BMC Bioinformatics

BioMed Central

Methodology article

Significance analysis of lexical bias in microarray data

Charles C Kim* and Stanley Falkow

Address: Microbiology and Immunology, Stanford University Medical Center, Stanford, CA, 94305, USA

Email: Charles C Kim* - cckim@stanford.edu; Stanley Falkow - falkow@stanford.edu

* Corresponding author

Published: 3 April 2003

BMC Bioinformatics 2003, **4**:12

Received: 8 January 2003

Accepted: 3 April 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/12

Abstract

Background: Genes that are determined to be significantly differentially regulated in microarray analyses often appear to have functional commonalities, such as being components of the same biochemical pathway. This results in certain words being under- or overrepresented in the list of genes. Distinguishing between biologically meaningful trends and artifacts of annotation and analysis procedures is of the utmost importance, as only true biological trends are of interest for further experimentation. A number of sophisticated methods for identification of significant lexical trends are currently available, but these methods are generally too cumbersome for practical use by most microarray users.

Results: We have developed a tool, LACK, for calculating the statistical significance of apparent lexical bias in microarray datasets. The frequency of a user-specified list of search terms in a list of genes which are differentially regulated is assessed for statistical significance by comparison to randomly generated datasets. The simplicity of the input files and user interface targets the average microarray user who wishes to have a statistical measure of apparent lexical trends in analyzed datasets without the need for bioinformatics skills. The software is available as Perl source or a Windows executable.

Conclusion: We have used LACK in our laboratory to generate biological hypotheses based on our microarray data. We demonstrate the program's utility using an example in which we confirm significant upregulation of SPI-2 pathogenicity island of *Salmonella enterica* serovar Typhimurium by the cation chelator dipyridyl.

Background

Many methods are available for the analysis of expression data from microarrays [1–3]. Most software falls into one of two categories: programs which generate a list of significantly differentially regulated genes and programs which perform an analysis based on such a significant genes list (SGL). The first category includes statistical methods (e.g. SAM [4]) and pattern-grouping methods (e.g. Cluster [5]), the end result of either type of analysis being a list of genes which appear to be significantly differentially regulated. The second category of software includes a variety of approaches which search for relationships between the members of the SGL, such as methods which identify common regulatory motifs [6] or establish significance based on genetic position [7].

In browsing a SGL, lexical trends are frequently observed as genes involved in similar metabolic functions are often coordinately regulated. Identification of such lexical trends is critical for the generation of hypotheses for further experimentation, but the large size of a typical microarray dataset makes objective evaluation by eye difficult, if not impossible. Additionally, the genetic content of the organism, the style of the genome annotation, the features selected for representation on the microarray, and experimental design all contain biases which are ultimately reflected in the SGL, which further confounds objective analysis. A number of publications have addressed some of these issues by designing lexical analyses based on functional classifications [8], Medline MeSH terms [9], or Medline titles and abstracts [10–12] that are associated with the genes in a SGL. Each of these approaches draw upon external information to overlay biological relationships onto the expression data, and the significance of these relationships can then be analyzed using standard statistical tools.

The described methods have all been used to identify previously undetected relationships of possible biological significance in published microarray datasets. However, each approach has drawbacks that limit their use as a generalized analysis tool for the average microarray user. Methods based on functional classifications require that the genes be categorized prior to the analysis. This is often not performed by the sequencing group and too laborious to be practical for most microarray end users. Medline-based methods cannot include genes with putative functions in their analyses, as there is no associated literature for such genes. As over 50% of the genes in most genomes are still of putative function, abstract-based approaches exclude a large proportion of the expression data from the analysis. All of the described approaches are also confounded by ambiguity and variability in the annotations used by different researchers (e.g. genes with multiple names). Most importantly, these methods require a significant degree of computer savvy to set up and execute, which makes use impractical for many microarray users. We developed a program, LACK, which performs a simple but flexible lexical analysis of microarray data that circumvents many of the described problems.

Results and Discussion

Observed Frequency of Search Terms in the SGL

The software requires three input files. The first file contains the SGL. This is usually the output of a clustering or statistical analysis which has been conducted to identify differentially regulated genes. The second file is the full, unanalyzed dataset. This dataset should be identical to the dataset used for the SGL generation to avoid introduction of any bias from the pre-analysis steps (array design, hybridization, feature identification, data filtering, etc.). The third file is a user-specified set of search terms to be used for the lexical analysis.

The software first tallies the number of genes in the SGL containing any of the search terms. The gene is counted as a match if any of the search terms are found within the annotation. The total tally of matches within the SGL represents the observed frequency of the search terms in the analyzed data. The software then identifies the frequencies at which the search terms would be found at random in datasets of identical size to the SGL. Two approaches for generating and analyzing the random datasets are available: binomial and Poisson. The choice is dependent on the size of the datasets being used due to computational limitations (described below).

