak detection setting. In the remaining spectra, we do detect the peaks in the query spectrum for some of its constituents; however, there are comparatively more distinctive peaks present in the query spectrum thus lowering the coefficient value (β is set to 0.05). To be consistent with our results, we chose to have the same peak detection parameter setting over the entire test set. Furthermore, from Figure 1b we observe that when the threshold levels are reduced the number of spectra with no hits reduces.

Figure 2 reports the average retrieval accuracy percentage over the test set and the number of query spectra with no hits (or zero retrieval accuracy) at the specified similarity threshold levels for the Micromass data set. The color scale shown in Figure 2b is inverted to the color scale shown in Figure 2a so that the same color (i.e., red) corresponds to the best performance, as we did for the Extended Chlorinated results. The Jaccard measure result is summarized in the Supporting Information. The best average retrieval accuracy for the Jaccard coefficient is 68.51% at a threshold of 0.3. There are 2 spectra for which nothing was detected. We observe the same behavior as with the Extended Chlorinated data set that with the increasing threshold the performance reduces. Figure 2a clearly indicates that there is an asymmetric pattern with better performance inclined toward the right i.e. when α is greater than β. From Figure 2b it is observed that when α is greater than β, the number of spectra for which no hits are retrieved are reduced along with the improvement in the accuracy. Therefore, it increases the possibility of finding true matches. The best average retrieval accuracy for the Tversky measure is 72.40% with the parameter setting (α = 0.90, β = 0.10) at the threshold of 0.7.

From the results, it is concluded that the Tversky measure used to represent our Contains concept yields better results when compared to the Jaccard similarity index metric. It is also concluded that it is worthwhile to examine and incorporate nonmetric measures in library search rather than being limited to metric measures: nonmetric measures can yield improvements in performance, and even where they do not help, they do not degrade the performance. The nonmetric measures also provide the advantage of an increased ability to comprehend similarity that is consistent with intuitive human ideas of similarity. It is worthwhile to note that by using the Tversky measure we increase the possibility of getting hits and also improve the retrieval accuracy, i.e. hits with true positives or hits with relevant spectra.

6. CONCLUSIONS AND FUTURE WORK

Similarity is the fundamental concept underlying spectral library search, so it is important to identify good measures for assessing similarity. Most of the similarity measures used in academic work and existing software follow the metric axioms, i.e., minimality, symmetry, and triangle inequality. Even those measures that are not metrics usually assume symmetry. The focus of this paper has been to suggest that metric measures are not appropriate in many practical spectroscopic applications. This is because the metric axioms that these measures obey are not necessary and in fact are often violated in practice. Also, the implicit assumption observed by these measures is that the queries are drawn from the same distributional space as of the training library, whereas in real spectroscopic applications, the query items may not be drawn from the same distribution as that of a reference library even if it is carefully constructed. In such scenarios, we propose to use our Contains concept which is inherently asymmetric as an alternative to similarity metrics. This concept asks which of the reference library spectra are contained in the given query spectrum rather than finding reference library spectra that are most similar to the given query spectrum. This paper provides a good starting point to understand the importance of defining and measuring the concept of similarity. The measure should be able to reflect the similarity appropriate to the domain. Contains is useful in experiments where there is a unique representation or fingerprint of spectrum in the training library and the user aims to identify the presence of these features in the query spectrum.

Our hypothesis is evaluated by adopting the Tversky contrast model which is appropriate to represent our Contains concept and comparing its performance to the Jaccard similarity index metric for spectral library search. Retrieval accuracy is used to measure the performance. The performance is evaluated on a Raman Spectroscopy and a Mass Spectrometry data set. From the results, it is seen that the Tversky measure yields better results with a retrieval accuracy of 87.00% at a threshold = 0.80 with parameters α = 0.95 and β = 0.05 for Extended Chlorinated data set, and a retrieval accuracy of 72.40% at a threshold = 0.70 with parameters α = 0.90 and β = 0.10 for Micromass data set. The Jaccard index did not perform as well because of the reduction in the similarity value due to the presence of distinguishing features. This study supports the hypothesis that spectral library search performance can be improved by using nonmetric measures. Although we tested our Contains concept on data transformed to a binary feature based representation, it is to be noted that it may not be limited to this representation. For future work, we will investigate the applicability of our Contains concept for continuous data. We also aim to present an approach to find the optimal values of α and β for a given data set that works well on unseen data.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Results in tabular format for the experiments performed in this paper. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00077.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: a.garg2@nuigalway.ie (A.G.).
*E-mail: catherine.enright@nuigalway.ie (C.G.E.).
*E-mail: michael.madden@nuigalway.ie (M.G.M.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by the European Commission under the Framework 7 Programme as contract number 284456.

■ REFERENCES

(1) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity-A Review. *QSAR Comb. Sci.* **2003**, *22*, 1006−1026.

(2) Brereton, R. G. *Chemometrics for Pattern Recognition*; John Wiley & Sons: West Sussex, United Kingdom, 2009.

(3) Khan, S. S.; Madden, M. G. New Similarity Metrics for Raman Spectroscopy. *Chemom. Intell. Lab. Syst.* **2012**, *114*, 99−108.

(4) Tversky, A. Features of Similarity. *Psychol. Rev.* **1977**, *84*, 327−352.

(5) Shepard, R. N. Multidimensional Scaling, Tree-fitting, and Clustering. *Science* **1980**, *210*, 390−398.

(6) Shepard, R. N. Toward a Universal Law of Generalization for Psychological Science. *Science* **1987**, *237*, 1317−1323.

