

Finding common genes in multiple cancer types through meta-analysis of microarray experiments: A rank aggregation approach

V. Pihur, Somnath Datta, Susmita Datta*

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

ARTICLE INFO

Article history:

Received 4 February 2008

Accepted 8 May 2008

Available online 20 June 2008

Keywords:

Microarray

Cancer

Meta-analysis

Rank aggregation

Cross-entropy

ABSTRACT

Discovering genes involved in multiple types of cancers is of significant therapeutic importance. We show that collective evidence for such genes can be obtained via a form of meta-analysis that aggregates the results (rankings and p values) from various cancer-specific microarray experiments. This method is illustrated by a combined analysis of 20 microarray experiments. In the aggregated list of top-50 genes, 36 of them have been implicated in cancer (often multiple cancers) genesis in past studies, which also suggests that this list may contain some novel cancer genes that may deserve further scrutiny in the future.

© 2008 Elsevier Inc. All rights reserved.

Learning from small-scale experiments, biologists have known for a long time of certain genes that are risk factors for multiple cancers. Identification of such genes is most important because targeting them could lead to prevention of many types of cancer. These genes include both oncogenes, which are amplified in cancers and activate the growth of tumors across different organs, and tumor suppressor genes, which have the opposite effect and, if active, prevent multiple types of tumor growth and development. For example, the oncogene *PIK3CA* belonging to the PI3K pathway has been implicated in numerous cancers [1,2], in particular ovarian, colorectal, and endometrial cancers [3]. Similarly, *p53*, which is a tumor suppressor gene whose activity prevents formation of tumors, clearly affects multiple cancers because of its fundamental role in tumor suppression [4,5]. Recently, Bagchi et al. [6] discovered a novel tumor suppressor gene, *CHD5*, and suggested that it could be a target for multiple cancer prevention.

The majority of cancer microarray experiments, however, are designed around a specific cancer type. By comparing gene expression levels between tumor and healthy samples, researchers can identify significantly differentially expressed genes that are usually system-specific and may not necessarily be implicated in other types of cancer. However, to put some of these genes into a greater perspective and construct an overall picture of the common genetic factors across multiple types of cancer, much more data needed to be collected. Thus it is perhaps not surprising that the surge in the numbers of microarray experiments in the last several years presented an opportunity for researchers to answer the question of the involvement of common

genes in multiple cancers in a more systematic way by means of microarray data meta-analysis [7–9].

Meta-analysis of microarray data coming from a number of microarray experiments can be attempted with two systematically different approaches. Although 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 to form a single dataset that can be analyzed as a separate entity without considering the origin of each sample (see, e.g., [10]). Performing a cluster 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 [11].

In this work we follow the second approach in which we combine the statistical evidence of differential expressions across various tissue types (e.g., normal and various gradations of cancer) from 20 different cancer experiments. Our aggregation uses both the rank orders of each gene (in terms of its statistical significance) and the (appropriately scaled) magnitude of the actual p values in each experiment. An advantage of our method is that it is applicable even when the individual experiments were run using different microarray platforms as long as a common set of candidate genes could be identified and then be subjected to the meta-analysis (rank aggregation).

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 [12]). The usefulness of the cross-entropy method for the rank aggregation problem and other biological

* Corresponding author.

E-mail addresses: vasyl.pihur@louisville.edu (V. Pihur), somnath.datta@louisville.edu (S. Datta), susmita.datta@louisville.edu (S. Datta).

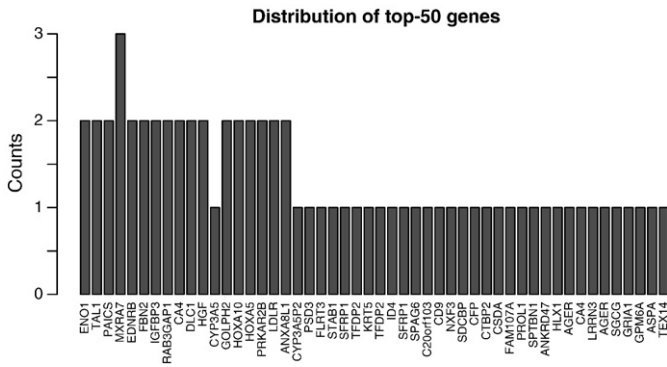


Fig. 1. The distribution of the top-50 genes from our overall list. The height of each bar represents the number of times (out of 20) each gene appears in individual top-50 lists.

applications was recognized by Lin et al. [13] and Pihur et al. [14]. 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!)

