

NORMALIZATION, BASELINE CORRECTION AND ALIGNMENT OF HIGH-THROUGHPUT MASS SPECTROMETRY DATA

Anne C. Sauve[#] and Terence P. Speed^{#*}

[#]Department of Statistics, University of California, Berkeley, ^{*}Division of Genetics and Bioinformatics, The Walter and Eliza Hall Institute, Australia

ABSTRACT

We propose several preprocessing steps to be used before biomarker clustering or classifying for high-throughput Mass Spectrometry (MS) data. These preprocessing steps for the mass spectra are multiple alignment of technical replicates, baseline correction and normalization along the mass/charge axis. While the benefits from baseline correction and alignment seem obvious we studied more carefully the benefit from normalizing using some human prostate cancer SELDI TOF MS data (obtained from the Virginia Prostate Center Tissue and body Fluid Bank and approved by the Eastern Virginia Medical School). We show on these data that our global normalization by scaling helps in distinguishing between different cancer groups as well as between cancer and non-cancer groups. We used the Between to Within sum of squares ratio introduced by Fisher as well as visual inspection to illustrate the improvement brought by the normalization.

1. INTRODUCTION

Retentate chromatography has been established as an important method to fractionate biological fluid sample. Chromatographic systems can be coupled to more sophisticated detection devices such as Mass Spectrometers. Together they have the capacity to generate voluminous spectra characterized by their retention axis and mass axis and quantified by the measured intensities [11]. Today's high-throughput MS detectors can generate numerous such spectra per patient. Because of non-linearity in the detector response, ionization suppression, minor changes in the mobile phase composition and interaction between analytes, undesirable variation may get introduced in the MS data. In order to perform an analysis of the entire collected data it is a prerequisite to align the group replicate profiles. In this paper we present three pre-processing steps that seem particularly important for the further analysis of the MS data. These three steps are multiple alignment, baseline correction and normalization. The next sections provide

more details about these three steps. These three steps have been addressed independently by [14].

2. DATA

The data are borrowed from and fully described in [1]. To summarize: the serum samples were obtained from the Virginia Prostate Center Tissue and Body Fluid Bank. The patient cohort was procured from the department of Urology, Eastern Virginia Medical School and the healthy men cohort was obtained from free screening clinics open to the general public.

2.1. SELDI-TOF technology

The protein profilings were obtained with Ciphergen Biosystems, Inc. Surface-enhanced laser desorption / ionization – time of flight (SELDI-TOF) mass spectrometry. Retentate chromatography is performed on the Ciphergen protein chips by varying chromatographic properties using different affinity surface chemistries (eg anion exchange, cation exchange). For this study, only the IMAC-Cu (Immobilized Metal Affinity Capture) metal binding protein chips were selected for their good serum profiles in terms of number and resolution of proteins.

2.2. Quality control and data selection

Despite the use of the same IMAC-Cu metal binding protein chip for all the profiles, we saw quite a big difference among profiles even in the same treatment group. Therefore we removed the spectra that visually were too different from a target spectrum previously chosen, keeping only 5 spectra in each group. In a clinical study setting, stringent QC procedures have been proven necessary for the Seldi-TOF IMAC ProteinChip [4].

3. MULTIPLE PAIRWISE ALIGNMENT OF MS SPECTRA USING DYNAMIC PROGRAMMING

As mentioned earlier, drifts that do not reflect any real sample variations get introduced along the mass axis between replicates. These need to be corrected for. Our proposed dynamic programming (DP) multiple alignment method addresses this problem.

3.1. Multiple pairwise alignment following a guide tree

As opposed to [3] and [7] our method does not assume a fixed profile to which all the profiles must align. In our multiple pair wise alignment method all the spectra pairs are progressively warped together according to a guide tree (either Neighbor-Joining or UPGMA). This guide tree is built ahead of time from the pairwise distance matrix between all the spectra.

3.2. Smoothed pairwise DP warping

DP optimization was first introduced to align speech spectra [10] but it has been used often since then to align chromatogram traces for gene sequencing [12][13]. Roughly, the principle of DP is, given a point wise alignment score between two spectra, fill in the accumulated score path matrix (until the last point of the spectra). Keep track of the previous “best” alignment point in a pointer. The best alignment path minimizes the total score and can be found by backtracking the accumulated score matrix from the point with smallest accumulated score in the rightmost column and following the pointer recursively, all the way back to the first column. In an effort to increase the smoothness of the alignment curves during our DP warping, we also smoothed the score and alignment paths as suggested in [13]. For our limited data set, the mass alignment is a small improvement as the spectra were initially relatively well aligned.