(7) Tversky, A.; Gati, I. Studies of Similarity. *Cogn. Categ.* **1978**, *1*, 79−98.

(8) Shepard, R. N. The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function.I. *Psychometrika* **1962**, *27*, 125−140.

(9) Goldstone, R. L. *The MIT Encyclopedia of the Cognitive Sciences*; Wilson, R. A., Keil, F. C., Eds.; MIT Press: Cambridge, MA, 1999.

(10) Goldstone, R. L. The Role of Similarity in Categorization: Providing a Groundwork. *Cognition* **1994**, *52*, 125−157.

(11) Nosofsky, R. M. Exemplar-based Accounts of Relations between Classification, Recognition, and Typicality. *J. Exp. Psychol.-Learn. Mem. Cogn.* **1988**, *14*, 700.

(12) Gower, J. C.; Legendre, P. Metric and Euclidean Properties of Dissimilarity Coefficients. *J. Classification* **1986**, *3*, 5−48.

(13) Attneave, F. Dimensions of Similarity. *Am. J. Psychol.* **1950**, *63*, 516−556.

(14) Tversky, A.; Gati, I. Similarity, Separability, and the Triangle Inequality. *Psychol. Rev.* **1982**, *89*, 123−154.

(15) Bradshaw, J. Introduction to Tversky Similarity Measure. Presented at *11th Annual Daylight MUG Meeting*, Laguna Beach, CA, Feb 25−28, 1997; http://www.daylight.com/meetings/mug97/Bradshaw/MUG97/tv_tversky.html.

(16) Bradshaw, J. YAMS—Yet Another Measure of Similarity. Presented at *Euromug01*, Cambridge, UK, Sept 13−14, 2001; http://www.daylight.com/meetings/emug01/Bradshaw/Similarity/YAMS.html.

(17) Mestres, J.; Maggiora, G. M. Putting Molecular Similarity into Context: Asymmetric Indices for Field-based Similarity Measures. *J. Math. Chem.* **2006**, *39*, 107−118.

(18) Chen, X.; Brown, F. K. Asymmetry of Chemical Similarity. *ChemMedChem.* **2007**, *2*, 180−182.

(19) Hahn, U.; Chater, N. Understanding Similarity: A Joint Project for Psychology, Case-based Reasoning, and Law. *Artif. Intell. Rev.* **1998**, *12*, 393−427.

(20) Medin, D. L.; Goldstone, R. L.; Gentner, D. Respects for Similarity. *Psychol. Rev.* **1993**, *100*, 254−278.

(21) Sloman, S. A.; Rips, L. J. Similarity as an Explanatory Construct. *Cognition* **1998**, *65*, 87−101.

(22) Shepard, R. N. Stimulus and Response Generalization: A Stochastic Model Relating Generalization to Distance in Psychological Space. *Psychometrika* **1957**, *22*, 325−345.

(23) Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. Targeted Data Extraction of the MS/MS Spectra Generated by Data-Independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics* **2012**, *11*, O111−016717.

(24) Matthiesen, R. *Mass Spectrometry Data Analysis in Proteomics*; Humana Press: Totowa, NJ, 2007; Vol. *1*.

(25) Conroy, J. Hazardous Substance Analysis Using Raman Spectroscopy and Chemometrics. Master's thesis, National University of Ireland, October, 2005.

(26) Lichman, M. UCI Machine Learning Repository. 2013; http://archive.ics.uci.edu/ml (accessed July 15, 2014).

(27) Mahé, P.; Arsac, M.; Chatellier, S.; Monnin, V.; Perrot, N.; Mailler, S.; Girard, V.; Ramjeet, M.; Surre, J.; Lacroix, B.; Belkum, A. V.; Veyrieras, J.-B. Automatic Identification of Mixed Bacterial Species Fingerprints in a MALDI-TOF Mass-Spectrum. *Bioinformatics* **2014**, *90*, 1280−1286.

(28) Flower, D. R. On the Properties of Bit String-based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.

(29) Grotch, S. L. Matching of Mass Spectra when Peak Height is Encoded to One Bit. *Anal. Chem.* **1970**, *42*, 1214−1222.

(30) Yoder, N. PeakFinder. http://www.mathworks.com/matlabcentral/fileexchange/file_infos/25500-peakfinder (accessed July 15, 2013).

(31) matlabcontrol A Java API to interact with MATLAB. https://code.google.com/p/matlabcontrol/ (accessed July 15, 2013)

(32) Gower, J. C. In *Encyclopedia of Statistical Sciences*; Kotz, S., Read, B. C., Balakrishnan, N., Vidakovic, B., Eds.; John Wiley & Sons: New York, USA, 1985; Vol. *5*; pp 397−405.

(33) Vandeginste, B. G. M., Massart, D. L., Buydens, L. M. C., de Jong, S., Lewi, P. J., Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part B*; Elsevier: Amsterdam, The Netherlands, 1998.

(34) Willett, P. *Similarity and Clustering in Chemical Information Systems*; John Wiley & Sons, Inc.: New York, USA, 1987.

(35) Monev, V. Introduction to Similarity Searching in Chemistry. *MATCH-Commun. Math. Co.* **2004**, *51*, 7−38.

(36) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(37) Willett, P. Similarity-based Approaches to Virtual Screening. *Biochem. Soc. Trans.* **2003**, *31*, 603−606.

(38) Chen, X.; Reynolds, C. H. Performance of Similarity Measures in 2D Fragment-based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1407−1414.