

Bayesian Decomposition Classification Analysis of the CAMDA 2003 Data Set

Andrew Kossenkov
Fox Chase Cancer Center
Philadelphia, PA 19111 USA
and
Moscow Physical Engineering
Institute
Moscow, Russian Federation
AV_Kossenkov@fccc.edu

Ghislain Bidaut
Fox Chase Cancer Center
Philadelphia, PA 19111 USA
and
Structural and Genetic Information Lab
CNRS-AVENTIS, Marseille, France
G_Bidaut@fccc.edu

Michael Ochs
Fox Chase Cancer Center
Philadelphia, PA 19111 USA
M_Ochs@fccc.edu

ABSTRACT

Cancer is a complex disease, comprising many different specific malfunctions within the body. Because numerous processes occur simultaneously within all cells, the gene expression related to tumor behavior is generally convoluted with expression due to routine metabolic processes. Bayesian Decomposition has been used to isolate expression signatures related to tissue type from background behaviors. Here, we present results of using Bayesian Decomposition coupled with gene ontology to identify differences in tumors and between tumors and normal tissue.

General Terms

Algorithms, Measurement

Keywords

Bayesian methods, cancer

1. INTRODUCTION

Despite numerous advances in treatment, cancer remains the second leading cause of death in the United States and throughout the Western world [Alison *et al.* 1997]. In order to understand cancer development in individual malignancies, the recovery of the process that led to the specific cellular malfunction present in the cancer cells must be identified. Such development generally involves the cellular signaling networks that control cell growth, differentiation, apoptosis, and motility [Kolch 2000; Jacks *et al.* 2002]. Because of the extreme underlying biological complexity of these pathways, observed cancers arise from a myriad of different cellular malfunctions [Cooper 1992; Macdonald *et al.* 1997]. It is from this complex background that microarray analysis attempts to glean insight to improve cancer treatment.

Understanding cancer at early stages remains a critical issue for improving patient survival. The studies in the CAMDA 2003 data set are primarily focused on refinement of the identification of the type of cancer using computational and statistical approaches, as was the focus of a number of early studies using microarrays [Golub *et al.* 1999; Alizadeh *et al.* 2000; Zhang *et al.* 2001]. These methods can also be extended to the discovery of biomarkers in the form of differential levels of production of mRNA [Carr *et al.* 2003; Kikuchi *et al.* 2003; Williams *et al.* 2003], which has the advantage of providing a more viable clinical protocol. In

general these methods aim to use microarray technology to detect disease state from tissue samples, and therefore aim to refine identification of suspect tissues after a biopsy has been performed. The refined identification of the form of cancer aides in tailoring treatment, as histologically different cancers require substantially different therapeutic regimens to maximize patient survival.

While the techniques noted above are useful, they have certain limitations as regards more advanced uses in cancer detection. Cancer is primarily a disease of signaling and newer therapeutics specifically target proteins involved in cellular signaling [Mauro *et al.* 2001; Repka *et al.* 2003; von Mehren 2003]. These therapies are not always effective, and the reason for failure, whether inherent poor interaction or complex cellular response, is unknown. It therefore becomes critical to understand how the protein target and its associated signaling pathways are affected during treatment, in order to identify, for instance, whether the activity of the target has not been affected or whether a different pathway has rescued the cell from apoptosis or stasis. In the first case, a different therapeutic would be needed to target the same pathway, while in the second case, additional, simultaneous therapeutics would be needed. Microarray measurements are being used to provide insight into these questions.

Here we focus on the use of Bayesian Decomposition [Bidaut *et al.* 2002; Moloshok *et al.* 2002; Moloshok *et al.* In Press], together with construction of relational trees between analyses in order to understand the processes at work in the development of cancer.

2. METHODS

The initial data were downloaded from the CAMDA web site (<http://www.camda.duke.edu/camda03/>). This data included Affymetrix CEL files for Harvard and Michigan data sets and files in GenePix format for Stanford data set. The Ontario data was not included because of lack of overlap in annotated genes with the other sets. Only adenocarcinoma and normal samples were picked for analysis. CEL files of Harvard and Michigan data sets were processed with dChip software [Cheng *et al.* 2001]. Expression levels were calculated using PM/MM difference model, and an array with median overall intensity was chosen as the baseline array against which other arrays are normalized. The standard error of each expression level was used as the uncertainty of the measurement. The mean background and foreground intensities of channel 1 and

channel 2 from the GenePix files of Stanford data set were used for normalization using the functional genomics data pipeline [Grant *et al.* In Press], using LOESS with a smoothing parameter of 0.9. Normalized expression levels from channel 2 and channel 1 were then used to calculate expression ratios. Uncertainty of each measurement was estimated from previous experience with microarrays at 30% of the corresponding expression level. For those data points where the ratio was negative or data was missing, the ratio was set to 1.0 and the uncertainty to 289 (equal to maximum ratio across all data points), effectively insuring sure that these data points do not affect the model.

