Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2011 ; 2011: 604–607. doi:10.1109/BIBM.
2011.113.

Hypergeometric Similarity Measure for Spatial Analysis in Tissue Imaging Mass Spectrometry

Chanchala Kaddi¹, R. Mitchell Parry¹, and May D. Wang¹

¹Wallace H. Coulter department of Biomedical Engineering, Georgia Institute of Technology,
Atlanta, GA, 30332

Abstract

Tissue imaging mass spectrometry (TIMS) is a data-intensive technique for spatial biochemical analysis. TIMS contributes both molecular and spatial information to tissue analysis. We propose and evaluate a similarity measure, based on the hypergeometric distribution, for comparing m/z images from TIMS datasets, with the goal of identifying m/z values with similar spatial distributions. We compare the formulation and properties of the proposed method with those of other similarity measures, and examine the performance of each measure on synthetic and biological data. This study demonstrates that the proposed hypergeometric similarity measure is effective in identifying similar m/z images, and may be a useful addition to current methods in TIMS data analysis.

Keywords

mass spectrometry; hypergeometric; imaging

I. Introduction

Tissue imaging mass spectrometry (TIMS) is a data-intensive technique for biochemical analysis in which a complete mass spectrum is acquired at multiple points on a tissue sample. TIMS directly links the biochemical information of the mass spectrum to the morphological characteristics of the tissue. Coupled molecular and spatial data is valuable in studying diseases where an individual tissue sample may contain diseased, marginal and disease-free regions, *e.g.* cancer. TIMS produces large, information-rich datasets, and the development of computational techniques for exploring these datasets is an important factor in effectively employing TIMS in biomedical research.

In mass spectrometry data analysis, researchers seek to observe relationships among m/z (mass-to-charge ratio) values. Identifying similarly expressed m/z values may aid in biologically characterizing disease states. In conventional (non-TIMS) mass spectrometry, patterns may be observed in terms of abundance or mass. TIMS enables assessment in terms of spatial distribution, extending upon histological staining. It can be applied to identify m/z

Correspondence May D Wang, Ph.D., Biomedical Engineering Department, Georgia Institute of Technology & Emory University, 313 Ferst Drive, UA Whittaker Building Suite 4106, Atlanta, GA, 30332, Tel: 404-3852954, Fax: 404-894-4243, maywang@bme.gatech.edu.

values present in the same tissue regions as molecules known to be of interest in a disease, thereby generating a shortlist of m/z values for further study. Recent papers have discussed techniques for identifying similarly distributed m/z values [1–2].

In this paper we propose a similarity measure for TIMS data, using the hypergeometric distribution as a basis, to identify spatially similar m/z values. The hypergeometric distribution has previously been used in bioinformatics to assess similarity in microarray functional analysis and tandem mass spectrometry [3–5]. In this paper we define the proposed hypergeometric similarity measure and compare it with cosine similarity and Pearson correlation in terms of desirable properties related to formulation and behavior. Cosine similarity and Pearson have previously been used to assess similarity in mass spectrometry data for tasks ranging from protein identification to quality control [1–2, 6–8]. We study the performance of the proposed similarity measure on synthetic data and provide examples showing its advantageous performance in identifying and ranking similarities. Then, we implement it on a biological dataset to assess its utility in identifying spatially similar m/z values. Results indicate that the proposed similarity measure is effective in meaningfully discriminating among m/z images, and can be a useful component of the analytical pipeline for TIMS data analysis.

II. Methods

A. Desirable properties of a similarity measure

The proposed similarity measure should sufficiently meet the following properties related to design and performance. The similarity measure should (1) be monotonically increasing between $[-1, 1]$, to facilitate interpretation and comparison with other measures; (2) have good power of discrimination, *i.e.* should identify differences where they exist; (3) be consistently defined, *i.e.* there should not be sets of valid (observable) inputs for which the similarity measure output is undefined, and valid inputs should utilize the full dynamic range of the output.

B. Binary representation of TIMS data

Spatial distribution, rather than abundance, is the informative factor in finding co-localized m/z values. We therefore utilize a binary representation of TIMS data: each m/z image indicates the presence, above a selected threshold, of the corresponding m/z value at each pixel. There are many different methods for selecting a threshold in practice in biomedical image processing. Here, we demonstrate performance over a range of thresholds by utilizing the percentile abundances of the mean spectrum of the TIMS dataset. The abundance at every 10th percentile between the 0th to the 100th is considered as a threshold.

