Finding cancer genes through meta-analysis of microarray experiments: Rank aggregation via the cross entropy algorithm

Vasyl Pihur, Susmita Datta[1], Somnath Datta

Department of Bioinformatics and Biostatistics, University of Louisville, KY 40292, USA

1 Introduction and a summary

Meta-analysis of microarray data coming from a number of microarray experiments can be attempted with two systematically different approaches. Though it is obvious that some sort of aggregation of the results is necessary to accomplish this, enough flexibility is left regarding what is aggregated and at what stage. In the first approach, different microarray experiments are put together forming a single dataset that can be analyzed as a separate entity without considering the origin of each sample. Performing a clustering analysis on a set of Probe IDs (genes) based on their expression profiles created as a result of such aggregation of microarray experiments is a perfect example of the first approach. In the second approach, however, instead of aggregating expression values (combining microarray samples directly), each individual microarray experiment is analyzed first and then the statistical results from all experiments are aggregated to produce the final meta-analysis results. It is the latter that is accomplished in this research where we combine the statistical evidence of differential expressions across various tissue types (e.g., normal and various gradations of cancer) from twenty different cancer experiments. An advantage of this approach is that it is applicable even if the individual experiments were run using different microarray platforms as long as the corresponding genes could be identified. We produce a list of fifty genes that are judged to be the most differentially expressed overall in these experiments when the rankings in terms of p-values are aggregated. The rank aggregation is formulated as a well defined minimization problem in terms of decision theory which is then solved by the cross entropy algorithm (originally due to Rubinstein (1999)). The remarkable property of this algorithm is that it is capable of performing the necessary stochastic search in a rather large space of possible lists (its size was of the order of 10^{148} for our problem!). The resulting top fifty list contains thirty-six genes that have been implicated in cancer (often with more than one cancer type) in the literature. The remaining genes are novel in terms of their connection to cancer. The current analysis suggests that perhaps they should be investigated further for their regulatory roles related to cancer activities.

A pervious attempt (DeConde *et al.*, 2006) of rank aggregation of genes from multiple microarray experiments (which we became aware of very recently) combined a relatively small number of ranked lists (five lists, all from prostate cancer, to be precise) using meta-search algorithms. The usefulness of cross entropy methods for the rank aggregation problem and other biological applications was recognized by Lin at el. (2006) and Pihur et al. (2007).

2 Meta-dataset

In this paper, we analyze part of the challenge dataset (META-analysis dataset) for CAMDA 2007. This contest meta-analysis microarray dataset consists of 5897 arrays collected from approximately 250 individual microarray experiments which study the whole range of different conditions in humans. All

[1] susmita.datta@louisville.edu

of the individual microarray studies were hybridized with the Affymetrix GeneChip Human Genome HG-U133A and record expression levels for 22,283 Probe IDs.

To make our final results meaningful and comprehensive at the same time, we decided to focus on microarray experiments that study different types of cancer in humans. The goal of our meta-analysis is to identify genetic factors which are common across different types and stages of cancer. For that purpose, we have selected 20 different cancer-related microarray experiments which are included in the contest meta-dataset and have explicit cell type groupings necessary for detecting differentially expressed genes. Here, we list the selected experiment IDs along with the number of samples in each experiment in the parenthesis: E-MEXP-72 (20), E-MEXP-83 (22), E-MEXP-76 (17), E-MEXP-97 (24), E-MEXP-121 (30), E-MEXP-149 (20), E-MEXP-231 (58), E-MEXP-353 (96), E-TABM-26 (57), E-MEXP-669 (24), GSE4475 (221), GSE1456 (159), GSE5090 (17), GSE1420 (24), GSE1577 (29), GSE1729 (43), GSE2485 (18), GSE2603 (21), GSE3585 (12), GSE4127 (29). The total number of selected arrays is 941 (about 1/6 of the overall number of samples in the meta-dataset). One can refer to ArrayExpress database which provides public access to the microarray data from these experiments (http://www.ebi.ac.uk/arrayexpress/).