In the next section, we report a list of 50 genes (in this work we loosely use “genes” for probe IDs) that are judged to be the most differentially expressed overall in these experiments when the rankings in terms of *p* values are aggregated. This includes a gene coding for a transmembrane anchor protein, *MXRA7*, which appeared in the top-50 list of three individual microarray experiments that we aggregated. The resulting top-50 list contains 36 genes that have been implicated in cancer (often in more than one cancer type) in the literature. The remaining genes are novel in terms of their connection to cancer (as mined from the existing literature). The current analysis

suggests that perhaps they should be investigated further for their regulatory roles related to cancer activities.

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, is then used to produce a single combined top-50 list via the weighted rank aggregation approach mentioned in the previous section. Further details of this procedure appear later in the article.

Applying the rank aggregation procedure to the 20 top-50 lists produces the following overall top-50 (ordered) gene list:

ENO1, TAL1, PAICS, MXRA7, EDNRB, FBN2, IGFBP3, RAB3GAP1, CA4, DLC1, HGF, CYP3A5, GOLM1, 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, KANK3, HLX1, AGER, CA4, LRRN3, AGER, SGCG, GRIA1, GPM6A, ASPA, TEX14.

In Fig. 1, a barplot shows the total number of times each gene in the aggregated list appears within the 20 individual lists. The maximum number of lists that any gene appeared in was 3; only one gene, *MXRA7*, which codes for a transmembrane anchor protein, attained this. Despite the fact that the microarray experiments were related, the majority of genes were on only one top-50 list. This small overlap between the individual top-50 lists may be surprising at first, but perhaps it is not unusual because the number of genes in each experiment is rather large. Indeed, DeConde et al. [11] observed similar patterns, where they aggregate five top-25 lists obtained from five microarray datasets (all prostate cancer). In their example, individual lists contained 89 unique genes, of which 23 appeared more than once

Table 1
Top-8 probe IDs from the combined list

Probe ID	Symbol	GO processes	Kegg pathways	Oncology	PubMed ID	Description
217294_s_at	ENO1	Negative regulation of cell growth, glycolysis, regulation of transcription		Breast carcinoma	7641187 9074493	Negative regulatory function by down-regulating c-myc expression Transcriptional repressor activity on c-myc promoter
206283_s_at	TAL1	Regulation of transcription, cell proliferation, cell differentiation	HSA04310: wnt signaling pathway	t-all	9695959 8208530	Selectively represses Bcl-xL expression in MCF-7 cells and induces mitochondrial involvement in the apoptotic process Disrupted by translocation or deletion (tal(d)) in up to 30% of T-cell acute lymphoblastic leukaemia (T-ALL)
201014_s_at	PAICS	Purine base biosynthetic process, de novo IMP biosynthetic process		Leukemia Cancer	2040693 17224163	tal-1 rearrangements This study provides essential structural information for designing PAICS-specific inhibitors for use in cancer chemotherapy
212509_s_at	MXRA7	Integral to membrane		Lung carcinoma	15246564	
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	16627760 15569975	Tissue matrix remodeling-like gene
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, regulation of cell growth, positive regulation of apoptosis		Breast cancer	8609661	Involved in the regulation of breast cancer cell growth
212932_at	RAB3GAP1	Regulating GTPase activity			10067859	

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.

and only one gene appeared in all five lists. This also suggests that a visual inspection of the individual top-50 list will not be enough to form the overall ranking of the top-50 genes and a mathematical procedure such as the one proposed here is needed to achieve this.

The full table is provided in Appendix A: Supplementary material. It shows the list of overall top-50 genes along with their Go biological functions mined through Osprey [15], Kegg pathways mined through David [16], and oncological implications with relevant PubMed IDs for the most part identified through Gene Cards available at <http://www.genecards.org>. In Table 1 we show the first 8 probe IDs from the larger table on the supplementary website.

Thirty-six of the fifty genes on our aggregated list have been previously implicated in different cancers, many of them, as we would expect, in multiple cancers. The first gene on the list, *ENO1*, which is responsible for the negative regulation of cell growth among its other functions, has been previously implicated in breast, cervical, and lung cancers. Please refer to Table 1 for the PubMed IDs of the relevant articles. Other genes at the top list, such as *TAL1*, *PAICS*, *EDNRB*, *FBN2*, and *IGFBP3*, have been associated with at least one form of cancer. Existing evidence in the literature suggests that gene *CYP3A5* (12 on our list) is linked with leukemia, prostate, breast, and colorectal cancers. Gene 14, *HOXA10*, plays a role in multiple forms of leukemia and endometrial cancer. Many more examples of genes' involvement in multiple cancers can be found. Gene *AGER* represented by probe IDs 42 and 45 appearing toward the end of the list claims partial responsibility for lung, colorectal, prostate, and pancreatic cancers.

