Visual analysis of statistical results from microarray studies of human breast cancer

DAVID M. REIF[1] and JASON H. MOORE[1,2]

[1]Computational Genetics Laboratory, Norris Cotton Cancer Center, Department of Genetics, [2]Department of Community and Family Medicine, Dartmouth Medical School, Lebanon, NH; Department of Biological Sciences, Dartmouth College, Hanover, NH; Department of Computer Science, University of New Hampshire, Durham, NH; Department of Computer Science, University of Vermont, Burlington, VT, USA

Received July 7, 2005; Accepted September 30, 2005

Abstract. Computational and statistical analysis of microarray data is a daunting challenge. Perhaps even more daunting is the biological interpretation of microarray data analysis results. We have previously developed the exploratory visual analysis (EVA) software and database for exploring data analysis results in the context of biological information on each gene available in public databases such as Entrez Gene. EVA brings a flexible combination of statistics and biological annotation to the user's desktop in a straightforward visual interface. Using a publicly available microarray dataset of gene expression response to chemotherapeutic agents in human breast cancer cell lines, we demonstrate the usefulness of the EVA system for interpreting statistical results. EVA can extend previous analyses as well as aid in making novel discoveries. Thus, we anticipate EVA will prove a useful addition to the repertoire of computational methods for microarray data analysis. The EVA software is freely available to academic users.

Introduction

Visualization as an analytical strategy. Modern experimental methods have generated an explosion in the volume of raw data that must be analyzed. The downside to this information abundance is that cancer researchers must wade through an analytical maze of spreadsheets and arbitrary statistical significance thresholds. Visualization is a proven solution to this challenge of scale. In his work on visualizing quantitative information, Tufte states that 'the most effective way to describe, explore, and summarize a set of numbers - even a very large set - is to look at pictures of those numbers. Furthermore, of all methods for analyzing and communicating statistical information, well-designed data graphics are usually the simplest and at the same time the most powerful' (1). Thus, effective analysis tools are needed to organize and display results so that readily distinguishable patterns emerge.

The novel combination of features in EVA supports comprehensive analysis. Other analysis tools have appeared in recent years, such as DAVID (2), FatiGO (3), GoMiner (4), GoSurfer (5), GOTree Machine (6), and Onto-Express (7). Each of these tools fills a specific niche in the community of genomic analysis methods. However, each has significant functional limitations that must be addressed. The drawbacks include inflexibility, dependence on only those statistical methods provided, inadequate information display density, sluggishness, and a lack of any mechanism to replicate analyses. These gaps in the functionality of other available tools leave a critical void in genomic analysis and bioinformatics. The EVA system was developed to address the limitations of other approaches to analysis of genomic results. Combining flexibility, speed, and visualization of both statistical and annotative information into a single package, EVA fulfills a crucial role in comprehensive microarray analysis.

EVA is designed for flexibility across a wide range of research goals - allowing a truly exploratory analysis. The software can handle any kind of statistical result(s) for any number of experiments. The user is thus free to use any statistic of choice or to define a custom statistic, rather than be limited by those implemented in the software. EVA's graphical results display can be organized into nested groupings for any combination of six biological categories: gene ontology (GO) (8), biopath (9), domain (10), map location (11), chromosome (11), and phenotype (12). Permutation testing is used to assess the statistical significance of certain biological groups that contain a higher proportion of differentially expressed genes relative to other groups. As a complement to statistical analysis, EVA links to multiple annotation sources via Entrez Gene (13). This aspect affords immediate evaluation of the biological relevancy of candidate genes or groups of genes. To ensure that the user can replicate findings, EVA incorporates a printable command log feature into the graphical user interface (GUI).

Correspondence to: Dr Jason H. Moore, HB7937, One Medical Center Dr., Dartmouth-Hitchcock Medical Center, Lebanon, NH 03756, USA
E-mail: jason.h.moore@dartmouth.edu

Key words: software, data mining, pathways, gene ontology

Speed enhances EVA's interactive flexibility. Switching between annotative groupings, statistics, significance levels, and display modes occurs seamlessly, because all of the data and links are loaded into memory upon opening the software. The permutation testing and reporting features provide an immediate complement to visual exploration.

Visualization binds the various components of EVA into a single coherent exploratory tool. Users can instantaneously modify color choices, sliding significance scales, category display thresholds, and other display parameters to highlight extremely significant genes, marginally significant genes, or both. Most importantly, the graphical nature of EVA facilitates interpretation of multitudes of information simultaneously. Since EVA translates numbers into pictures and vice-versa, graphical discoveries can be verified statistically, and statistical significance can be verified graphically. In order to have confidence that all aspects of the data have been sufficiently explored, both types of information are essential. The ability of graphics to rationally condense vast amounts of data is crucial for evaluating biological systems in which the concerted action of numerous contributing factors is the final determinant of phenotype.