4. BASELINE CORRECTION

The baseline adjustment method we use relies on a non-linear filter known as the “top-hat” operator in the mathematical morphology literature. The top-hat operator consists in subtracting from the original spectrum its morphological opening. By removing objects larger than the structuring element (whose size must be carefully chosen) typically top-hat operators remove the slow trends thus performing a contrast enhancement. For more information about mathematical morphology we refer to Jean Serra’s course publicly available at: <http://cmm.enscm.fr/~serra/cours/index>.

Such a morphological filter has already been used with success to correct for the background in microarray experiments [2][5]. Morphological filters, being very simple as they do not require any polynomial fitting, are very fast filters and are therefore especially appropriate for high-throughput data. Figures 3.1 and 3.2 illustrate our baseline correction method on the human prostate dataset. The same four spectra are plotted before baseline correction on Fig. 3.1 and after on Fig. 3.2. On these data sets our baseline correction method is very satisfactory.

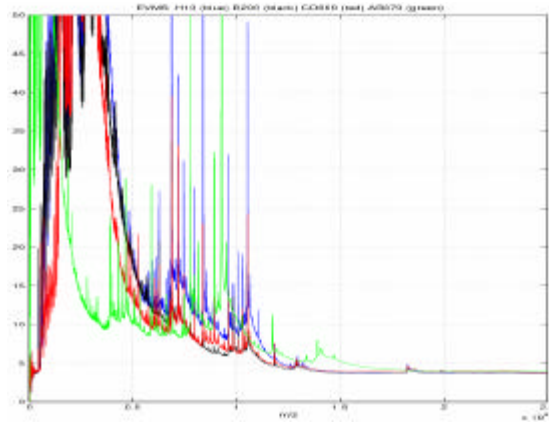


Fig. 4.1: Four MS spectra in need of baseline correction

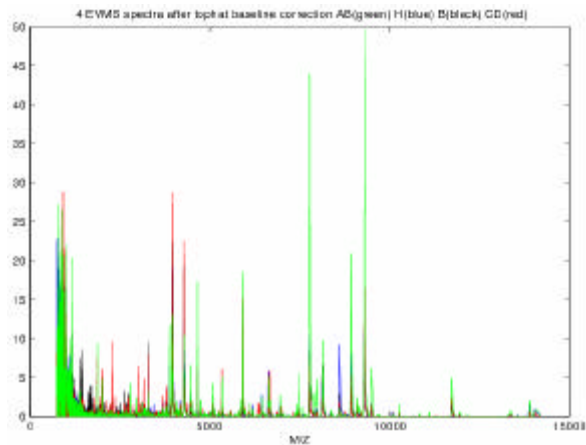


Fig 4.2: Same four MS spectra after the top-hat baseline correction

5. NORMALIZATION

The purpose of inter-spectrum normalization is to identify and remove sources of systematic variation between spectra due for instance to varying amounts of protein or degradation over time in the sample or even variation in the instrument detector sensitivity. The first normalization methods to consider are the global normalizations. Such global normalizations assume that the sample intensities are all related by a constant factor. A common choice for this rescaling coefficient is the spectrum median or the mean. The assumption behind the global normalization is that on average, the number of proteins that are over expressed is approximately equal to the number of proteins under expressed and the number of proteins whose expression level change is few relative to the total number of proteins. For mass spectrometric data, where each protein concentration is measured by the Area Under the Curve (AUC) of its peak, a natural choice for rescaling coefficient is the Average AUC also called the Total Ion Current (TIC). Therefore, we normalized each of our mass

spectra as in [6] by the dividing coefficient: AUC of spectrum / average AUC over all spectra.

5.1. Normalization study

The log scale is an appropriate scale to study the differences between mass spectra. We borrowed the following notations and plotting techniques from the microarray literature. As in [5], M denotes the point wise log ratio between two mass spectra and A their point wise log geometric mean. Therefore, MA plots are useful tools to study intensity-dependent variations between mass spectra. Similarly, M versus m/z plots allow us to study m/z dependent variation between spectra.

