Published in final edited form as:

Comput Stat Data Anal. 2010 April 1; 54(4): 1124. doi:10.1016/j.csda.2009.09.002.

Quantile map: simultaneous visualization of patterns in many distributions with application to tandem mass spectrometry

George C. Tseng^{*}

^{*}George C. Tseng is Assistant Professor, Department of Biostatistics and Human Genetics, University of Pittsburgh, Pittsburgh, PA 15261 (ctseng@pitt.edu).

Abstract

High-throughput experiments have become more and more prevalent in biomedical research. The high-dimensional data have brought new challenges. Effective data reduction, summarization and visualization are important keys to initial exploration in the data mining. In this paper, we introduce a visualization tool, namely quantile map, to present information contained in a probabilistic distribution. We demonstrate its use as an effective visual analysis tool through the application of a tandem mass spectrometry data set. Information of quantiles of a distribution is presented in gradient colors by concentric doughnuts. The width of the doughnuts is proportional to the Fisher information of the distribution to present unbiased visualization effect. A parametric empirical Bayes (PEB) approach is shown to improve the simple maximum likelihood estimate (MLE) approach when estimating the Fisher information. In the motivating example from tandem mass spectrometry data, multiple probabilistic distributions are to be displayed in two-dimensional grids. A hierarchical clustering to reorder rows and columns and a gradient color selection from a Hue-Chroma-Luminance model, similar to that commonly applied in heatmaps of microarray analysis, are adopted to improve the visualization. Both simulations and the motivating example show superior performance of quantile map in summarization and visualization of such high-throughput data sets.

Keywords

pattern visualization; distribution summarization; statistical graphics; computer graphics; Fisher information

1 Introduction

Consider the problem of how to suitably present and communicate the information contained in multiple (few to hundreds) related sets of continuously distributed data. Basic summary statistics often used include (1) the mean, median and mode for measure of location (2) the variance, interquartile range and absolute deviation for measure of dispersion and (3) the skewness, kurtosis and higher-order moments for measure of shape. Such measures attempt to capture important information in an observed distribution and the procedure can be viewed as a data reduction from a set of observations to a few summary statistics. Such summarization is, however, not always effective and sometimes leads to biased information presentation and decision making. Graphics and plots provide a better alternative for such a task. For example,

© 2009 Elsevier B.V. All rights reserved.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

the histogram, density plot and popular box-plot (see Tukey 1977; Benjamini 1988; Frigge et al. 1989; and even a bivariate version in Rousseeuw et al. 1999) are all commonly used in exploratory data mining.

Box-plot is a simple yet powerful tool for displaying a distribution of observations. It has several virtues including: (1) displaying important summary statistics such as the smallest observation, lower quartile, median, upper quartile, the largest observation as well as outliers; (2) allowing parallel comparison of multiple distributions; (3) ease of computation and plotting by hand. However, when it comes to presenting hundreds of distributions, box-plots are no longer feasible. In the research of many new complex and high-throughput data sets from fields such as genomic experiments, biomedical applications, internet problems and business surveys, it is important to develop novel and systematic graphical presentations to concisely, richly and unbiasedly convey information in the data sets. A good example is the application of heatmap to display a data matrix of continuous intensities in a two-dimensional expression profile (Eisen et al, 1998). The rows usually represent thousands of genes in the genome under study and the columns contain tens to hundreds of biological tissue samples under different conditions. The quantitative values in each cell of the data matrix are usually represented by gradient colors: gradient red color for positive values, gradient green color for negative values and black for zero.

In this paper, we consider the problem of simultaneously displaying hundreds of distributions. We propose a method, namely quantile map, that plots quantile information of a distribution in concentric doughnuts within a circle. Fisher information is used to shrink the circle to an ellipse and to adjust the visualization impact for the different amount of information carried in each distribution. The basic concept of quantile map has been briefly introduced and applied in a previous publication (Tseng, 2007) to visually demonstrate the effectiveness of the proposed clustering method. The rigorous definition, concept of Fisher information adjustment for visualization impact and related properties were, however, not covered. In this paper, we aim to provide a systematic introduction of this method and explore the use of this method as an effective visual analysis tool, especially through the application of a tandem mass spectrometry (MS/MS) data set. In Section 2, we first discuss the motivating example from a biological MS/MS experiment. Section 3 introduces the quantile map method. Practical computational issues of Fisher information estimation is covered in Section 4. A simple maximum likelihood estimate (MLE) approach and an improved parametric empirical Bayes (PEB) approach are proposed. In Section 5, we demonstrate the application and performance of quantile map by MLE and PEB approaches in simulated data and in the motivating example. Section 6 contains final discussions.

2 Motivating example

The motivation of this paper comes from a case study of proteomic data. Proteins are the ultimate gene products and major participants in most biological activities. Proteomics is the large-scale study of proteins, particularly their structures and functions, encoded in a genome. Currently, mass spectrometry (MS) experiment is a widely used tool in proteomics analysis to identify the existence and expression abundance of thousands of proteins simultaneously. Among the many variations of MS techniques, tandem mass spectrometry (MS/MS) is one of the most popular platforms.