EVA brings together diverse abilities that allow the kind of comprehensive analysis necessary to answer complex biological questions. The annotative groupings and immediate expert-knowledge links provided by EVA are essential to understanding diseases of a systemic nature, such as cancer (14). The synergistic pieces of EVA coalesce into an analysis tool wherein the visual, numerical, and annotative components are entirely complementary. EVA encourages a truly integrated analysis - beyond spreadsheets of flat statistical results.

Materials and methods

Details of the EVA software package. Versatility, speed, and user-friendliness are the key design elements of the EVA database. The user downloads the client, which provides a portal to the EVA web server via the custom GUI. This architecture provides ease of security, distributability, and expandability. Thus, once the user has downloaded the client, updates or expansions of the EVA modules are transparent from the user's perspective. Performance is not limited by demand on the web server because information about a particular experiment is stored in memory upon loading. This aspect negates the query lag time typical of other analysis packages.

The graphical user interface (GUI) is the portal through which the user manipulates the components of the EVA package. Upon opening the software, the user supplies a username and secure password. Login grants the user access to interfaces for various administrative tasks, including creating, updating, deleting, or loading experiments and results. Defining a new experiment involves deciding upon a descriptive name, choosing the type of gene identifier (Affymetrix, Genbank, etc.), selecting or defining the statistical tests used, and uploading the text file of results. Once a new experiment has been defined or an existing experiment selected, all of the results and links for that experiment are loaded for viewing on the user's desktop. An abbreviated overview of features is provided in Fig. 1, and a more complete description of the EVA interface is given in ref. 15. Additionally, a comprehensive, illustrated help menu is included with the software. It is important to note that the EVA database is designed to store statistical results (e.g. p-values) and not raw microarray data. As such, there are no confidentiality concerns since only summary statistics are stored.

Application of EVA to a breast cancer microarray dataset. The dataset to which we applied EVA is described fully in ref. 16 and is publicly available at the UNC Microarray Database (https://genome.unc.edu). Four breast cancer cell lines (ME16C, HME-CC, MCF-7, and ZR-75) were treated with the chemotherapeutics 5-fluorouracil (5FU) or doxorubicin (DOX). In total, there were 25 samples treated with 5FU and 26 samples treated with DOX (Table I). Cy5- and Cy3-labeled cDNA was synthesized using mRNA harvested from treated or cell-line-specific reference samples. Expression was measured on custom microarrays created in the University of North Carolina at Chapel Hill Genomics Core Facility and spotted with human oligonucleotides representing ~22,000 genes.

The significance analysis of microarrays (SAM) procedure was used to identify genes that were differentially expressed between treatment classes (17). Prior to analysis, genes were excluded that did not have median intensity greater than twice the median background for both the red (Cy5) and green (Cy3) channel in ≥70% of the experiments. The $\log_2$ ratio of the median red intensity over median green intensity was calculated for each gene. Imputation of missing data was carried out using the k-nearest neighbors procedure included with the SAM software. A two-class, unpaired SAM analysis compared all 5FU-treated samples to all DOX-treated samples. SAM computes a q-value for each gene, which can be interpreted in a manner similar to the familiar p-value. The q-value represents the chance that the observed treatment group difference is really a false positive, and it is the lowest false discovery rate at which that gene can be called significantly different between treatment classes (see ref. 18 for a discussion of the false discovery rate as it applies to multiple testing problems). The resulting list of q-values was loaded into EVA.

Results

EVA can extend previously published results. To identify genes that distinguish DOX- from 5FU-treated breast cell lines, the original authors used SAM on subclasses of the full dataset, as well as a combination of statistical methods including prediction analysis of microarrays (19), k-nearest neighbors, and the gene selection approach described by Dudoit and Fridlyand (20). While their approach is certainly valid, it demands a great deal of statistical expertise and requires the researcher to bridge the usual gap between statistical results and biological knowledge. By using EVA to examine the results of a simplified application of the SAM procedure, we were able to identify treatment-specific gene expression changes having simultaneous statistical and biological significance.