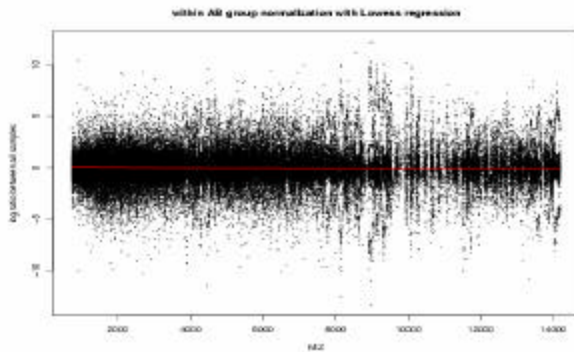


Fig 5.1.1: M versus m/z plot after normalization between all mass spectra pairs in AB cancer group with fitted Lowess regression curve in red.

Fig.5.1.1 illustrates the variations of normalized log-ratios within one of the cancer group as a function of m/z. Note that the log ratios are evenly distributed around zero across the m/z range. Such an even pattern was found for each single group and therefore we concluded that our normalized spectra have no location-dependent variation within groups and so do not need any location-dependent normalization.

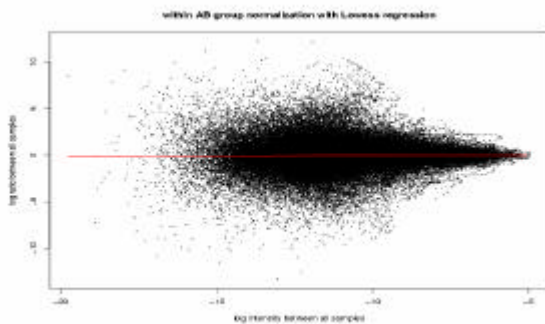


Fig 5.1.2: MA plot after normalization between all mass spectra pairs in AB cancer group with fitted Lowess regression curve in red.

Fig.5.1.2 illustrates the normalized log-ratios between mass spectra pairs in the same cancer group as above. Note again that the log ratios are evenly distributed around zero across the range of intensities. Such a pattern was found for each single group in our study and we therefore

conclude that our normalized spectra have no intensity-dependent variation within groups and so do not need any intensity-dependent normalization. At this point we would want to confirm these findings on more datasets but for the prostate cancer data set we used global normalization seems to be a satisfactory normalization. Ultimately a major goal is to find and classify biomarkers in serum [1][9]. We now address the question: does our normalization help in finding differentially expressed biomarkers between the different cancer groups?

5.2. Benefits from this global normalization

Our data set consists of 4 groups of 5 patients each. The spectra have been previously aligned and baseline corrected. We propose to compare the F ratios (between groups / within group sum of squares) across the m/z range for this data set, before and after normalization.

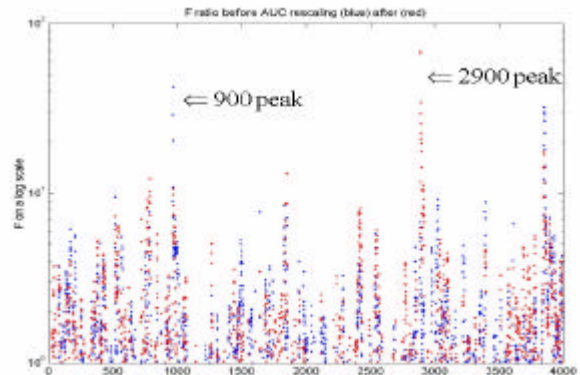


Fig 5.2.1: F ratios versus masses for our prostate data set. Blue: before normalization, Red: after normalization.

Peaks with high F ratio are of potential interest. As pointed out on Fig. 5.2.1, two peaks in the F occur around masses 900 and 2900 before and after normalization. While mass 900 has its F value decreased after normalization, mass 2900 sees its F value increase. Is mass 2900 a true difference between the 4 groups that is amplified after normalization?. Also is mass 900 a false discrepancy between the groups that is attenuated after normalization? In an attempt to answer this question we overlap the 5 patients' spectra for each group before (blue) and after (red) normalization in the two mass regions of interest, see Fig 5.2.2.