In MS/MS, proteins are involved in two stages of fragmentation and MS operation, and intensities of their corresponding mass-to-charge ratios (m/z) are recorded. Proteins of interest are first isolated from tissues (step (1) in Figure 1) and sent to the first fragmentation where each protein is digested by a specific enzyme into peptides (normally 8–20 amino acids in length) at pre-specified positions (step (2) in Figure 1). Normally trypsin enzyme that produces

peptides ending with K or R is used at this step. Peptides can then be isolated and detected by m/z ratio in the first MS operation (step (3) in Figure 1). At the second stage, the peptides of the same m/z ratio are isolated and redirected to the second fragmentation, collision-induced dissociation (CID), where low-energy ions collide with the peptides causing them to randomly dissociate into two smaller peptides (step (4) in Figure 1). All of the CID-fragmented small peptides are then sent to the second MS operation to obtain a MS spectrum (step (5) in Figure 1). For example, a group of peptides “AAIANAPR” isolated and obtained from the first enzymatic fragmentation and MS operation are directed to CID. Under CID, some of them are dissociated at the “I-A” position resulting in two peptides “AAI” and “ANAPR”. Others can break at “N-A” yielding “AAIAN” and “APR”, and so on. The intensity and mass-to-charge ratio (m/z) of each type of small peptides can be measured in step (5). It is usually assumed that the intensities measured are proportional to the original abundance of peptides and thus reflect the relative probabilities of dissociation at the corresponding positions in CID fragmentation. Computational algorithms are then applied to match the intensity and m/z information to an existing protein sequence database to identify the original protein product (step (6) in Figure 1).

The left panel of Figure 2 shows an ideal intensity versus m/z spectrum plot. If the differences of two consecutive weights unambiguously match to an amino acid, the peptide sequence can easily be identified in a sequential manner (see right panel; the molecular weight of R is 156, P 116, L 113, etc.). In practice, many of the presumably existing signals are missing and the identification of underlying peptide sequence is a difficult problem. Currently most methods adopt score functions that quantify the correlation of the observed spectrum to each of the theoretical spectra in the protein database. Only a low percentage (10~35%) of proteins can be identified with high confidence (Resing et al. 2004; Elias et al. 2004). The score function methods applied in most protein identification algorithms adopt an implicit assumption of uniform dissociation probability in the CID fragmentation. Recently it has been found that peptides of different sequence motifs can result in different dissociation patterns in CID (Huang et al. 2005), which violates the assumption in score function methods and is the major cause of the low identification accuracy. For example, peptides having a +1 positive charge, containing P in the middle, no H in the middle and ending with R (denoted as $[...P...R]^{1+}$) are found to have higher fragmentation probabilities at “D-X” and “E-X” positions where ‘X’ represents any amino acid. On the other hand, peptides of the same sequence motif with a +2 positive charges (denoted as $[...P...R]^{2+}$) are found to have higher fragmentation probabilities at “X-P” positions (see Figure 4 & Figure 8 and explanation in later text). In published studies so far, prior chemistry knowledge of the sequence motifs is needed, *a priori*, to enhance discovery of fragmentation patterns; yet the total number of known sequence motifs is still very limited. In our recent publication (Huang et al. 2008), a pure data mining scheme was introduced to re-discover important sequence motifs responsible for various fragmentation patterns without the aid of prior chemistry knowledge. The result elucidated more uncovered sequence motifs and underlying chemical interactions. Incorporation of such information will ultimately improve protein identification algorithms. The visualization tool we propose in this paper was essential in the development of the data mining scheme in Huang et al. (2008).

The complete data to be analyzed are described below. The data set contains MS intensities of 28,330 peptides for which the charge state and peptide sequence information was rigorously verified by other techniques (Huang et al. 2005). Each peptide is a sequence of residues of variable length (usually 8–20) and each residue takes one of the 20 possible amino acids, $AA=\{AA_i, i=1, \dots, 20\}=\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$. The experiment produces a sequence of intensities with their corresponding mass-to-charge (m/z) values that can be mapped to a fragmentation location by calculating the total weight of amino acids and charge. We denote by $S = \{s_i, i=1, \dots, 28,330\}$ the total peptides analyzed, where

$s_i = \{a_1^{(i)}, a_2^{(i)}, \dots, a_{n(s_i)}^{(i)}, c^{(i)}\}$ represents the sequence and charge state ($c^{(i)}$) of peptide i , $a_j^{(i)} \in AA$ and $n(s_i)$ is the length of the peptide. For each peptide s_i , the fragmentation intensity¹ at the " $a_j^{(i)} - a_{j+1}^{(i)}$ " location is denoted as $\lambda_{a_j^{(i)} a_{j+1}^{(i)}}^{(i)}$, $1 \leq j \leq n(s_i) - 1$. We note that intensities are normalized within each peptide so that $\sum_j \lambda_{a_j^{(i)} a_{j+1}^{(i)}}^{(i)} = 1$ and the normalized intensities are reasonable estimates of the CID dissociation probabilities. For a given subset of peptides $S' \in S$, the dissociation distributions are defined as the collection of all dissociation intensities at each AA_m-AA_n location in peptide set S' : $\Lambda_{mn}(S')$

$$= \left\{ \lambda_{a_j^{(i)} a_{j+1}^{(i)}}^{(i)}, a_j^{(i)} = AA_m, a_{j+1}^{(i)} = AA_n, s_i \in S', 1 \leq j \leq n(s_i) - 1 \right\}, 1 \leq m, n \leq 20.$$

Our goal is to provide an adequate visualization tool to present the patterns of the 400 distributions, $\Lambda_{mn}(S')$.