C. Definition of similarity measure

For a dataset with N pixels in each image, the reference m/z value has an image which contains n_1 pixels with intensities greater than a selected threshold. When binarized, this image will have n_1 ‘on’ pixels and $N - n_1$ ‘off’ pixels. A second, query m/z image has n_2 ‘on’ pixels. The total number of pixels at which both images are ‘on’ is k . The significance of overlap can be defined in terms of the probability of observing k given N , n_1 , and n_2 . If k

of the n_1 pixels from the first m/z image overlap k of the n_2 pixels from the second m/z

image, those k pixels in the first image may be arranged in $\binom{n_1}{k}$ ways. In the second

image, the $(n_2 - k)$ pixels which do not overlap may be arranged in $\binom{N - n_1}{n_2 - k}$ ways in the image space N . Thus, the total number of ways in which an overlap of k pixels can occur,

given n_1, n_2 and N , is $\binom{n_1}{k} \binom{N - n_1}{n_2 - k}$. When divided by the number of ways in which the n_2 'on' pixels in the second image could be arranged if k of them were not constrained, this becomes the pdf of the hypergeometric distribution. We propose a similarity measure $h(k, n_1, n_2, N)$ which is defined, for any k , as the difference between the lower and upper tails of the hypergeometric distribution, as shown in equation (1).

$$h = \sum_{i=0}^k \frac{\binom{n_1}{i} \binom{N - n_1}{n_2 - i}}{\binom{N}{n_2}} - \sum_{j=k}^N \frac{\binom{n_1}{j} \binom{N - n_1}{n_2 - j}}{\binom{N}{n_2}} \quad (1)$$

The lower tail of the hypergeometric distribution has previously been utilized as a similarity measure [9]. Since the population is discrete, the upper tail offers additional information about k not available from only the lower tail, *i.e.* the probability of observing overlap at least as extreme. The upper and lower tails can be considered p -values of hypothesis tests. In both cases, the null hypothesis H_0 is that the observed overlap occurred by chance, given that both images are drawn from an urn model without replacement. The alternative hypotheses are that the observed overlaps are, respectively, larger or smaller than would be expected to occur at random for such an image pair. This implies that the images may be related, *i.e.* notably similar or dissimilar. Through the difference between the tails, the proposed measure provides a scaled description of the unexpectedness of any observed overlap. The tails of the hypergeometric distribution also have upper bounds [10–11]. For some parameter sets, the value of the hypergeometric pdf may be so small as to encounter machine resolution limits. Then, the proposed similarity measure may be implemented in terms of the upper bounds, as shown in equation (2).

$$\left(\left(\frac{p_1}{p_1 + t_1} \right)^{p_1 + t_1} \left(\frac{1 - p_1}{1 - p_1 - t_1} \right)^{1 - p_1 - t_1} \right)^n - \left(\left(\frac{p_2}{p_2 + t_2} \right)^{p_2 + t_2} \left(\frac{1 - p_2}{1 - p_2 - t_2} \right)^{1 - p_2 - t_2} \right)^n \quad (2)$$

Here, $n = n_1$, $p_1 = \frac{N - n_2}{N}$ and $t_1 = \left(\frac{n_1 - k}{n_1} - p_1 \right)$, such that $t_1 \geq 0$. Similarly, $p_2 = \frac{n_2}{N}$ and $t_2 = \left(\frac{k}{n_1} - p_2 \right)$, such that $t_2 \geq 0$.

D. Similarity measure comparison

First, the cosine similarity and Pearson correlation for binary vectors are expressed with the same variables as the hypergeometric pdf, allowing direct comparison of their formulae. Second, the proposed similarity measure is compared with the cosine similarity and Pearson correlation by evaluating their output for binary image pairs having varying degrees of overlap. Third, a synthetic dataset of binary vectors (images) with dimension $N = 10$ was created by considering all combinations of n_2 , n_1 and k such that $N \geq n_2 \geq n_1 \geq k$, and such that k is greater than or equal to its minimum for any given n_1 , n_2 and N . This dataset consists of 150 vector pairs. The three similarity measures were evaluated for each pair, and their outputs compared in terms of relative rankings. Fourth, the similarity measures were implemented on a TIMS dataset acquired from a mouse model of Tay-Sachs/Sandhoff disease. The experimental protocol for data acquisition is described in [12]. The image corresponding to m/z 890 was selected as a reference to due to its distinctive spatial pattern. The TIMS dataset has a spectral dimension of 4,438 m/z values. It was manually inspected in non-binary mode to identify m/z values with similar spatial patterns; 47 were selected. The top 47 values selected by each similarity measure were compared to these values. The percent correspondence of the two lists was calculated for each similarity measure, and repeated for each of the 11 binarization thresholds. The upper bounds of the tails of the hypergeometric distribution were used in this assessment due to the large parameter values involved.