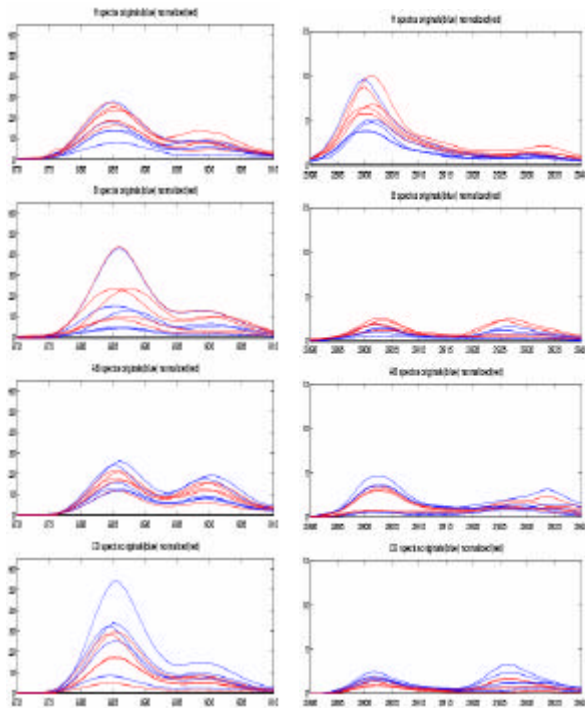


Fig 5.2.2: For each of the 4 groups of patients, Left: zoom around mass 900, Right: zoom around mass 2900.

At mass 2900, group H clearly rose above the other groups after normalization making it a possible biomarker candidate. On the other hand, at mass 900, normalization seems to have uniformized the groups making us reject mass 900 as a possible biomarker. From this small study, we believe that the rescaling is a useful preprocessing step before classification but we would now need to check that these results are biologically meaningful following the method described in [8].

6. CONCLUSION

With our limited data set we observe that the replicate alignments bring some improvement even when the replicates are initially well aligned. The baseline correction seems essential. Finally, we show that normalization can possibly help discriminating between groups.

7. REFERENCES

[1] Bao-Ling Adam, Yinsheng Qu et al., “Serum Protein Fingerprinting Coupled with Pattern-matching Algorithm Distinguishes Prostate Cancer From Benign Prostate Hyperplasia and Healthy Men”, *Cancer Research* 62 pp. 3609-3614, July. 2002.

[2] Jesus Angulo and Jean Serra, “Automatic analysis of DNA microarray images using mathematical morphology”, *Bioinformatics* 19(5) pp. 553-562, 2003.

[3] Dan Bylund, Rolf Danielsson et al. “Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modeling of liquid chromatography-mass spectrometry data”, *Journal of Chromatography* 961, pp. 237-244, Jul 2001.

[4] Kevin R. Coombes, Herbert A. Fritsche et al. “Quality Control and Peak Finding for Proteomics Data Collected from Nipple Aspirate Fluid by Surface-Enhanced Laser Desorption and Ionization”, *Clinical Chemistry* 49:10, pp. 1615-1623, 2003.

[5] Sandrine Dudoit, Yee Hwa Yang et al. , “Statistical Methods for identifying differentially expressed genes in replicated cDNA microarray experiments”, *Statistica Sinica* 12 (1) pp. 111-139, January 2002.

[6] Eric T. Fung and Cynthia Enderwick, “ProteinChip Clinical Proteomics: Computational Challenges and Solutions”, *Computational Proteomics Supplement* 32 pp. S34-S41, March 2002.

[7] Niels-Peter Vest Nielsen, Jens Michael Cartensen et al., “Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimized warping”, *Journal of Chromatography A*, 805 pp. 17-35, 1998.

[8] M. S. Pepe, R. Etzioni et al., “Phases of biomarker development for early detection of cancer”, *Journal of the National Cancer Institute* 93(14) pp. 1954-1061, July. 2001.

[9] Emmanuel F. Petricoin III, Ali M. Ardekani et al., “Use of proteomic patterns in serum to identify ovarian cancer”, *The Lancet* 359 pp. 572-577, Feb. 2002.

[10] L.R. Rabiner and B.H Juang, “Fundamentals of Speech Recognition”. Prentice Hall, 1993.

[11] Gary Siuzdak, “Mass Spectrometry for biotechnology”. Academic Press, March. 1996.

[12] J.D. Thompson, D.G. Higgins and T.J. Gibson, “*CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*”, *Nucleic Acids Res.*,22(22),pp 4673-80, Nov. 1994.

[13] Pratyaksha Jagad Wirapati, “An automated Allele-calling System for High-throughput Microsatellite Genotyping”, *Ph.D. dissertation Department of Medical Biology University of Melbourne*, Jan. 2003.

[14] Hongyu Zhao, “Statistical issues in using mass spectra for disease classification”, IPAM workshop: High throughput technologies and methods of analysis, March 2004.