3 Quantile map

We propose the following quantile map for visualizing a random variable $X: \Omega \rightarrow \mathcal{R}$. Denote by $F_X(x)$ the corresponding cumulative distribution function of X . A quantile map consists of d concentric doughnuts, $D_1, \dots, D_d$. For each doughnut D_i ($1 \leq i \leq d$), the color $C(F_X^{-1}((i - 0.5)/d))$ is applied where the color system $C(\cdot)$ transforms the support of X to a series of gradient colors usually from bright (small value) to dark (large value). More specifically, we specify $C(\chi_{min}) = (R_{min}, G_{min}, B_{min})$ the RGB color presented for the low-end displaying value, χ_{min} , and similarly $C(\chi_{max}) = (R_{max}, G_{max}, B_{max})$ for the high-end displaying value, χ_{max} . We will discuss later how to select χ_{min} and χ_{max} when visualizing multiple random variables. Then, throughout this paper, we define the following gray-scale visualization $C(x) = (R(x), G(x), B(x))$ where $(R_{min}, G_{min}, B_{min}) = (256, 256, 256)$, $(R_{max}, G_{max}, B_{max}) = (0, 0, 0)$ and

$$R(x)=G(x)=B(x)=\begin{cases} R_{max} & \text{if } x > \chi_{max} \\ R_{min} + (R_{max} - R_{min}) \cdot \frac{x - \chi_{min}}{\chi_{max} - \chi_{min}} & \text{if } \chi_{max} \geq x \geq \chi_{min} \\ R_{min} & \text{if } x < \chi_{min} \end{cases} \quad (1)$$

Figure 3 shows several examples of quantile maps. The upper two examples are unimodal distributions with concentrated weights in the maximum or minimum region. The bottom example is a bimodal distribution. We note that it is possible to define a different color scale by varying different $(R_{min}, G_{min}, B_{min})$ and $(R_{max}, G_{max}, B_{max})$. For example, $(R_{min}, G_{min}, B_{min}) = (0, 0, 0)$ and $(R_{max}, G_{max}, B_{max}) = (256, 0, 0)$ gives a red-scale visualization with bright red coloring for high values of x and black coloring for small values. It has been found that gradient colors visually may not always be the most informative. Discrete qualitative classes from cluster analysis provide more concise presentation in many applications (Gribov et al. 2006). The estimation of the number of classes, however, poses another difficulty. In general, the choices of colors and glyph representations depend on the biological goals of an analysis. The default gray color scheme seems to serve the biological purpose we desire for the MS/MS project. In Section 5.2, we will discuss a gradient color scheme from a Hue-Chroma-Luminance model for improved visualization.

¹There are two commonly seen ion types, y-ion and b-ion, measured in MS/MS data. After a peptide is fragmented into two sub-peptides, the y-ion type measures the intensity of the C-terminal fragment (i.e. the downstream sub-peptide) and the b-ion measures the N-terminal fragment (i.e. the upstream sub-peptide). For simplicity, we only consider the y-ion data in this paper and the fragmentation intensities represent the intensities of the down-stream sub-peptides measured. For more details, refer to Huang et al. 2005.

In most applications, we have i.i.d. observations $\{x_1, \dots, x_n\}$ of a random variable X . The quantile map can be used to visualize the empirical distribution. For example, if $d = 10$, the quantile map simply draws the corresponding colors of the 5%, 15%, ..., 95% quantiles in 10 concentric doughnuts; this is recommended in most applications. If $d = n$, every observation is presented in the quantile map.

A proper choice of the displaying range $[\chi_{min}, \chi_{max}]$ depends on the multiple empirical distributions to be visualized simultaneously. In the motivating example in Section 2, we set χ_{min} to be the 5% quantile of all the observations $\Lambda_{mn}(S)$ and χ_{max} to be the 95% quantile.

4 Fisher information and its estimation

The quantile map in Section 3 is introduced for visualizing a single distribution. When visualizing multiple distributions simultaneously, we should be aware of the different amount of information each single distribution contributes to visualization. For example, the quantile map of 10 i.i.d. samples from $N(2,1)$ is very similar to 1000 random samples of $N(2,1)$. The statistical information contained in the later distribution, however, is much stronger and should be reflected in the visualization. The variance of the distribution is also relevant to the statistical information contained. Thus we design the width of quantile map to be proportional to the Fisher information. Intuitively the more observations available and the less variation in a distribution, the larger Fisher information it contains and thus the larger quantile map should be plotted when multiple distributions are displayed together for visualization.

4.1 Fisher information and simple MLE

The Fisher information measures the amount of information that distribution X carries about the underlying parameter θ upon which the likelihood function of X , $f(X; \theta)$, depends on. From now on, we consider visualization of I observed distributions: $Y_i = y_{i1}, \dots, y_{iJ_i}$ ($1 \leq i \leq I$), where J_i is the number of observations in the i -th distribution. Under the normal assumption, we have $Y_i = y_{i1}, \dots, y_{iJ_i} \sim N(\mu_i, \sigma_i^2)$, where we assume μ_i unknown and σ_i^2 known. The Fisher information that Y_i contains for μ_i is $FI_i = J_i / \sigma_i^2$. Replacing σ_i^2 with the maximum likelihood estimate (MLE) sample variance $\hat{\sigma}_{iMLE}^2 = \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_i)^2 / J_i$ (where $\bar{y}_i = \sum_{j=1}^{J_i} y_{ij} / J_i$), the estimated Fisher information becomes $\hat{FI}_{iMLE} = J_i / \hat{\sigma}_{iMLE}^2$.