3 Methodology

The proposed meta-analysis approach to microarray data is a two-step procedure:

1. **Individual Analysis.** By analyzing each microarray dataset individually, a set of "interesting" genes (top-50 Probe IDs) that exhibit the largest differences in terms of expression values between the groups is obtained for each dataset.

2. **Rank Aggregation.** Aggregation of the individual lists from Step 1 based on the rankings of genes within each list is performed to produce a "super"-list of 50 Probe IDs which would reflect the overall importance of genes as judged by the collective evidence of all experiments.

In the first step, one-way ANOVA analysis is performed on each Probe ID for each dataset (20 in our case). The usual F-test statistic and the corresponding p-value is computed for each Probe ID. The smaller the p-value, the stronger the evidence for the involvement of the corresponding Probe ID in cancer-related processes. If we rank Probe IDs according to the p-values assigned by ANOVA from the smallest to the largest, the top most Probe IDs are of primary interest to biologists as revealed through that microarray experiment. Thus, for example, a top-50 Probe ID list can be obtained for each microarray experiment. Since the experiments are related, combining the top-50 lists is the next logical step to do. The rank aggregation method provided in Pihur *et al.* (2007) can be easily adapted to achieve this. The method is based on Cross-entropy Monte Carlo algorithm proposed by Rubinstein for solving large combinatorial optimization problems (Rubinstein, 1999). The objective function we want to minimize in our aggregation problem is relatively simple. Find the aggregated ordered list

$$\delta^* = \arg\min_{\delta} \sum_{M} d(\delta, L_M),$$

where M indices the microarray experiments, L_M are the ordered lists to be combined, δ is any ordered list of size $k = |L_M|$, and d is a distance function which, in our case, is the weighted Spearman footrule distance to be defined next.

Let $p(1, M), \ldots, p(k, M)$ be the p-values corresponding to the top-k Probe IDs and $r^M(G)$ be the rank of Probe ID G under M (1 means "best") if G is within top k, and be equal to $k+1$, otherwise.

The weighted Spearman footrule distance then can be defined as

$$d(\delta, L_M) = \sum_{t \in L_M \cup \delta} |p(r^{\delta}(t), M) - p(r^{L_M}(t), M)| \times |r^{\delta}(t) - r^{L_M}(t)|.$$

We omit the details of the CE algorithm from this extended abstract. It can be found in earlier papers (e.g., De Boer *et al.* (2005), Lin *et al.* (2006), Pihur *et al.* (2007)). In the present context, the CE algorithm searches for δ^*, the list which is the "closest" to all L_M lists simultaneously.