Not all of the genes in our list have been previously associated with any form of cancer. These remaining 14 genes, among which are *RAB3GAP1*, *GOLM1*, *ANXA8L1*, and *CYP3A5P2*, may have to be further studied in connection with cancer development. For most of them, not much biological information in the form of Go annotations and pathways is available at present.

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 among them was less than 0.033%. In the above, we reported the list corresponding to the lowest value of the objective function out of the three. The algorithm converged in 47 iterations. The lists were very similar in terms of which probe IDs were included (at least 40 probe IDs from the reported list were included in the other two lists) but differed in the actual ordering, especially at 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.

Alternative aggregation methods

For the purpose of finding the optimal solution to this minimization problem we also tried the genetic algorithm (GA), but it failed to converge, probably because it was unable to successfully explore the search space of this size (10^{148}). Both CE and GA algorithms for our weighted rank aggregation are implemented in an R package, *RankAggreg*, publicly available on CRAN (the Comprehensive R Archive Network).

In addition, we implemented one of the rank aggregation algorithms discussed in [11], namely, the MCT algorithm, and obtained an aggregated list. A comparison with the list obtained from the cross-entropy algorithm revealed 16 probe IDs in common. It is important to note, however, that the MCT algorithm does not take into consideration the p values obtained from the F tests (or t tests) according to which the individual lists themselves are ordered.

Discussion

The weighted rank aggregation method based on the cross-entropy Monte Carlo algorithm proves to be a useful tool for 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 based on multiple experiments. Here we limit our investigation to $k=50$. A larger list can also be produced at a greater computational cost.

Using cross-entropy Monte Carlo algorithms with Spearman's footrule distance as a measure of "closeness" between two ordered lists is one of the many possible approaches to rank aggregation. It must be noted here that the problem of rank aggregation has a long history, which has its origins in voting theory. The Borda count is probably the most famous and intuitive rank aggregation scheme in which each element in each ordered list is given a score depending on its rank and then these weights are summed up across all such lists. Elements in the aggregated list are given in descending order according to the overall scores. An alternative to the Borda count, the Condorcet method, can also be considered. It performs aggregation based on the number of pairwise wins of each element. The more wins, the higher the ranking in the final list. These two approaches represent the competing philosophies on rank aggregation. While the Borda count seeks "consensus" among the lists to be ordered and will usually ensure that elements that are consistently at the top of individual lists surface to the top of the aggregated list, the Condorcet method gives the advantage to elements favored by the majority of the lists, neglecting the few where these particular elements are ranked at the very end.

Both Borda and Condorcet methods can be easily applied when the ordered lists to be aggregated are mere permutations of each other. When this is not the case, as with our ordered lists of genes, both Borda scores and pairwise wins will result in a very large number of ties that cannot be easily resolved. This is one reason our method for rank aggregation makes use of additional information available in the form of weights (p values, in this case), which virtually eliminate the possibility of such ties. The second advantage of the rank aggregation method implemented here is the formal framework of a well-defined minimization problem, which has more appeal than somewhat arbitrary criteria of both the Borda and Condorcet methods. This rank aggregation can also be extended by considering different distance measures between lists, for example, Kendall's tau distance. As shown here the cross-entropy algorithm is suitable for solving the underlying optimization problem.

Materials and methods

Meta-dataset

In this article, we describe our analysis of 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 that study the whole range of different conditions in humans. All of the individual microarray studies were hybridized with the Affymetrix GeneChip Human Genome HG-U133A and record expression levels for 22,283 probe IDs.

The goal of our meta-analysis is to identify genetic factors that are common across different types and stages of cancer. For that purpose, we selected 20 different cancer-related microarray experiments (leukemia, neuroblastoma, thyroid carcinoma, lung cancer, prostate cancer, breast cancer, and some other tissues) that 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 parentheses: 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/>).