4.2 Parametric empirical Bayes estimator

When visualizing many empirical distributions at the same time, it is common to encounter extreme situations where the number of observations is too small or the standard error is close to 0, in which case the MLE estimator becomes unavailable or unstable. To alleviate such a problem, we introduce the following parametric empirical Bayes (PEB) approach (Carlin and Louis 2000) to estimate the standard deviation, σ . In addition to the Gaussian assumption in Section 4.1, we assume a uniform prior for μ_i and conjugate prior for $\sigma_i^2: \mu_i \propto 1$ and $\sigma_i^2 \mid \nu_0, \sigma_0^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$. The hyperparameters, ν_0 and σ_0^2 , are estimated from the conditional distribution:

$$(\hat{\nu}_0, \hat{\sigma}_0^2) = \arg \max_{\nu_0, \sigma_0^2} f(y_{ij})$$

$$, i=1, \dots, I, j=1, \dots, J_i \mid \nu_0, \sigma_0^2 \quad (2)$$

The posterior modes of (μ_i, σ_i^2) can then be estimated by the complete log-likelihood function:

$$(\hat{\mu}_{i:PEB}, \hat{\sigma}_{i:PEB}^2) = \arg \max_{\mu_i, \sigma_i^2} \log f(\mu_i, \sigma_i^2 | y_{ij}, i=1, \dots, I, j=1, \dots, J_i, \hat{v}_0, \hat{\sigma}_0^2) \quad (3)$$

The estimated Fisher information is then defined as $\widehat{FI}_{i:PEB} = J_i / \hat{\sigma}_{i:PEB}^2$. Detailed inference and optimization for the empirical Bayes procedures in equation (2) and (3) are presented in the Appendix.

5 Simulation and application

5.1 Simulation

We first perform the following simulation that exactly follows the assumptions in our PEB approach: $f(y_{ij} | \mu_i, \sigma_i^2) \sim N(\mu_i, \sigma_i^2)$ ($1 \leq i \leq I$ and $1 \leq j \leq J_i$), $J_i \sim \text{Poisson}(\lambda)$, $\mu_i \propto 1$ and $\sigma_i^2 | v_0, \sigma_0^2 \sim \text{Inv} - \chi^2(v_0, \sigma_0^2)$. We let $I = (5, 400)$, $\lambda = (5, 10, 40)$, $v_0 = 10$ and $\sigma_0^2 = (0.05, 0.1, 0.2)$. Given the simulated data set, the Fisher information estimators $\widehat{FI}_{i:MLE}$ and $\widehat{FI}_{i:PEB}$ ($1 \leq i \leq I$) are calculated. Since the true Fisher information for each distribution, $FI_i = J_i / \sigma_i^2$, is known, we define a standardized root mean squared error (S-RMSE) to measure the deviation of the estimated Fisher information from the truth:

$$S\text{-}RMSE = \sqrt{\frac{\sum_{i=1}^I \left(\frac{\widehat{FI}_i - FI_i}{FI_i} \right)^2}{I}}.$$

Table 1 shows the mean and standard error of S-RMSE over 1000 simulated data sets under different λ and σ_0^2 settings. We can clearly see that when λ is small, the number of observations of each distribution is limited and the simple MLE estimator is unstable with large S-RMSE. When λ is large, MLE and PEB perform similarly.

We next examine the sensitivity when both Gaussian and hierarchical model assumptions are violated in MLE and PEB approaches. Suppose the data are from the family of the gamma distribution: $f(y_{ij} | \alpha, \theta_i) \sim \text{Gamma}(\alpha, \theta_i)$ ($1 \leq i \leq I$ and $1 \leq j \leq J_i$) with a known shape parameter α and unknown scale parameter θ_i , $J_i \sim \text{Poisson}(\lambda)$ and $\theta_i \sim \text{Unif}(0, 3)$. We perform simulations on $\lambda = (5, 10, 40, 100)$ and $\alpha = (1, 10, 100)$. We can easily show that the Fisher information for the mean $\mu_i = \alpha \cdot \theta_i$ becomes $FI_i = J_i / \alpha \cdot \theta_i^2 = J_i / \sigma_i^2$, where $\sigma_i^2 = \alpha \cdot \theta_i^2$ is the variance of Gamma (α, θ_i). As a result, $\widehat{FI}_{i:MLE}$ is essentially a method of moment estimator (MME) for FI_i . Table 2 shows the S-RMSE of $\widehat{FI}_{i:MLE}$ and $\widehat{FI}_{i:PEB}$. Again we see that $\widehat{FI}_{i:MLE}$ works poorly when λ is small. When λ and α are extremely large, $\widehat{FI}_{i:PEB}$ can be slightly worse than $\widehat{FI}_{i:MLE}$ because of violation of distributional assumptions. In general, PEB approach is robust and outperforms MLE estimate in a wide range of parameter settings.

5.2 Application to the motivating example

We investigate the application of our quantile map to the motivating example introduced in Section 2. We will concentrate on the presentation of the $S_1 = [\dots P \dots R]^1$ and $S_2 = [\dots P \dots R]^2$ data sets. Figure 4 and Figure 8 show median plots for S_1 and S_2 that are similar to the original presentation by our collaborators (Huang et al. 2004). For each dissociation

position, the median of the distributions are presented according to the grey-scale color scheme. It is clearly seen that “D-X” and “E-X” have high dissociation intensities in S_1 and “X-P” in S_2 . In Figure 5 and Figure 9, quantile maps with Fisher information calculated from MLE estimation are shown for S_1 and S_2 respectively. We note that “C-G” and “C-P” in S_1 both have only one observation with value 1 and thus have high median value in Figure 4 but are not displayed in Figure 5. Compared to Figure 8’s median plot in S_2 , the quantile map in Figure 9 shows more clearly the bimodal patterns in “L-X” and “V-X” and maybe “I-X”.