For example, from their extensive lists of statistically significant genes, the original authors highlighted the fact

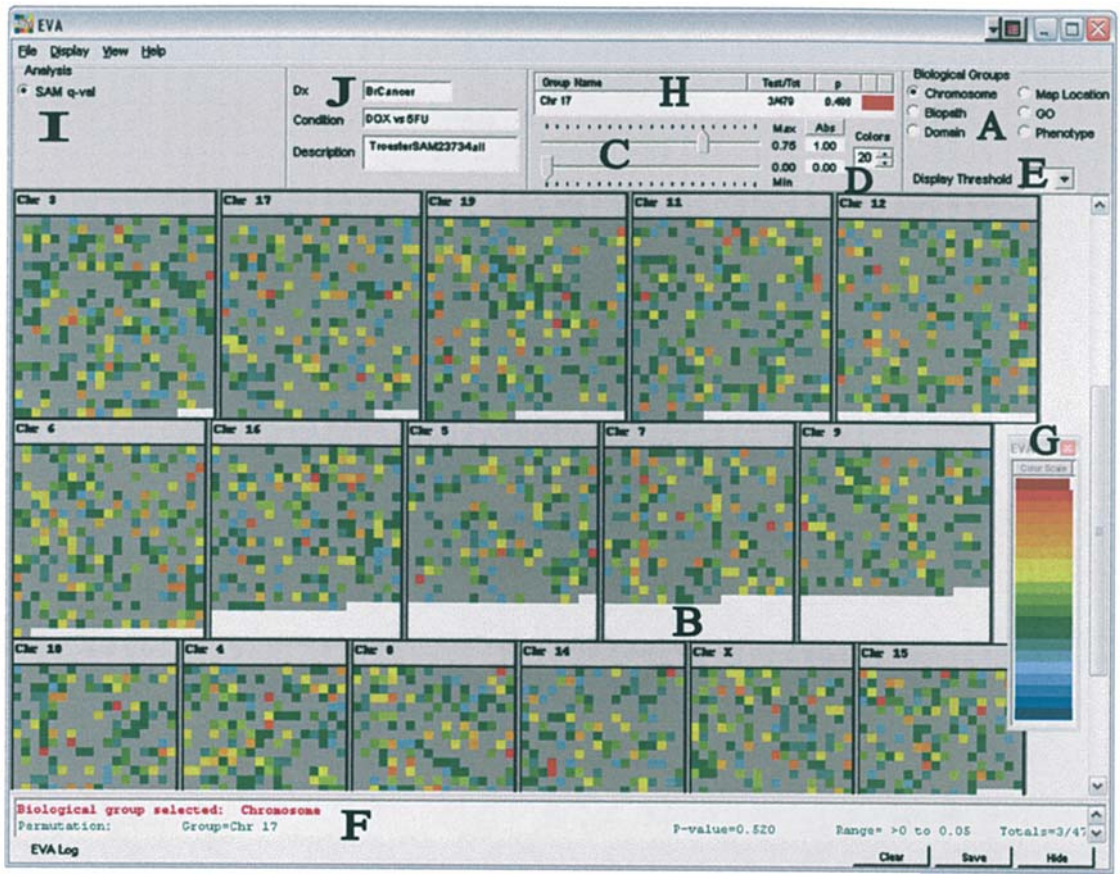

Figure 1. Brief overview of the EVA graphical user interface. Each colored square represents a single gene. Visible features include: (A) biological annotation categories, (B) group boxes, (C) significance threshold color sliders, (D) number of color intervals displayed, (E) display threshold, (F) log, (G) significance range selector for permutation testing, (H) permutation testing results (also shown in log), (I) analysis types, (J) description of current experiment. In addition to options under the file, display, view, and help menus, right-clicking accesses a number of features not illustrated here.

Table I. Samples in each treatment group.

	Treatment group	
	5FU	DOX
MCF-7	6	7
ZR-75	7	6
ME16C	6	6
HME-CC	6	7
Total	25	26

The number of samples of each human cell line (MCF-7, ZR-75, ME16C, or HME-CC) and the total number treated with each chemotherapeutic agent (5FU or DOX) is given.

that MLH-1, CCNE1, MYBL2, E2F, and ID3 made biological sense. Taking advantage of EVA's biological group enrichment (batch permutation) testing feature, the 'nucleus' gene ontology category - which includes all of the above genes - was identified as significantly enriched. Thus, EVA also compiled this list and automatically assigned biological relevancy. The original authors also found that several ribosomal proteins showed differential expression between treatment groups. At first pass, visual exploration with EVA did not show a remarkable number of significant genes within the ribosomal groupings of either the GO or biopath annotative categories. However, when the significance display range was softened from 0.05 to 0.06, these same ribosomal categories attracted immediate visual interest. Importantly, other biologically plausible ribosomal genes in these categories had q-values just beyond the significance threshold and would have been missed by a purely statistical analysis.

