Probe design optimization of HLA microarray

Data cleaning of probe signals from cDNA tiling microarray: outlier detection, noise reduction, and identification of uninformative probes in HLA typing application

Esma Dilek
Computer Science Department
Metropolitan College
Boston University
Boston, MA, USA
esmadilek@gmail.com

Guang Lan Zhang
Cancer Vaccine Center
Dana-Farber Cancer Institute
Boston, MA, USA
guanglan_zhang@dfci.harvard.edu

Jae Young Lee
Computer Science Department
Metropolitan College
Boston University
Boston, MA, USA
jaeylee@bu.edu

Tanya Zlateva
Computer Science Department
Metropolitan College
Boston University
Boston, MA, USA
zlateva@bu.edu

Lou Chitkushev
Computer Science Department
Metropolitan College
Boston University
Boston, MA, USA
ltc@bu.edu

Vladimir Brusic
Cancer Vaccine Center
Dana-Farber Cancer Institute,
Boston, MA, USA
vladimir_brusic@dfci.harvard.edu

Abstract

Abstract— **Custom made tiling cDNA microarrays have been developed for high resolution HLA genotyping. The array comprises tens of thousands of sequence-specific oligonucleotide probes (SSOP) that hybridize with HLA sequences. Sophisticated methods for analyzing the multidimensional complex data are needed due to large amounts of data generated from the microarray experiments. Moreover, the inconsistencies, noise, and outliers in the probe signals add additional complexities. We proposed data cleaning methods for the identification of uninformative and misinformative probes to improve the performance of HLA typing process.**

Keywords-HLA typing; cDNA micorarray; outlier detection; noise reduction

I. INTRODUCTION

Human leukocyte antigens (HLA) proteins are present on the surface of every nucleated cell in human body. HLA molecules bind short peptides and display them for recognition by T cells of the immune system. Peptides presented by HLA molecules originate from degradation of intracellular (presented by HLA class I) or extracellular (presented by HLA class II) proteins [1]. The ability of the immune system to respond to a particular antigen varies of fine differences that are often single nucleotide variations between individuals according to their specific pattern of HLA genes. Each human individual expresses up to six HLA class I molecules and more than ten HLA class II molecules [2]. HLA genes show extensive variability. More than 5,000 variants of HLA class I and more than 1500 variants of HLA class II molecules have been characterized and named to date (IMGT/HLA database, September 2011). Precise identification of HLA profiles (HLA typing) is important in transplantation, assessment of susceptibility to certain disease, and prediction of immune responses to infection and vaccination [3]. An average individual expresses two HLA-A, two HLA-B, and two HLA-C molecules. There are a total of 1,698 reported variants of HLA-A, 2271 variants of HLA-B, and 1213 variants of HLA-C, making the theoretical number of HLA class I profiles as large as [4]. Identification of HLA profiles from samples requires analysis of large number of highly similar sequences and interpretation of fine differences that are often single nucleotide variations.

Custom made tiling cDNA microarrays have been developed for high resolution HLA genotyping. This array comprises tens of thousands of sequence-specific oligonucleotide probes (SSOP) that hybridize with HLA sequences. Fine differences are encoded by probes that have a small number of differences. Because of large amounts of data generated from the microarray experiments, sophisticated methods for analyzing the multidimensional complex data are needed [5]. Moreover, the inconsistencies, noise, and outliers in the probe signals add additional complexities. The development of the information

system for removal of inconsistencies and noise in probe signals is essential for the downstream probe signal analysis. In this study, we focused on the identification of uninformative and misinformative probes using computational methods to improve the performance of HLA typing process.

II. MATERIAL AND METHOD

The HLA typing experiments were performed using Agilent custom designed 4×44K slide design. Each cDNA microarray covered 2927 HLA variants. From the alignments of HLA sequences, we identified 850,309 probe-variant pairs that were used for data analysis. HLA profiles from four samples were also provided along with as observed experimental results of probe signals. Predicted HLA variants of individuals were compared with the actual HLA profiles to measure how many HLA variants were correctly predicted after applying the data cleaning methods. A relational database was built using Microsoft SQL Server 2005 to store the research dataset and results of data cleaning methods. This system allows us to retrieve and manipulate data in a highly efficient manner.