Quantile maps with Fisher information from PEB estimation are shown in Figure 6 and Figure 10. To compare the MLE and PEB estimations, the scatter plots of log-transformed square root Fisher information estimation from MLE and PEB are shown in Figure 7 and Figure 11. In general, when the number of observations is large, Fisher information estimations from MLE and PEB are similar (points around $x=y$ dotted line). The PEB approach tends to stabilize unstable MLE estimations of distributions when the number of observations is small. For example, “P-R” and “Y-Y” in S_1 both have only two observations but very small standard deviations. Thus the estimations $\widehat{FI}_{MLE}$ are extremely large, which is unreasonable since each distribution only contains two observations. $\widehat{FI}_{PEB}$ are, however, stabilized to more reasonable smaller values (Figure 7 and Table 3). An extreme case happens at “E-W”. The distribution contains two observations, (1,1), resulting in $\widehat{FI}_{MLE}=\infty$. $\widehat{FI}_{PEB}$ is, however, stabilized to 9.62 (Table 3). The opposite situation of small MLE estimation but large PEB estimation happens at “N-W” with two observations, (0.032, 1). The very large standard deviation gives a small MLE estimation ($\widehat{FI}_{MLE}=2.92$) but PEB approach stabilizes it to a larger estimation ($\widehat{FI}_{PEB}=5.54$). Similar examples are also found for S_2 data in Figure 11 (See details in Table 3).

To further improve the visualization, we first sort the rows and columns in quantile map plots by hierarchical clustering (with complete linkage), a common practice in heatmap. In the hierarchical clustering implementation, we define a measure of two vectors of random variables

$$D((V_1, \dots, V_K), (U_1, \dots, U_K))$$

by
$$= (\sum_{k=1}^K d(V_k, U_k)) / K$$
, where $D(\cdot, \cdot)$ denotes a distance measure of two random vectors and $d(\cdot, \cdot)$ denotes a measure of two random variables. One can think of any number of statistics to measure the difference of two independent random variables. Here we adopt an

intuitive measure $d(U, V) = \int_0^1 |F_U^{-1}(p) - F_V^{-1}(p)| dp$. It is easily seen that $d(U, V) = 0$ if $U \stackrel{D}{=} V$, $d(U+b, V) = b$ if $U \stackrel{D}{=} V$ and b a constant, and $d(U, V) \leq d(U, W) + d(W, V)$ for any random variables U, V and W . Secondly, instead of using gray-scale color scheme, we adopt the perceptually based color model of Hue-Chroma-Luminance (HCL) discussed in Zeileis et al. (2009). Figure 12 and 13 shows the new quantile maps after deleting amino acids that have more than half missing Fisher information estimates, reordering rows and columns by hierarchical clustering and adopting the new HCL color. The result shows similar patterns of “D-X” and “E-X”, “X-G” and “X-P”, and “X-D” and “X-E” in S_1 . In S_2 , “X-P” has strong dissociation intensities, very different from all the other amino acids. “I-X” and “V-X” (and probably “L-X”) have a similar bimodal pattern as discussed before. It should be noted that I, V and L are the three most hydrophobic amino acids (hydropathy index I=4.5, V=4.2 and L=3.8) that might contribute to the similar dissociation behavior. For a given hierarchical tree, there are many possible serialiations to reorder the amino acids in each dimension. Systematic serialiations for specific visualization preferences can be referred to Hahsler et al. (2008).

6 Discussion

In this paper, we consider the problem of simultaneous visualization of many distributions to capture the underlying patterns in the data. The quantile map presented in this paper can

virtually convey entire information of a distribution within the area of a circle. The relative information (Fisher information) contained in each distribution is reflected by shrinking the width of the circles. A parametric empirical Bayes approach is introduced to alleviate the unstable MLE estimates in extreme situations of the distributions. Both simulated and real data demonstrate better performance of the proposed method.

We propose to shrink the width of the circles by Fisher information in this paper. A simple alternative may use sample size instead (i.e. dropping the variance term from Fisher information since information of variance is already displayed in the quantiles). This alternative is, however, problematic in some extreme situations. For example, an observation (1,1,1,1) is very informative and should be displayed with larger width although it has a small number of four samples. Despite such extreme cases, simple shrinkage by sample size can be useful enough in some applications. There are two major limitations in our proposed method. Firstly, the support of each distribution should be in similar range, as we have seen in the MS motivating example, so the quantile information of each distribution can be shown by gradient colors. Otherwise some normalization or scaling should be applied. When the supports greatly vary across distributions, such as that we commonly encounter in microarray data discussed below, the quantiles of a spot are likely mapped to the same color and the Fisher information becomes important to control the visualization impact. Secondly, current Fisher information calculation depends on normality assumption, which is true in many experimental data sets. When the data set follows a different parametric distribution, the Fisher information calculation should be modified accordingly.

The proposed Fisher information shrinkage can also find applications in microarray studies. Many data from the cDNA microarray platform contain contaminated or scratched cDNA probes. Normally some quality index (such as the standard deviation of intensities of the hundreds of pixels in a probe) estimated from image analysis is assessed and an arbitrary threshold may be applied to flag problematic probes as missing. Since such filtering may be arbitrary, another alternative is to incorporate such quality index in down-stream analysis. For example, Yeung et al. (2003) applied variability-weighted distance to gene cluster analysis when repeated measurements (i.e. multiple probes correspond to one gene) exist. Applying Fisher information shrinkage at the log-transformed intensity scale can help reflect the information unbiasedly. If a probe is scratched or printed non-uniformly, the width of the circle will shrink to reduce the visualization impact. It is worth notice that microarray data are found to have variations proportional to mean measurements. The log-transformation can usually alleviate the problem and Fisher information can still apply.

