



# Effective Information Integration from Disparate Microarray Datasets

Duygu Ucar<sup>1</sup>, Sarah Javaid<sup>2</sup>, and Srinivasan Parthasarathy<sup>1,2</sup>

<sup>1</sup> Department of Computer Science,

<sup>2</sup> Biophysics Graduate Program,

The Ohio State University, Columbus, OH, USA

## Motivation:

To discover genes that might play a role in the existence of lung cancer, an analysis was done on two disparate microarray lung cancer datasets. The datasets were integrated based on common probesets, and co-clustered both individually and together. MedlineR and Common Subsets were used to evaluate the final clusters.

## Approach:

Although most genes in human DNA have been sequenced, the functional relationships between the genes are not fully understood. The problem with obtaining information about which genes are linked to a specific type of cancer and how these genes are interrelated is an important topic. With the development of DNA microarray and other biological devices, expression levels of thousands of genes can be measured at the same time. Due to such technological advances, data is produced at a very fast rate. Analyzing and integrating this data efficiently is thus paramount. However, the extraction of biologically relevant data is a challenging task and the small number of patients is a limiting factor. Information was integrated from studies performed at different institutions using various microarray technologies. All four datasets from CAMDA 2003 were initially considered but Harvard and Michigan datasets were chosen. Thus, 289 patients were considered, 203 from Harvard and 86 from Michigan. This required us to integrate the Harvard and Michigan datasets, which used two different Affymetrix oligonucleotide microarrays (HU6800 and HGU95a Chips). The bioconductor packages implemented in R was used to obtain gene expression values from the CEL files. The datasets were RMA normalized and inactive genes with a small standard deviation were eliminated. The comparison spreadsheet obtained from the Affymetrix website was used to integrate the resulting datasets resulting in 1,709 common probesets for 289 patients. The integrated dataset was re-normalized with RMA in order to eliminate any experimental bias caused by different environments and technologies.

Quite a large number and type of clustering methods have been applied on microarray datasets such as k-means clustering, hierarchical clustering and self-organizing maps (SOM). However, most of these algorithms focus on clustering along one dimension. Typically, the microarray data is arranged as a matrix. Thus, one would like a system that can simultaneously cluster both dimensions of a matrix by exploiting both the rows and the columns. Co-clustering differs from clustering along one dimension in that at all stages, row clusters incorporates column cluster information and vice versa. We applied the minimum sum squared residue co-clustering to simultaneously cluster rows (probesets) and columns (patients) at the same time (Cho et al., 2004). We picked 25 as the number of probeset cluster because it resulted in reasonable sized clusters. Having too many probesets inside a cluster leads to a large number of accidental relations and decreases the reliability and interpretability of the final probeset clusters. Patient clusters were chosen based on the different types of diseases on each dataset (5 and 2 were selected as column cluster size for Harvard and Michigan respectively). The resultant final probeset clusters obtained from the co-clustering algorithm formed the basis for our analysis.

Clusters were obtained on the individual datasets as well as the integrated data. The probesets were matched with their respective genes and a set of genes that co-occurred in all three datasets (Harvard, Michigan, and Integrated) were extracted. We introduced the terminology common subsets to denote these common sets of genes that occurred in each dataset. We believe that common subsets have strong correlations amongst each other because they ended up in the same cluster for each data. This method helped to reduce the number of genes to be considered and eliminated the high dimensionality problem in microarrays. Validation of common subsets was done by querying Medline abstracts. Among all the Medline querying tools, we employed MedlineR, an open source library, which was implemented in R (Lin et al., 2004). MedlineR produced a co-occurrence

matrix which corresponds to the pair-wise co-occurrences of querying terms (genes) in at least one Medline abstract. If two or more genes occur in an abstract together, they are assumed to have a relationship. Also, a co-occurrence graph was obtained from the matrix and constructed in Pajek (visualization software) to examine the graph interactively (Batagelj and Mrvar, 2004). In each graph, each vertex represents a gene whereas an edge between two vertices corresponds to the co-occurrence between the two genes in at least one Medline abstract.

### **Results:**

Preliminary results show that known and unknown associations between genes might be discovered by using this technique. We evaluated some of the common subsets that were extracted and some associations have been discovered amongst these genes. The associations between PCNA with FEN1, RPA1, CKS2, TOPBP1, and YES1 (in one common subset) are all due to genomic replication and stability. Mutation of one or more of these genes might play a role in cancer development. For example, PC-SPES is an herbal mixture which has clinical efficacy against prostate cancer. Analysis has shown that PC-SPES causes decreased expression of cyclin B, Nedd8, cdc2, skp1, PCNA, MAD2L1, cyclin H, CKS2, E2F, Rbx1, MCM2, MCM5, Mpp2, Cullin-Cul4A, Cks1p9 and McM7, which are involved in

cell cycle progression. This shows that PCNA and CKS2 are by some means related. Top2A, SMC1L1, RRM1, MCM2, KIAA0042, TYMS, E2F3, MCM7, and ADPRT are elements of another common subset that was obtained. Medline abstracts show that TOP2A and E2F3 are proven to be upregulated together in an experiment conducted on kidney cancer cases. In another study, TOP2A and TYMS are found to be amongst the 12 drug response genes for 8 anti-cancer drugs. Another abstract concluded TYMS and RRM1 play important roles in the G1/S transition, and down-regulation of human hepatocarcinoma Hep3B cells. We also inferred that inhibition of thymidylate synthase (TYMS) is an important target for cancer chemotherapy and MCM7 is essential for the initiation of eukaryotic genome replication. We discovered that PARP (ADPRT) is activated by DNA strand breaks and is implicated in DNA repair, recombination, DNA replication, and transcription. We also analyzed that SMC1L1 interacts with BRCA1, which is known to be a cancer causing gene.

We proposed a method to extract information from integrated microarray datasets. Although dense co-occurrence graphs were not obtained, we argue that using the proposed method is a valid approach to discover prognostic genes.