We first preprocessed the input dataset and stored in a relational database so that we could perform multiple analyses efficiently. Positions of each probe were calculated using multiple sequence alignment files and all provided research dataset were stored in database tables to enable systematic analyses and categorization of probes. Data preprocessing allowed us to query probe signal distributions of HLA variants and to visualize them easily by the produced charts. It also facilitated identification and handling the outliers among probe data.

After preprocessing, we performed a number of visual analyses to find out similarities and variations between probe signal patterns belonging to different individuals. We produced several charts and compared them to recognize both common and distinct patterns of probe signals. Comparison of probe signal distribution charts of HLA variants belonging to different individuals showed that some probes produced high signals across all arrays (*i.e.* for all individual samples). Since experiments were carried out with samples obtained from four different samples, it was not expected that a probe has global high signals across all arrays unless it was experimentally proved that it existed in all individuals. We observed that among those probes which have high signals across all arrays, less than 3% of them are present in all individuals, and about 25% of them are not present in any of the individuals, and for about 25% of them, we did not have information whether they are present or not in the sampled individuals. Based on statistical analysis and visualization of results, we categorized probes which produce global high signals across all arrays as "uninformative probes". Since these probes are not selective, they are not beneficial for identifying HLA variants within these individuals. They were considered to be the global false positives within the SSOP data. The selection algorithm is as following. Let threshold, t, be the threshold used in the global high signals elimination approach to detect uninformative probes. If a probe's average signal across all arrays was greater than or equal to the threshold, we categorized the probe as "uninformative probe". If a probe's average signal across all arrays was less than the threshold, we categorized that probe categorized as "informative probe". In the study, multiple values between 1,000 and 20,000 have been explored to be the predefined threshold. After identification of uninformative probes, system performance was measured by filtering out the uninformative probes and using the remaining informative probes as input to the objective function.

We identified outliers in the probe dataset by employing K-Means clustering-based approach. Since outlier probes were considered to have distinctly different signals to their neighboring probes, the expected outcome of clustering-based analysis was that they were assigned to clusters with small number of probes. If a resulting cluster had sufficient number of probes assigned to it, we considered the probes in that cluster to be informative. If a resulting cluster was of small size, (i.e. had not enough probes assigned to it), then we designated a probe in that cluster a candidate outlier. Then, we further analyzed these candidate probes by adding them to other clusters and examined the change of standard deviations of the clusters to which they were added. If it produced a statistically significant increase when added to other clusters, we considered that probe to be an outlier; otherwise we considered that probe to be informative.

To categorize a probe by K-Means clustering approach, we employed majority voting and categorized the probe based on its votes. If its votes as an outlier probe exceeded its votes as an informative probe, then we categorized that probe as "misinformative probe" and cleaned from probe dataset for HLA prediction; otherwise we categorized that probe as "informative probe" and used for HLA variant prediction.

As an alternative to clustering-based approach to detect outliers (i.e. misinformative probes), we studied statistical-based outlier detection approach and applied "3σ edit rule" to probe dataset. We changed the original "3σ" standard deviation threshold which was used to determine outliers in "3σ edit rule", and the number of data points considered during detection of

outliers in a systematic manner to find the best set of informative probes for HLA prediction.

Let S_{all} be the set of all available probes, and let $S_{informative}$ be the set of identified informative probes, and $S_{uninformative}$ be the set of identified uninformative probes, and $S_{misinformative}$ be the set of identified misinformative probes used in this study. The list of predicted HLA variants for each sample was produced using S_{all} as input to the objective function as an initial step. After that step, $S_{uninformative}$ and $S_{misinformative}$ were excluded from consideration and the performance of the developed data cleaning algorithms was measured by using $S_{informative}$ as input to the objective function. This procedure was repeated by systematically changing the predefined threshold and parameters used in algorithms for finding out the optimal set of probes which were produced by objective function and useful for HLA variant prediction.