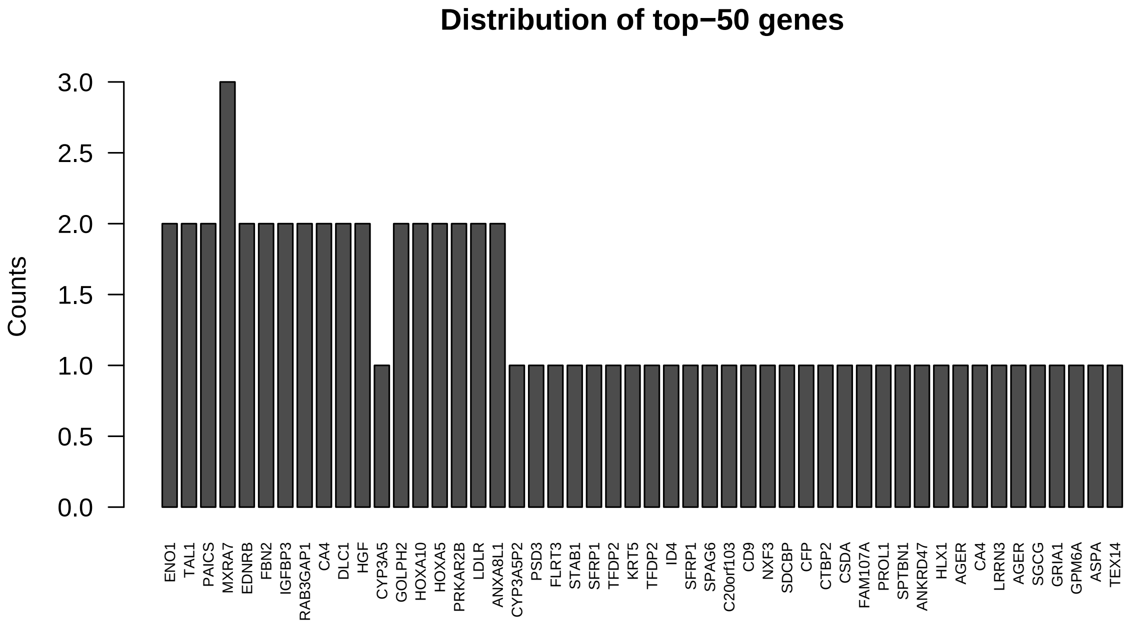

Figure 1: The distribution of the top-50 genes from our overall list. The height of each bar represents the number of times (out of 20) each gene appears in individual top-50 lists.

4 Results

For each of the 20 microarray datasets, we obtain a top-50 list of Probe IDs ranked according to their p-values from an ANOVA analysis. The lists along with the corresponding p-values, which play the role of weights, are then used to produce a single combined top-50 list via the weighted rank aggregation approach discussed in the previous section. We obtained n = 966 unique Probe IDs in the twenty top-50 lists and therefore the solution space of ordered $k = 50$ lists has an astonishing 4.880507×10^{148} many elements. Applying the rank aggregation procedure to the twenty top-50 lists produces the following overall gene list: *ENO1, TAL1, PAICS, MXRA7, EDNRB, FBN2, IGFBP3, RAB3GAP1, CA4, DLC1, HGF, CYP3A5, GOLPH2, HOXA10, HOXA5, PRKAR2B, LDLR, ANXA8L1, CYP3A5P2, PSD3, FLRT3, STAB1, SFRP1, TFDP2, KRT5, TFDP2, ID4, SFRP1, SPAG6, C20ORF103, CD9, NXF3, SDCBP, CFP, CTBP2, CSDA, FAM107A, PROL1, SPTBN1, ANKRD47, HLX1, AGER, CA4, LRRN3, AGER, SGCG, GRIA1, GPM6A, ASPA, TEX14.* The algorithm converged in 47 iterations with the objective function value of 36952.215. In Figure 1, a barplot shows the total number of times each gene in the aggregated list appears within the 20 individual lists. All but one of the first 18 genes were in at least two top-50 lists. The maximum number of lists that any gene appeared in was 3 (only one gene *MXRA7*). Despite the fact that the microarray experiments were related, the majority of genes were in only one top-50 list.

In Table 1 we show the first 8 Probe IDs from the combined list along with their GO biological functions mined through Osprey (Breitkreutz *et al.*, 2003), KEGG pathways mined through DAVID (Dennis *et al.*, 2003) and oncological implications with relevant PubMed IDs for the most part identified through Gene Cards available at http://www.genecards.org. The full table is available at http://www.somnathdatta.org/Supp/Camda/supp.htm in Excel format. Thirty-six of the fifty genes in our aggregated list have been previously implicated in different cancers, many of them, as we would expect, in multiple cancers. For the remaining fourteen genes we were unable to find any previously published results that would indicate their involvement in cancer development. For most of them, not much biological information in the form of GO annotations and pathways is available. Additional discussions of this table will be provided in the paper.

5 Discussion

The weighted rank aggregation method based on the Cross-entropy Monte Carlo algorithm proves to be a useful tool in carrying out meta-analysis of microarray experiments. Top-k lists of genes, the usual results of microarray analyses, can be successfully aggregated to form a single list of genes which is based on multiple experiments.