Novel patterns identified using EVA. In order to illustrate the novelty of EVA in packaging a complementary set of visual, statistical, and annotative analysis tools, an example of findings capitalizing upon each of these aspects is given.

Visualizing meaningful patterns in a sea of results is the core of EVA's design philosophy. With the results organized according to gene ontology as in Fig. 2, it is immediately apparent that the 'DNA binding' group has a relatively high number of genes that are statistically significant at the $q<0.05$ level. Performing a permutation test on this category statistically verifies this visual impression that the 'DNA binding' group is enriched with respect to the number of differentially expressed genes it contains. This conclusion also makes intuitive biological sense since it is expected that chemotherapeutic agents such as DOX - a topoisomerase poison - would affect the expression of genes involved in binding DNA (21).

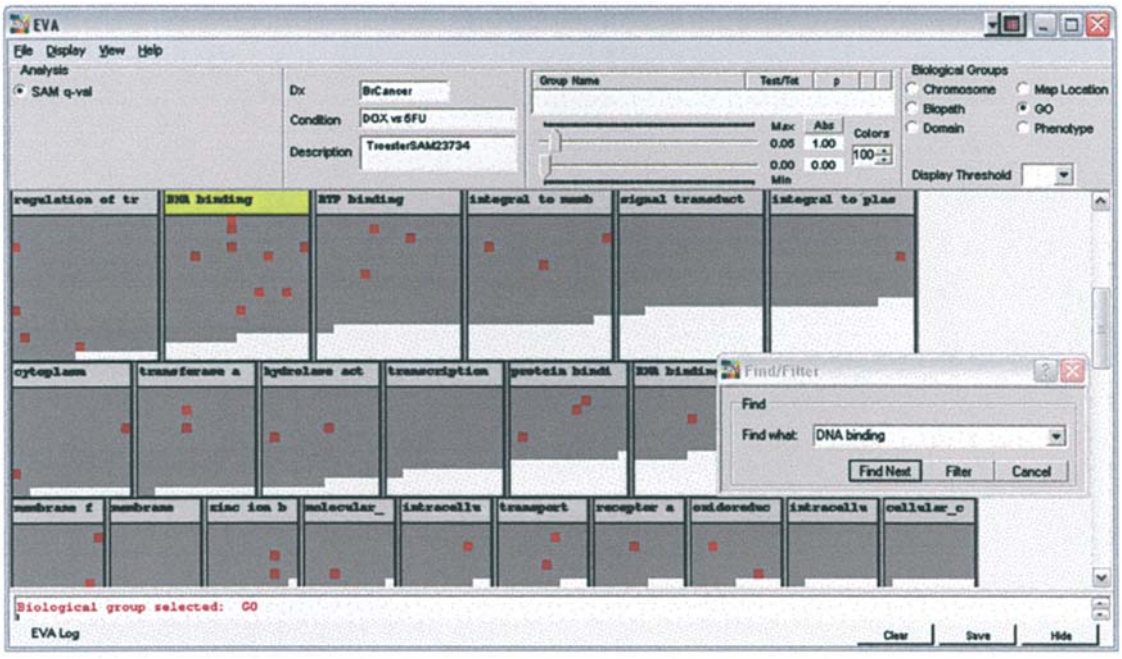

Figure 2. Visual exploration of results organized according to gene ontology groups. With the display slider adjusted to color genes significant at the q<0.05 level, it is apparent that the 'DNA binding' group contains a relatively high number of differentially expressed genes. EVA's find tool is used to highlight the 'DNA binding' group.

The point-and-click interface for statistically testing the relative enrichment of certain biological groups is a useful mechanism for scanning an entire dataset with a single action. The significant results of such a batch test are highlighted in the EVA log and can be written to an external file for further analysis. Using this strategy, interesting biopath groups involved in carbon metabolism were identified. Inspecting the genes contained in these groups revealed several mitochondrial genes. The biological relevancy of this finding is that anthracyclines (such as DOX) are known to affect cellular respiration.

EVA naturally incorporates biomedical knowledge into the analysis via its biologically-organized display and direct links to gene-specific annotation through Entrez Gene. For these data, it is reasonable to assume that differences in expression profiles of cells treated with DOX or 5FU might reflect alterations in transcriptional machinery. Not surprisingly, many differentially expressed genes appear in domain groups representing common transcription factor domains. Additionally, GO groups involving transcriptional initiation and regulation contain many genes identified as significant. Using EVA's permutation testing capabilities statistically reinforces the observed high number of differentially regulated genes in these biological groups relative to others.