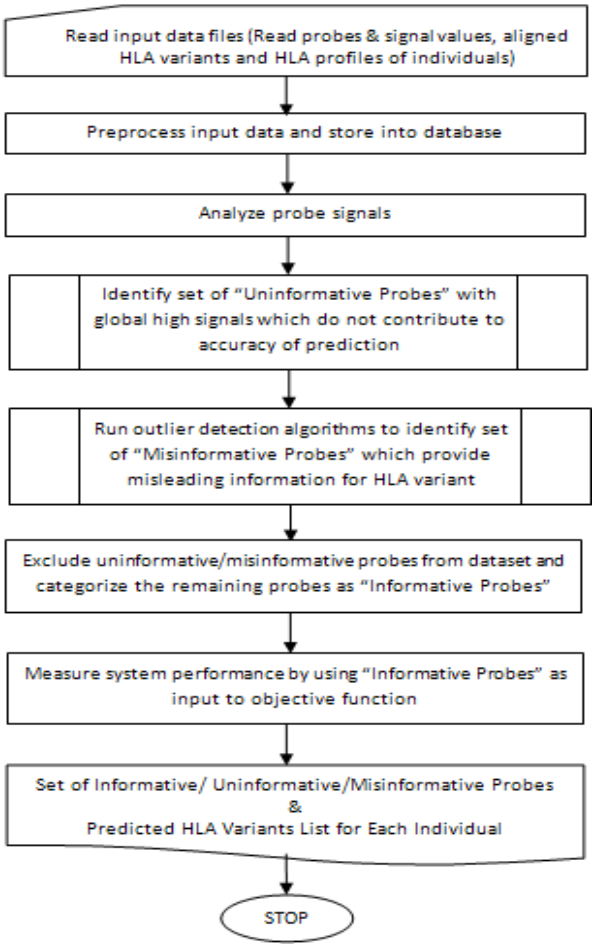

Figure 1. Schematic diagram of the overall method.

Figure 1 illustrates the schematic diagram of the overall system which we designed to clean SSOP data with basic steps.

III. RESULTS AND CONCLUSION

We proposed data cleaning methods for a specific SSOP application to do accurate HLA typing and focused on detecting uninformative and misinformative probes. A significant improvement in the number of falsely identified HLA was achieved when array-specific thresholds and parameters were used for data cleaning. Results show that the simulations which yielded the highest sensitivity and specificity rates were not the same for all arrays although for some of the samples it was consistent across arrays (Simulation results which belong to the same sample are shown with the same color). This indicates that most of noise has biological, sample-specific origin and not the setting or conditions of experiments. We can conclude that the vast majority of noise originates from biochemical properties of individual probes and combinatorial properties of HLA variants from each sample. Simulation results showed that each array needs to be optimized locally and optimal filtering parameters are sample-dependent. Different arrays that target samples

Our major finding was that using local filtering thresholds and parameters is critical to identify informative probes. The outputs of our analyses can be used for future design of HLA microarray so that unnecessary or detrimental probes will not be involved in subsequent analysis.

REFERENCES

[1] M. Colonna, M. Bresnahan, S. Bahram, J.L. Strominger, T. Spies, "Allelic variants of the human putative peptide transporter involved in antigen processing", *Proc. Natl. Acad. Sci. USA*, 1992; 89, pp. 3932-3936.

[2] R. Coico, *Immunology: a short course*: Wiley-Liss; 2003.

[3] C. Feng, C. Putonti, M. Zhang, R. Eggers, R. Mitra, M. Hogan, K. Jayaraman, Y. Fofanov, "Ultraspecific probes for high throughput HLA typing", *BMC Genomics*, 2009, 10:85.

[4] Anthony Nolan Research Institute http://hla.alleles.org/nomenclature/stats.html.

[5] S. Durinck. "Pre-processing of microarray data and analysis of differential expression", *Methods in Molecular Biology*, 2008, 452, pp. 89-110.