We ran the aggregation method three times with different starting seeds for the Monte carlo sampler and each time obtained a slightly different aggregated list. Looking at the values of the objective function for each resulting list, we noted that the difference amongst them was less than 0.033%. In the result section we reported the list corresponding to the lowest value of the objective function out of the three. The lists were very similar in terms of which Probe IDs were included (at least forty Probe IDs from the reported list were included in the other two lists) but differed in the actual ordering, especially, in the tail ends. It is perhaps not surprising that multiple runs of the stochastic Cross-entropy algorithm do not produce identical ordered lists, especially given that the size of the search space is extremely large.

References

Breitkreutz, B. J., Stark, C., and Tyers, M. (2003). Osprey: a network visualization system. *Genome Biol*, **4**(3), R22.

De Boer, P., Kroese, D., Mannor, S., and Rubinstein, R. (2005). A tutorial on the cross-entropy method. *Ann. Oper. Res.*, **134**, 19–67.

DeConde, R., Hawley, S., Falcon, S., Clegg, N., Knudsen, B., and Etzioni, R. (2006). Combining results of microarray experiments: a rank aggregation approach. *Stat Appl Genet Mol Biol*, **5**(1), Article15.

Dennis, G., J., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. (2003). David: Database for annotation, visualization, and integrated discovery. *Genome Biol*, **4**(5), P3.

Lin, S., Ding, J., and Zhou, J. (2006). Rank aggregation of putative microrna targets with cross-entropy monte carlo methods. (Preprint, presented at the IBC 2006 conference, Montral).

Pihur, V., Datta, S., and Datta, S. (2007). Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach. *Bioinformatics*, **23**(13), 1607–1615.

Rubinstein, R. (1999). The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, **1**, 127–190.

Table 1: Top-8 Probe IDs from the combined list. All but the last one have been implicated in playing a role in different cancers in the past. PubMed IDs are given for further references.

ProbeID	Symbol	GO Process	Kegg Pathways	Oncology	PubMed ID	Description
217294_s_at	ENO1	negative regulation of cell growth, glycolysis, regulation of transcription		breast carcinoma	7641187	negative regulatory function by down-regulating c-myc expression
					9074493	transcriptional repressor activity on c-myc promoter
				cervical carcinoma	7641187	
				lung carcinoma	10853020	
				carcinoma	11973636	selectively represses Bcl-xL expression in MCF-7 cells and induces mitochondrial involvement in the apoptotic process
206283_s_at	TAL1	regulation of transcription, cell proliferation, cell differentiation	HSA04310: wnt signaling pathway	t-all	9695959	disrupted by translocation or deletion (tal(d)) in up to 30% of T-cell acute lymphoblastic leukaemia (T-ALL)
					8208530	
				leukemia	2040693	tal-1 rearrangements
201014_s_at	PAICS	purine base biosynthetic process, 'de novo' IMP biosynthetic process		cancer	17224163	this study provides essential structural information for designing PAICS-specific inhibitors for use in cancer chemotherapy
				lung carcinoma	15246564	
212509_s_at	MXRA7	integral to membrane		MRD	16627760	tissue matrix remodeling-like gene
206701_x_at	EDNRB	G-protein signaling, coupled to IP3 second messenger, signal transduction, peripheral nervous system development	HSA04020: calcium signaling pathway, HSA04080: neuroactive ligand-receptor interaction	bladder cancer	15569975	
203184_at	FBN2	anatomic structure morphogenesis	pancreatic cancer		15951052	loss of FBN2 expression due to promoter methylation was recently identified in pancreatic cancer
212143_s_at	IGFBP3	positive regulation of myoblast differentiation, regualtion of cell growth, positive regulation of apoptosis		breast cancer	8609661	involved in the regulation of breast cancer cell growth
					10067859	
212932_at	RAB3GAP1	regulating GTPase activity				