The genes present in the Harvard, Michigan and Stanford data sets were annotated for gene ontology information [Ashburner *et al.* 2000] using the automated sequence annotation pipeline [Kossenkov *et al.* 2003]. Only genes with annotations in all three data sets were retained for further analysis, because we would like to compare across multiple studies. Since the goal of the analysis is to look for differences in expression between tissue types, variation coefficients (standard deviation/average) were calculated for each transcript of all three data sets, and only genes for which variation coefficients were more than 35% were retained. The final data comprised 1216 transcripts from the Harvard data, 1088 from the Michigan data set, and 1337 transcripts from the Stanford data set. These numbers differ, as there are multiple transcripts that refer to the same UniGene clusters, which were taken to be measurements of the same gene.

The samples from the three data sets were classified by tumor stage according to provided patient information. The Harvard data set comprised 6 classes (76 first stage, 24 second stage, 10 third stage, 3 fourth stage, 17 normal, 12 undefined), the Michigan data set comprised 3 classes (67 first stage, 19 third stage, 10 normal) and the Stanford data set comprised 4 classes (17 second stage, 15 third stage, 5 normal, 7 undefined).

Bayesian Decomposition was used to analyze the data, exploring different number of patterns. The Harvard data set was analyzed positing from 6 to 10 patterns, the Michigan data set was analyzed positing from 3 to 7 patterns, and the Stanford data set was analyzed positing from 4 to 8 patterns. The number of patterns was chosen to allow additional patterns to explain the non-tumor related behaviors (e.g., routine metabolism) present in the behaviors, as done previously [Moloshok *et al.* 2002].

Bayesian Decomposition is a matrix decomposition algorithm that allows the encoding of additional prior information within a Bayesian framework [Besag *et al.* 1995]. The input is a set of data in the form of a matrix, D , which describes the measurements of expression levels for genes (rows) over various conditions (columns), which are different samples in this data. In addition, there is a matrix \square that provides estimates of the uncertainty or noise for each individual measurement in D . From this data, two matrices are constructed such that

$$D = AP + \square$$

where P contains k rows giving k patterns within the data across the conditions and A provides a measure of how strongly each gene contains each pattern.

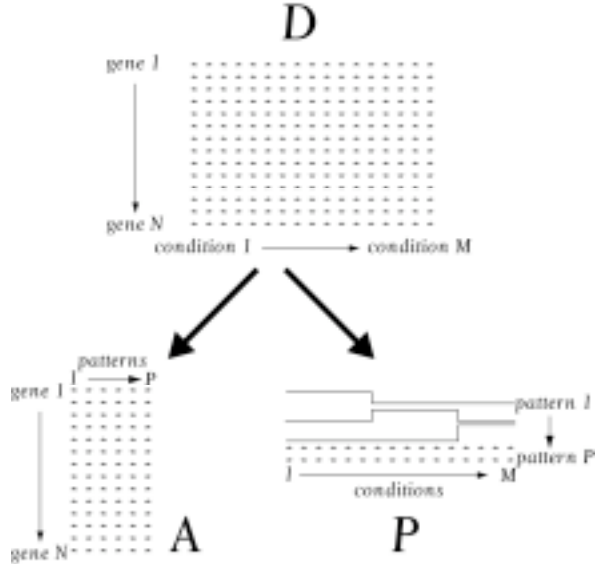


Figure 1: The decomposition performed by Bayesian Decomposition on the data. The conditions are separate samples and the patterns explain variations across the samples. The first three patterns here are enforced to be related to the tumor staging or type, and this number will vary depending on the data set.

The results were compared across different number of posited patterns independently for each data set to identify stable patterns and genes linked to these patterns. Then the results for each data set were compared to other data sets to identify the consistency of the analyses across different data sets. This allowed identification of genes that were consistently associated with tumor stage at different sites and with different platforms.

3. RESULTS

The resulting tree for the Michigan data positing 3 to 7 solutions is shown in figure 2. The first three patterns in each set are locked to the tissue annotation in a way analogous to figure 1. The first group (all 1 in nodes in figure 2) is stage 1 tumor, the second group (2) is third stage tumor, and the third group (3) is normal tissue. The analysis was performed using ClutrFree, a visualization tool allowing global comparison of patterns and clusters in terms of shapes and gene retention (manuscript submitted).

The key issue from figure 2 is the fact that for the Michigan data, variability appears to be primarily explained as variations in behaviors in stage 3 tumors. The relative stability of normal tissue and stage 1 tumors is likely due to the variability of stage 3 tumors due to onset of genetic instability. This variability would then swamp lower level variability in the other tissues. To explore this, we are repeating the analysis with the stage 3 tumors removed from the analysis.