III. Results

A. Similarity measure comparison

Similarities and differences among the measures may be observed through their formulae; the binary expressions for cosine similarity and Pearson correlation are shown in equations (3) and (4). These are derived by noting that for binary vectors V_1 and V_2 , the dot product is equivalent to k , and the norms to $\sqrt{n_1}$ and $\sqrt{n_2}$. Equation (4) is equivalent to Matthew's correlation coefficient [9, 13].

$$\frac{V_1 \cdot V_2}{\|V_1\| \|V_2\|} = \frac{k}{\sqrt{n_1} \sqrt{n_2}} \quad (3)$$

$$\frac{(V_1 - \bar{V}_1) \cdot (V_2 - \bar{V}_2)}{\|V_1 - \bar{V}_1\| \|V_2 - \bar{V}_2\|} = \frac{k - \frac{n_1 n_2}{N}}{\sqrt{n_1(1 - \frac{n_1}{N})} \sqrt{n_2(1 - \frac{n_2}{N})}} \quad (4)$$

Cosine similarity and Pearson correlation are both linear with respect to k , and behave nonlinearly with respect to n_1 and n_2 . Cosine similarity is independent of N , while mean-centering in Pearson correlation brings N into consideration. Pearson correlation asymptotically approaches cosine similarity for large N . The proposed measure, like Pearson correlation, considers N , but like cosine similarity, does not mean-center the data. Unlike both, it considers how unlikely it is to observe k by chance. Figure 1 demonstrates that the

proposed similarity measure satisfies criterion (1) regarding the desired properties of monotonicity and range. The hypergeometric similarity measure and Pearson correlation share a range of $[-1, 1]$, while for positive data the range of cosine similarity is $[0, 1]$. The extremes of the hypergeometric similarity measure represent the limits of observable overlap k for a given parameter set N , n_1 and n_2 .

B. Synthetic data

Figure 2 shows the performance of the similarity measures over the synthetic dataset. The rankings show that the proposed similarity measure fulfills criterion (2), which addresses discrimination of differing cases. In particular, we examine the extreme cases of (a) no overlap, (b) complete overlap and (c) ‘unsurprising’ overlap. (a) Cosine similarity assigns 0 to all vector pairs with no overlap; a large segment of the dataset is so labeled with no additional sorting. Pearson correlation and the proposed similarity measure both sort this subset of vectors. However, only the proposed similarity measure recognizes that $k = 0$ is more surprising when n_1 and n_2 are larger, because there is more opportunity for at least some overlap. (b) The treatment of cases with complete overlap ($k = n_1 = n_2$) is also more favorable with the proposed similarity measure because it orders them in a meaningful manner. It identifies $k = 5$ as the most ‘surprising’ case of complete overlap, since there are the most opportunities for non-overlap to occur. The probability that $k = n_1 = n_2 = 5$ is equal to $\left(\frac{5}{10} \frac{4}{9} \frac{3}{8} \frac{2}{7} \frac{1}{6}\right)$. It also recognizes that $k = 6$ and $k = 4$ are equally ‘surprising’, since the probability of arranging $n_2 = 6$ pixels to completely overlap $n_1 = 6$ pixels is the same as arranging $(N - n_2) = 4$ pixels to completely overlap $(N - n_1) = 4$ pixels. The same pairings are observed for $k = 7$ and $k = 3$, and $k = 8$ and $k = 2$. In contrast, both cosine similarity and Pearson correlation assign 1 to this set of vectors without further sorting. (c) The proposed similarity measure also meets criterion (3) regarding definition over the parameter space. Pearson correlation is not defined for vector pairs in which $n_1 = N$ or $n_2 = N$; this is evident from equation (4). The proposed similarity measure assigns 0 to these cases, because by definition, $k = n_1$. Thus, even though complete overlap occurs, it is not unexpected.

C. Biological data

Figure 3 describes similarity measure performance, assessed as the percent agreement between the top m/z values selected by each similarity measure and the manual selections. This comparison was carried out across multiple binarization thresholds based on the abundance percentiles of the mean spectrum. For this dataset, the 90th percentile yields top selections from the similarity measures which correspond most closely to the manual selections. The selections of the proposed measure and Pearson correlation correspond highly with the manual selections, and also with each other. The selections of cosine similarity consistently differed from the other two, and from the manual selections.