Many modern data mining problems and statistical applications are high-throughput in nature. By better utilizing the growing power of computing and plotting/printing capacities, more visualization tools can be realized to better interpret and analyze the data.

Acknowledgments

The author is partially supported by the National Institute of Health (1 KL2 RR024154-02); University of Pittsburgh (Central Research Development Fund, CRDF; Competitive Medical Research Fund, CMRF). Data used in the motivating example were provided by Vicki Wysocki's lab at the University of Arizona. The author would like to thank Yingying Huang and Shinsheng Yuan for discussion and insightful comments. The collaboration of the motivating project was initiated in the Institute of Pure and Applied Mathematics (IPAM) at UCLA, funded by NSF. The author would like to thank the associate editor and two reviewers for many insightful comments and suggestions to improve the paper.

Appendix

1. Empirical Bayes to estimate hyperparameters

$$\begin{aligned}
 & f(y_{ij} \mid \mu_i, \sigma_i^2) \sim N(\mu_i, \sigma_i^2), \mu_i \propto 1 \text{ and } \sigma_i^2 \mid \nu_0, \sigma_0^2 \sim \text{Inv} \\
 \text{Assume } & -\chi^2(\nu_0, \sigma_0^2) \quad . \text{ Then} \\
 & = \int \cdots \int f(y_{ij}, \forall i, j \mid \mu_i, \sigma_i^2, \forall i) \cdot f(\mu_i, \sigma_i^2 \mid \nu_0, \sigma_0^2) d\mu_1 \cdots d\mu_I d\sigma_1^2 \cdots d\sigma_I^2 \\
 & = \int \cdots \int \prod_{i=1}^I \prod_{j=1}^{J_i} f(y_{ij} \mid \mu_i, \sigma_i^2) \cdot \prod_{i=1}^I f(\mu_i, \sigma_i^2 \mid \nu_0, \sigma_0^2) d\mu_1 \cdots d\mu_I d\sigma_1^2 \cdots d\sigma_I^2 \\
 & = \prod_{i=1}^I \left[\int \int \prod_{j=1}^{J_i} f(y_{ij} \mid \mu_i, \sigma_i^2) \cdot f(\mu_i, \sigma_i^2 \mid \nu_0, \sigma_0^2) d\mu_i d\sigma_i^2 \right] \\
 & = \prod_{i=1}^I \left[\frac{h(\nu_0, \sigma_0^2)}{h(\nu_0 + J_i - 1, \frac{\nu_0 \sigma_0^2 + (J_i - 1) \hat{\sigma}_i^2}{\nu_0 + J_i - 1})} \right] \tag{4}
 \end{aligned}$$

$$\begin{aligned}
 h(\nu, \sigma^2) &= \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \cdot (\sigma^2)^{\nu/2} \\
 \text{where } \hat{\sigma}_i^2 &= \frac{\sum_{j=1}^{J_i} (y_{ij} - y_{i\cdot})^2}{J_i - 1} \text{ and } y_{i\cdot} = \frac{\sum_{j=1}^{J_i} y_{ij}}{J_i}. \text{ We will obtain estimate of} \\
 & \text{hyperparameters } (\nu_0, \sigma_0^2) \text{ by}
 \end{aligned}$$

$$\begin{aligned}
 & (\hat{\nu}_0, \hat{\sigma}_0^2) \\
 & = \arg \max_{\nu_0, \sigma_0^2} \log f(y_{ij}, 1 \leq i \leq I, 1 \leq j \leq J_i \mid \nu_0, \sigma_0^2) \\
 & = \arg \max_{\nu_0, \sigma_0^2} \sum_{i=1}^I \log (h(\nu_0, \sigma_0^2)) - \log \left(h(\nu_0 + J_i - 1, \frac{\nu_0 \sigma_0^2 + (J_i - 1) \hat{\sigma}_i^2}{\nu_0 + J_i - 1}) \right) \tag{5}
 \end{aligned}$$

Since the maximizer of (ν_0, σ_0^2) in equation (5) does not have a closed form solution, we obtain a crude estimator for (ν_0, σ_0^2) from method of moments and apply it as an initial value to solve equation (5) with a Newton-type minimization algorithm (“nlm”) in R. The method of moments solves the following equation

$$\begin{cases} E(\sigma_i^2) = \frac{\nu_0}{\nu_0 - 2} \cdot \sigma_0^2 = \tilde{\sigma}^2 \\ \text{Var}(\sigma_i^2) = \frac{2\nu_0^2}{(\nu_0 - 2)^2(\nu_0 - 4)} \cdot \sigma_0^4 = \frac{\sum_{i=1}^I (\hat{\sigma}_i^2 - \tilde{\sigma}^2)^2}{I - 1}, \end{cases}$$

where $\tilde{\sigma}^2 = \sum_{i=1}^I \hat{\sigma}_i^2 / I$, $\hat{\sigma}_i^2 = \sum_{j=1}^{J_i} (y_{ij} - \hat{\mu}_i)^2 / (J_i - 1)$ and $\hat{\mu}_i = \sum_{j=1}^{J_i} y_{ij} / J_i$. The initial value for

“nlm” to solve equation (5) is $\hat{\nu}_0^{(0)} = 4 + \frac{2(I - 1)(\tilde{\sigma}^2)^2}{\sum_{i=1}^I (\hat{\sigma}_i^2 - \tilde{\sigma}^2)^2}$ and $\hat{\sigma}_0^{2(0)} = \frac{\hat{\nu}_0 - 2}{\hat{\nu}_0} \cdot \tilde{\sigma}^2$.