Additional available information can be used to interpret these results. The association of genes with specific nodes and the persistence of those genes across multiple levels of the trees can provide a reliability estimate of the assignment of a gene to a pattern. The gene ontology can then be used to interpret the functions present within the patterns. In this case, one gene ontological group appeared strongly suppressed in the tumor samples only. This was the MAPKKK cascade ontology term.

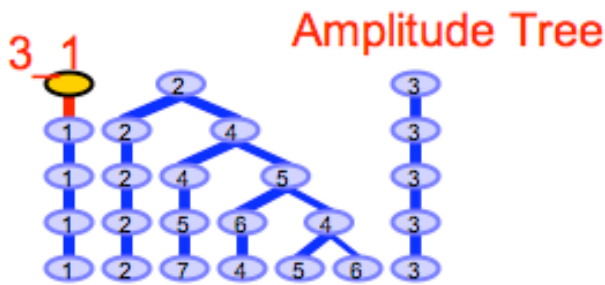


Figure 2: The relationship in the patterns deduced in gene assignment to the patterns positing 3 to 7 patterns. The thickness of the blue line shows the consistency in retention of the genes between different nodes. The red line is an artifact of that node being chosen for further analysis when the image was exported.

Full results involving all data sets will be presented at the meeting. This will include analyses across multiple patterns as in figure 2 and a full discussion of ontological enhancement in the patterns. Finally, comparisons across data sets will be shown.

4. REFERENCES

Alison, M. and C. Sarraf. Understanding Cancer. Cambridge, Cambridge University Press. 1997.

Alizadeh, A. A., M. B. Eisen, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403(6769) (2000): 503-11.

Ashburner, M., C. A. Ball, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1) (2000): 25-9.

Besag, J., P. Green, et al. Bayesian computation and stochastic systems. *Statistical Science* 10(1) (1995): 3 - 66.

Bidaut, G., T. D. Moloshok, et al. Bayesian Decomposition analysis of gene expression in yeast deletion mutants. in K. Johnson and S. Lin. Methods of Microarray Data Analysis II. Boston, Kluwer Academic (2002): 105-122.

Carr, K. M., M. Bittner, et al. Gene-expression profiling in human cutaneous melanoma. *Oncogene* 22(20) (2003): 3076-80.

Cheng, L. and W. H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc Natl Acad Sci U S A* 98 (2001): 31-36.

Cooper, G. M. Elements of Human Cancer. Boston, Jones and Bartlett Publishers. 1992.

Golub, T. R., D. K. Slonim, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286(5439) (1999): 531-7.

Grant, J. D., L. A. Somers, et al. FGDP: functional genomics data pipeline for automated, multiple microarray data analyses. *Bioinformatics* (In Press).

Jacks, T. and R. A. Weinberg. Taking the study of cancer cell survival to a new dimension. *Cell* 111(7) (2002): 923-5.

Kikuchi, T., Y. Daigo, et al. Expression profiles of non-small cell lung cancers on cDNA microarrays: identification of genes for prediction of lymph-node metastasis and sensitivity to anti-cancer drugs. *Oncogene* 22(14) (2003): 2192-205.

Kolch, W. Meaningful relationships: the regulation of the Ras/Raf/MEK/ERK pathway by protein interactions. *Biochem J* 351 Pt 2 (2000): 289-305.

Koskenkov, A., F. J. Manion, et al. ASAP: automated sequence annotation pipeline for web-based updating of sequence information with a local dynamic database. *Bioinformatics* 19(5) (2003): 675-676.

Macdonald, F. and C. H. J. Ford. Molecular Biology of Cancer. Oxford, BIOS Scientific Publishers, Ltd. 1997.

Mauro, M. J. and B. J. Druker. STI571: targeting BCR-ABL as therapy for CML. *Oncologist* 6(3) (2001): 233-8.

Moloshok, T. D., D. Datta, et al. Bayesian Decomposition classification of the Project Normal data set. in K. Johnson and S. Lin. Methods of Microarray Data Analysis III. Boston, Kluwer Academic (In Press).

Moloshok, T. D., R. R. Klevecz, et al. Application of Bayesian Decomposition for analysing microarray data. *Bioinformatics* 18(4) (2002): 566-75.

Repka, T., E. G. Chiorean, et al. Trastuzumab and interleukin-2 in HER2-positive metastatic breast cancer: a pilot study. *Clin Cancer Res* 9(7) (2003): 2440-6.

von Mehren, M. Recent advances in the management of gastrointestinal stromal tumors. *Curr Oncol Rep* 5(4) (2003): 288-94.

Williams, N. S., R. B. Gaynor, et al. Identification and validation of genes involved in the pathogenesis of colorectal cancer using cDNA microarrays and RNA interference. *Clin Cancer Res* 9(3) (2003): 931-46.

Zhang, H., C. Y. Yu, et al. Recursive partitioning for tumor classification with gene expression microarray data. *Proc Natl Acad Sci U S A* 98(12) (2001): 6730-5.

[1]