Discussion

The broad utility of EVA for oncology. As shown by these results, EVA provides an easy interface for incorporating biological knowledge into large-scale analysis challenges, such as microarrays. The EVA package rationally condenses volumes of data into a graphical display that assures the analysis is grounded within a biological context. The visual, statistical, and annotative abilities of EVA permit a flexible analysis approach - reflecting the expertise or prior notions of the particular user. Thus, EVA has the power to build conclusions supported by a cohesive array of sources without necessitating advanced statistical expertise. This novel combination of visualization, statistics, and biological knowledge in a speedy, user-friendly desktop software package allows for a truly comprehensive exploration of results. We anticipate EVA will be a useful addition to the repertoire of computational methods for microarray data analysis.

Availability and further development. The current Windows-compatible prototype of EVA is available at no charge to academic users. An open-source and platform-independent Java version with additional functionality is in production. More information about the open-source version of EVA can be found at http://www.epistasis.org/open-source-eva-project.html.

Acknowledgements

The authors wish to thank Dr Charles M. Perou and Katherine A. Hoadley for their assistance with the dataset. The authors also wish to thank Scott Dudek, Christian Shaffer, and Janey Wang for their excellent technical assistance. This project was supported through generous support from the Norris Cotton Cancer Center at Dartmouth Medical School and NIH grants RR018787 (J.H.M.) and GM062758 (D.M.R.). The development of the EVA software prototype was previously supported by NIH grant LM007613 (J.H.M.).

References

1. Tufte ER (ed): The Visual Display of Quantitative Information. Graphics Press, Cheshire, CT, 2001.

2. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC and Lempicki RA: DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 4: R60.1-R60.11, 2003.
3. Al-Shahrour F, Diaz-Uriarte R and Dopazo J: FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 20: 578-580, 2004.
4. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC and Weinstein JN: GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 4: R28.1-R28.8, 2003.
5. Zhong S, Storch F, Lipan O, Kao MJ, Weitz C and Wong WH: GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. Appl Bioinformatics 3: 261-264, 2004.
6. Zhang B, Schmoyer D, Kirov S and Snoddy J: GOTree machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics 5: 1-8, 2004.
7. Draghici S, Khatri P, Bhavsar P, Shah A, Krawetz SA and Tainsky MA: Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. Nucleic Acids Res 31: 3775-3781, 2003.
8. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC and Dwight SS: The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32: D258-D261, 2004.
9. Kanehisa M, Goto S, Kawashima S, Okuno Y and Hattori M: The KEGG resource for deciphering the genome. Nucleic Acids Res 32: D277-D280, 2004.
10. Marchler-Bauer A, Anderson JB, De Weese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler GH, Mazumder R, Nikolskaya AN, Panchenko AR, Rao BS, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y and Yamashita RA: CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Res 31: 383-387, 2003.
11. Pruitt KD and Maglott DR: RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29: 137-140, 2001.
12. McKusick-Nathans Institute for Genetic Medicine: Online Mendelian Inheritance in Man. National Center for Biotechnology Information, National Library of Medicine, 2004.
13. Maglott D, Ostell J, Pruitt KD and Tatusova T: Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 33: D54-D58, 2005.
14. Hood L: Systems biology: integrating technology, biology, and computation. Mech Ageing Dev 124: 9-16, 2003.
15. Reif DM, Dudek S, Shaffer C, Wang J and Moore JH: Exploratory visual analysis of pharmacogenomic results. In: Pacific Symposium on Biocomputing. Altman R (ed). World Scientific Publishing Co., Singapore, pp296-307, 2005.
16. Troester MA, Hoadley KA, Parker JS and Perou CM: Prediction of toxicant-specific gene expression signatures after chemotherapeutic treatment of breast cell lines. Environ Health Perspect 112: 1607-1613, 2004.
17. Tusher VG, Tibshirani R and Chu G: Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98: 5116-5121, 2001.
18. Benjamini Y and Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc 57: 289-300, 1995.
19. Tibshirani R, Hastie T, Narasimhan B and Chu G: Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 99: 6567-6572, 2002.
20. Dudoit S and Fridlyand J: A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol 3: R0036.1-R0036.21, 2002.
21. Mizutani H, Tada-Oikawa S, Hiraku Y, Kojima M and Kawanishi S: Mechanism of apoptosis induced by doxorubicin through the generation of hydrogen peroxide. Life Sci 76: 1439-1453, 2005.