IV. Discussion

In this study, we propose a hypergeometric similarity measure as a tool for exploration and analysis of TIMS data, with the specific application of identifying m/z values - and hence molecules - with similar spatial distributions in tissue. Due to its definition as the difference

between the upper and lower tails of the hypergeometric distribution, the proposed similarity measure explicitly defines the unexpectedness of any observed overlap. Using synthetic data, the proposed similarity measure was compared with cosine similarity and Pearson correlation in terms of three criteria related to design and performance, and it was shown to perform favorably. Tests on a biological TIMS dataset showed that the proposed similarity measure is effective in identifying visually notable spatial similarities. Together, these results indicate that the proposed similarity measure can play a useful role in the analytical pipeline for TIMS data. Through its application in combination with existing processing techniques, it may be possible to extract additional experimentally meaningful information from TIMS datasets.

Abundance is a key feature of mass spectrometry data, but binary data ignores this information. The selection of appropriate thresholds is an open problem in biomedical image processing, with many different algorithms in use. Results on this biological dataset indicate that increasing the threshold can increase agreement with the set of manually selected m/z values. However, the potential effects of inter-dataset variation on the performance of all three similarity measures have not yet been studied. This will provide more insight into threshold effects and lead to more systematic recommendations for TIMS data. Future work will also address this issue by modifying the similarity measure to more explicitly incorporate abundance. Additionally, future testing will more extensively compare the proposed similarity measure with other measures, including non-linear correlations.

Manual selection of m/z values was necessary to provide a measure of performance for biological data in this study, but subjectivity can hinder interpretation of the results. This issue will be avoided in future work by using labeled TIMS data, *i.e.* where the identity and relationships of m/z values are known. The ‘gold standard’ list of m/z values could be built from molecules with distributions known to be similar from experimental results. This would provide support to spatial analysis by measuring the degree to which similarity measure choices indicate biologically related molecules.

Acknowledgments

We thank Dr. Alfred H. Merrill, Jr., Dr. M. Cameron Sullards and Dr. Yanfeng Chen for sharing the TIMS dataset used in this study.

References

1. McDonnell LA, et al. Mass spectrometry image correlation: quantifying colocalization. *Journal of Proteome Research*. 2008; 7:3619–3627. [PubMed: 18570456]
2. Van de Plas R, et al. Spatial querying of imaging mass spectrometry data: a nonnegative least squares approach. presented at the Neural Information Processing Systems Workshop on Machine Learning in Computational Biology. 2007
3. Curtis RK, et al. Pathways to the analysis of microarray data. *Trends in Biotechnology*. 2005; 23:429–435. [PubMed: 15950303]
4. Drăghici S, et al. Global functional profiling of gene expression. *Genomics*. 2002; 81:98–104.
5. Sadygov RG, Yates JR III. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Analytical Chemistry*. 2003; 75:3792–3798. [PubMed: 14572045]

6. Alfassi ZB. Vector analysis of multi-measurements identification. *Journal of Radioanalytical and Nuclear Chemistry*. 2005; 266:245–250.
7. Hong H, et al. Quality control and quality assessment of data from surface-enhanced laser desorption/ionization (SELDI) time-of-flight (TOF) mass spectrometry (MS). *BMC Bioinformatics*. 2004; 6(Suppl2)
8. Stein SE, Scott DR. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*. 1994; 5:859–866. [PubMed: 24222034]
9. Li X, Dubes RC. A probabilistic measure of similarity for binary data in pattern recognition. *Pattern Recognition*. 1989; 22:397–409.
10. Chvátal V. The tail of the hypergeometric distribution. *Discrete Mathematics*. 1979; 25:285–287.
11. Hoeffding W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*. 1963; 58:13–30.
12. Chen YF, et al. Imaging MALDI mass spectrometry using an oscillating capillary nebulizer matrix coating system and its application to analysis of lipids in brain from a mouse model of Tay-Sachs/Sandhoff disease. *Analytical Chemistry*. 2008; 80:2780–2788. [PubMed: 18314967]
13. Lund, O., et al. *Immunological Bioinformatics*. Cambridge, MA: The MIT Press; 2005.

Figure 1.
Hypergeometric similarity measure (solid), cosine similarity (dot) and Pearson correlation (dash) for $N = 100$, $n_1 = n_2 = 50$.

Figure 2.

Vector pair description (I) and corresponding similarity measure score (r). Color coding for the vector pair segments is as follows. Dark gray: k , number of overlapping 'on' pixels. Black: $n_1 - k$, number of non-overlapping 'on' pixels in image 1. Light gray: $n_2 - k$, number of non-overlapping 'on' pixels in image 2. White: $N - n_1 - n_2 + k$, total number of 'off' pixels. All vectors have length $N = 10$.

Figure 3.
Percentage agreement between similarity measure and manual m/z selections across percentile-based binarization thresholds.