Rank aggregation

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

- 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 (and their order of importance) 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 are 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. These individual ranked lists can be combined to produce a top- k list using the rank aggregation method provided in [14]. In the present context it can be expressed as the following optimization problem. Find the aggregated ordered list δ^* that minimizes

$$\sum_M d(\delta, L_M),$$

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 $p(1, M), \dots, p(k, M)$ be the p values corresponding to the top- k probe IDs and $r^{L_M}(G)$ 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

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

The above optimization problem is solved using the cross-entropy Monte Carlo algorithm proposed by Rubinstein for solving large combinatorial optimization problems [12]. An ordered list of k out of n genes is expressed as an $n \times k$ matrix of entries 0 and 1 satisfying certain constraints and a stochastic search is conducted in a sequential manner. In our analysis reported here, $k=50$ and $n=966$ is the number of unique probe IDs in the union of the 20 top-50 lists. We omit further details of the CE algorithm as they can be found in the earlier papers (e.g., [17,13,14]).

Acknowledgment

This research was partially supported by grants from the United States National Science Foundation to S.D. and S.D.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.05.003.

References

- [1] F.I. Raynaud, S. Eccles, P.A. Clarke, A. Hayes, B. Nutley, S. Alix, A. Henley, F. Di-Stefano, Z. Ahmad, S. Guillard, L.M. Bjerke, L. Kelland, M. Valenti, L. Patterson, S. Gowan, A. de Haven Brandon, M. Hayakawa, H. Kaizawa, T. Koizumi, T. Ohishi, S. Patel, N. Saghir, P. Parker, M. Waterfield, P. Workman, Pharmacologic characterization of a potent inhibitor of class I phosphatidylinositol 3-kinases, *Cancer Res.* 67 (2007) 5840–5850.
- [2] Y. Samuels, Z. Wang, A. Bardelli, N. Silliman, J. Ptak, S. Szabo, H. Yan, A. Gazdar, S.M. Powell, G.J. Riggins, J.K. Willson, S. Markowitz, K.W. Kinzler, B. Vogelstein, V.E. Velculescu, High frequency of mutations of the *pik3ca* gene in human cancers, *Science* 304 (2004) 554.
- [3] I.G. Campbell, S.E. Russell, D.Y. Choong, K.G. Montgomery, M.L. Ciavarella, C.S. Hooi, B.E. Cristiano, R.B. Pearson, W.A. Phillips, Mutation of the *pik3ca* gene in ovarian and breast cancer, *Cancer Res.* 64 (2004) 7678–7681.
- [4] M. Hollstein, D. Sidransky, B. Vogelstein, C.C. Harris, *p53* mutations in human cancers, *Science* 253 (1991) 49–53.
- [5] N.S. Pellegata, G.N. Ranzani, The significance of *p53* mutations in human cancers, *Eur. J. Histochem.* 40 (1996) 273–282.
- [6] A. Bagchi, C. Papazoglu, Y. Wu, D. Capurso, M. Brodt, D. Francis, M. Bredel, H. Vogel, A.A. Mills, *Chd5* is a tumor suppressor at human 1p36, *Cell* 28 (2007) 459–475.
- [7] D.R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, A.M. Chinnaiyan, Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression, *Proc. Natl. Acad. Sci. USA* 101 (2004) 9309–9314.
- [8] E. Segal, N. Friedman, D. Koller, A. Regev, A module map showing conditional activity of expression modules in cancer, *Nat. Genet.* 36 (2004) 1090–1098.
- [9] X. Yang, S. Bentink, R. Spang, Detecting common gene expression patterns in multiple cancer outcome entities, *Biomed. Microdevices* 7 (2005) 247–251.
- [10] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA* 95 (1998) 14863–14868.
- [11] R. DeConde, S. Hawley, S. Falcon, N. Clegg, B. Knudsen, R. Etzioni, Combining results of microarray experiments: a rank aggregation approach, *Stat. Appl. Genet. Mol. Biol.* 5 (2006) Article 15.
- [12] R. Rubinstein, The cross-entropy method for combinatorial and continuous optimization, *Method. Comput. Appl. Probability* 1 (1999) 127–190.
- [13] S. Lin, J. Ding, J. Zhou, Rank aggregation of putative microRNA targets with cross-entropy monte carlo methods (Preprint, presented at the IBC 2006 conference, Montreal).
- [14] V. Pihur, S. Datta, S. Datta, Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach, *Bioinformatics* 23 (2007) 1607–1615.
- [15] B.J. Breitkreutz, C. Stark, M. Tyers, Osprey: a network visualization system, *Genome Biol.* 4 (2003) R22.
- [16] Jr G. Dennis, B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane, R.A. Lempicki, David: Database for annotation, visualization, and integrated discovery, *Genome Biol.* 4 (2003) P3.
- [17] P. De Boer, D. Kroese, S. Mannor, R. Rubinstein, A tutorial on the cross-entropy method, *Ann. Oper. Res.* 134 (2005) 1967.