2. Calculating the estimated Fisher information of each distribution

With the estimated hyperparameters $(\widehat{\nu}_0, \widehat{\sigma}_0^2)$, the posterior mean and variance of each distribution can be obtained by

$$\begin{aligned}
 L &= f(\mu_i, \sigma_i^2, 1 \leq i \leq I | y_{ij}, 1 \leq i \leq I, 1 \leq j \leq J_i, \widehat{\nu}_0, \widehat{\sigma}_0^2) \\
 &\propto f(y_{ij}, 1 \leq i \leq I, 1 \leq j \leq J_i | \mu_i, \sigma_i^2, 1 \leq i \leq I) \cdot f(\mu_i, \sigma_i^2, 1 \leq i \leq I | \widehat{\nu}_0, \widehat{\sigma}_0^2) \\
 &= \prod_{i=1}^I \prod_{j=1}^{J_i} f(y_{ij} | \mu_i, \sigma_i^2) \cdot \prod_{i=1}^I f(\mu_i, \sigma_i^2 | \widehat{\nu}_0, \widehat{\sigma}_0^2) \\
 &\propto \prod_{i=1}^I \left[\left(\frac{1}{\sigma_i} \right)^{J_i} \cdot \exp \left(-\frac{\sum_{j=1}^{J_i} (y_{ij} - \mu_i)^2}{2\sigma_i^2} \right) \cdot (\sigma_i^2)^{-(\widehat{\nu}_0/2+1)} \cdot \exp \left(-\frac{\widehat{\nu}_0 \cdot \widehat{\sigma}_0^2}{2\sigma_i^2} \right) \right] \\
 &= \prod_{i=1}^I \left[(\sigma_i)^{-(\frac{\widehat{\nu}_0+J_i}{2}+1)} \cdot \exp \left(-\frac{\sum_{j=1}^{J_i} (y_{ij} - \mu_i)^2 + \widehat{\nu}_0 \widehat{\sigma}_0^2}{2\sigma_i^2} \right) \right]
 \end{aligned}$$

A good estimate of (μ_i, σ_i^2) can be obtained by the posterior modes $(\widehat{\mu}_i, \widehat{\sigma}_i^2)$ where

$$\begin{aligned}
 (\widehat{\mu}_i, \widehat{\sigma}_i^2) &= \arg \max_{\mu_i, \sigma_i^2} L \\
 &= \arg \max_{\mu_i, \sigma_i^2} \log f(\mu_i, \sigma_i^2, 1 \leq i \leq I | y_{ij}, 1 \leq i \leq I, 1 \leq j \leq J_i, \widehat{\nu}_0, \widehat{\sigma}_0^2) \\
 &= \arg \max_{\mu_i, \sigma_i^2} \sum_{i=1}^I - \left(\frac{\widehat{\nu}_0 + J_i + 2}{2} \right) \cdot \log(\sigma_i^2) - \frac{\sum_{j=1}^{J_i} (y_{ij} - \mu_i)^2 + \widehat{\nu}_0 \widehat{\sigma}_0^2}{2\sigma_i^2}.
 \end{aligned}$$

Taking partial derivatives, we obtain

$$\begin{cases} \frac{\partial L}{\partial \mu_i} = 2 \sum_{j=1}^{J_i} (y_{ij} - \mu_i) = 0 \\ \frac{\partial L}{\partial \sigma_i^2} = -\frac{\widehat{\nu}_0 + J_i + 2}{2\sigma_i^2} + \frac{\sum_{j=1}^{J_i} (y_{ij} - \mu_i)^2 + \widehat{\nu}_0 \widehat{\sigma}_0^2}{2(\sigma_i^2)^2} = 0. \end{cases}$$

$$\text{Thus } (\widehat{\mu}_i, \widehat{\sigma}_i^2) = \left(\frac{\sum_{j=1}^{J_i} y_{ij}}{J_i}, \frac{\sum_{j=1}^{J_i} (y_{ij} - \widehat{\mu}_i)^2 + \widehat{\nu}_0 \widehat{\sigma}_0^2}{\widehat{\nu}_0 + J_i + 2} \right) \text{ and } \widehat{FI}_i = \frac{J_i}{\widehat{\sigma}_i^2}.$$

References

- Benjamini Y. Opening the Box of a Boxplot. The American Statistician 1988;42:257-262.
- Carlin, BP.; Louis, TA. Bayes and Empirical Bayes Methods for Data Analysis. Second edition. Chapman & Hall/CRC; 2000.
- Eisen MB, Spellman PT, Brown PO, Botstein D. Genetics Cluster analysis and display of genome-wide expression patterns. PNAS 1998;95:14863–14868. [PubMed: 9843981]
- Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP. Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nature Biotechnology 2004;22:214–219.
- Frigge M, Hoaglin DC, Iglewicz B. Some Implementations of the Boxplot. The American Statistician 1989;43:50-54.

- Gribov A, Unwin A, Hofmann H. About Glyphs and Small Multiples: Gauguin and the Expo. *Statistical Computing and Graphics Newsletter* 2006;17:18–22.
- Hahsler M, Hornik K, Buchta C. Getting things in order: An introduction to the R package seriation. *Journal of Statistical Software* 2008;25:1–34.
- Huang Y, Triscari JM, Wysocki VH, Pasa-Tolic L, Anderson GA, Lipton MS, Smith RD. Dissociation Behaviors of Doubly-Charged Tryptic Peptides: Correlation of Gas-Phase Cleavage Abundance with Ramachandran Plots. *J. Am. Chem. Soc* 2004;126:3034–3035. [PubMed: 15012117]
- Huang Y, Triscari JM, Tseng GC, Pasa-Tolic L, Lipton MS, Smith RD, Wysocki VH. Statistical Characterization of Charge State and Residue Dependence of Low Energy CID Peptide Dissociation Patterns. *Analytical Chemistry* 2005;77:5800–5813. [PubMed: 16159109]
- Huang, Y. Dissertation. Department of Chemistry, University of Arizona; 2005. Data mining of peptide MS/MS spectra to elucidate gas phase peptide dissociation mechanisms and improve protein identification.
- Huang Y, Tseng GC, Yuan S, Pasa-Tolic L, Lipton MS, Smith RD, Wysocki VH. A data mining scheme for identifying peptide structural motifs behind different MS/MS fragmentation intensity patterns. *Journal of Proteome Research* 2008;7:70–79. [PubMed: 18052120]
- Resing KA, Meyer-Arendt K, Mendoza AM, Aveline-Wolf LD, Jonscher KR, Pierce KG, Old WM, Cheung HT, Russell S, Wattawa JL, Goehle GR, Knight RD, Ahn NG. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Analytical Chemistry* 2004;76:3556–3568. [PubMed: 15228325]
- Rousseeuw PJ, Ruts I, Tukey JW. The Bagplot: A Bivariate Boxplot. *The American Statistician* 1999;53:382–387.
- Tseng GC. Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data. *Bioinformatics* 2007;23:2247–2255. [PubMed: 17597097]
- Tukey, JW. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley; 1977.
- Yeung KY, Medvedovic M, Bumgarner RE. Clustering gene-expression data with repeated measurements. *Genome Biology* 2003;4:R34. [PubMed: 12734014]
- Zeileis A, Hornik K, Murrell P. Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis* 2009;53:3259–3270.

Figure 1.

Introduction to tandem mass spectrometry experiment. This figure is modified from Yingying Huang's dissertation (2005).

Figure 2.
Example of an ideal MS spectrum (left) and how the peptide sequence can be inferred (right).

Figure 3.

Examples of quantile map, density plot and box-plot for displaying observed i.i.d. samples. Upper: A unimodal distribution concentrated around 1. Middle: A unimodal distribution concentrated around 0. Lower: A bi-modal distribution concentrated around 0 and 1.

Figure 4. Median plot for S_1 . The amino acids down the left side are the residues before the dissociated peptide bond and the amino acids across the top are the residues after the dissociated position.

Figure 5.
Quantile map by MLE estimator for S_1 .

Figure 6.
Quantile map by PEB estimator for S_1 .

Figure 7. Scatter plot of log-transformed square root Fisher information estimated by MLE and PEB in S_1 .

Figure 8. Median plot for S_2 . The amino acids down the left side are the residues before the dissociated peptide bond and the amino acids across the top are the residues after the dissociated position.

Figure 9.
Quantile map by MLE estimator for S_2 .

Figure 10.
Quantile map by PEB estimator for S_2 .

Figure 11. Scatter plot of log-transformed square root Fisher information estimated by MLE and PEB in S_2 .

Figure 12. Quantile map by PEB estimator for S_1 after row and column reordering by hierarchical clustering.

Figure 13.
Quantile map by PEB estimator for S_2 after row and column reordering by hierarchical clustering.

Table 1

Simulation results when data are from a Gaussian hierarchical model.

I	σ_0^2	$\lambda=5$ MLE	PEB	$\lambda=10$ MLE	PEB	$\lambda=40$ MLE	PEB
400	0.05	15.2(4.6)	.45(.003)	.63(.09)	.30(.002)	.125(.0005)	.125(.0005)
	0.1	9.5(2.2)	.46(.003)	.71(.09)	.30(.002)	.126(.0007)	.126(.0007)
	0.2	9.3(2.0)	.45(.003)	.66(.07)	.30(.002)	.128(.0006)	.127(.0006)
5	0.05	2.6(.93)	.43(.009)	.40(.02)	.27(.004)	.119(.001)	.120(.002)
	0.1	1.5(.12)	.44(.009)	.45(.06)	.27(.005)	.119(.001)	.119(.001)
	0.2	2.6(.76)	.44(.009)	.39(.01)	.28(.005)	.118(.001)	.119(.001)

Table 2
Simulation results when Gaussian and hierarchical model assumptions are violated.

I	α	$\lambda=5$ MLE	PEB	$\lambda=10$ MLE	PEB	$\lambda=40$ MLE	PEB	$\lambda=100$ MLE	PEB
400	1	5.8E7	2.39	2.4E4	1.63	.617	.628	.32	.403
	10	2.2E8	1.76	362	1.08	.314	.455	.175	.33
	100	1.5E5	1.69	48	1.03	.281	.439	.156	.323
5	1	1979	2.50	3.39	1.35	.537	.484	.295	.283
	10	353	1.58	3.55	.81	.282	.277	.165	.164
	100	149	1.39	1.10	.79	.254	.257	.148	.149

Table 3

Selected distributions from the charge 1 and charge 2 data sets.

data set	dissociation position	observations	$\hat{FI}_{MLE}$	$\hat{FI}_{PEB}$
[...P...H...R] ¹⁺	P-R	(0.037, 0.044)	363.93	9.62
	Y-Y	(0.012, 0.023)	253.22	9.62
	E-W	(1, 1)	∞	9.62
	N-W	(0.032, 1)	2.92	5.54
[...P...H...R] ²⁺	C-G	(0, 0.014)	195.23	7.70
	C-P	(0.975, 1)	113.60	7.70
	V-C	(0.024, 1)	2.90	5.23