Binomial Statistics

The default statistical method employed is the binomial distribution to model the data. For each member of the test sample, the binary criterion (contains a search term / does not contain any search terms) is applied. Furthermore, the population from which the test sample (SGL) is chosen is already known (the full dataset), so the precise frequencies of genes containing a search term (p = number of genes containing a match / number of genes) or not containing any of the search terms ($q = 1 - p$) can be calculated. For a given population, we can calculate the probability of the observed number of matches (or fewer) in the SGL (k) occurring in a random sample of size N by the cumulative binomial probability function:

$$P(X \le k) = \sum_{j=0}^{k} \frac{N!}{j!(N-j)!} p^j q^{N-j}$$

By using the binomial calculation, there is no need for random sampling, which has the advantages of higher precision in the p-values and reduced computation time. The probability of obtaining j matches, $P(j)$, is calculated for every value of j which generates $P(j) = 0.00001$ and provided in the output in order to generate enough values for plotting both tails of a histogram.

Poisson Analysis

Because microarray datasets are often very large, the factorial computations in the binomial coefficient can easily exceed the capabilities of modern desktop computers. We therefore provide an option to perform Poisson statistics, which does not require very large integers. Random samples (RSs) that are identical in size to the SGL are repeatedly chosen from the full dataset with replacement, and the number of genes containing any of the search terms is tallied for each RS. The cumulative Poisson probability is calculated for the observed number of matches in the analyzed dataset as compared to the distribution of matches in the RSs:

Table 1: Variation in cumulative Poisson probabilities in response to number of random samples and SGL matches

	Matches in SGL					
RSs	**8**	**7**	**6**	**5**	**4**	**3**
10	0.993 ± 0.014	0.981 ± 0.045	0.957 ± 0.057	0.898 ± 0.126	0.799 ± 0.167	0.626 ± 0.240
100	0.996 ± 0.003	0.987 ± 0.008	0.963 ± 0.017	0.914 ± 0.033	0.812 ± 0.059	0.643 ± 0.072
1000	0.996 ± 0.001	0.987 ± 0.003	0.964 ± 0.005	0.913 ± 0.011	0.810 ± 0.017	0.647 ± 0.021

Poisson analysis was performed on the described dataset 100 times using 10, 100, or 1000 random samples. The mean of the cumulative probabilities is reported with two standard deviations. The original SGL contained 8 matches (first column); the additional columns were generated using a synthetic SGL modified to contain the specified number of matches.

$$P(X \le k) = \sum_{j=0}^{k} \frac{e^{-\mu}\mu^{j}}{j!}$$

where k is the number of matches in the SGL and μ is the mean number of matches for all RSs.

We have observed that 100 RSs are generally sufficient to accurately mimic the *p*-values generated by binomial analysis (Table 1). However, using 1000 RSs requires only slightly more computational time, which is easily outweighed by the benefit of a more accurate *p*-value. It is also noteworthy that for more significant numbers of matches, there is less variation in the *p*-values. Therefore, we recommend performing multiple trials when using Poisson statistics, especially in cases where the *p*-value borders on the selected significance threshold.

Additionally, in Poisson output mode, a raw output option can be selected that includes the tally data in the output file for analysis by other statistical methods.

Validation of LACK

We have successfully used LACK to generate and test hypotheses generated from our microarray data. Using our *Salmonella enterica* serovar Typhimurium LT2 microarray [13], we observed that members of *Salmonella* Pathogenicity Island 2 (SPI-2) [14,15] appeared 8 times in the top 256 genes significantly upregulated by the ferrous iron chelator dipyridyl. Because there are 34 genes predicted to be within SPI-2 [14,15], it seemed possible that this number of matches might appear at random. We constructed a list of the 34 genes in SPI-2 and performed LACK analysis on the dataset. The binomial distribution of the calculated expected frequency of matches and the probability values show that the overrepresentation of SPI-2 in the dataset is unlikely to be random (Figure 1). We have confirmed that SPI-2 is indeed induced by dipyridyl treatment using a SPI-2 promoter (*ssa*G) fused to GFP (data not shown). During the course of this work, an independent report also confirmed that SPI-2 is indeed upregulated in response to cation starvation [16].

Conclusions

We have demonstrated one example of how we have employed LACK for statistical analysis of lexical trends. In analyzing our microarray data, we realized that few simple tools are available for assessing significance of observed trends. While there are a number of other lexical analysis methods available, they employ different goals (*i.e.* de novo identification of lexical trends) than LACK (significance analysis of specific lexical trends). The primary advantages of LACK over other lexical analysis methods for the typical microarray user are the simplicity of use and the flexibility in searchable concepts. The inputs are text files that are generated as part of a typical microarray analysis, and the output is provided as tab-delimited text; no external information is required apart from the microarray annotation. The inherent flexibility in user-specified search terms makes LACK potentially useful for many microarray users who are interested in determining whether or not specific trends are significant in their data, rather than performing a global search of all possible trends. In addition, this flexibility makes lexical analysis possible for datasets that are difficult to analyze by other lexical methods due to inconsistencies in annotation nomenclature, as the user can specify terms to encompass multiple naming conventions.

It is important to note that with this flexibility comes an additional point at which bias can be introduced: careless selection of search terms can lead to false determination of significance. For example, search terms should not be chosen only from the SGL itself, but rather from the full dataset. To demonstrate such an inappropriate example, we selected a search list based on metabolites found in the top 20 genes of the SGL described above. While there is no clear relationship between the search list terms (glucosamine, methionine sulfoxide, citrate, 2-aminoethylphosphonate, ethanolamine, UDP-glucose, cysteine), LACK indicates a high level of significance in

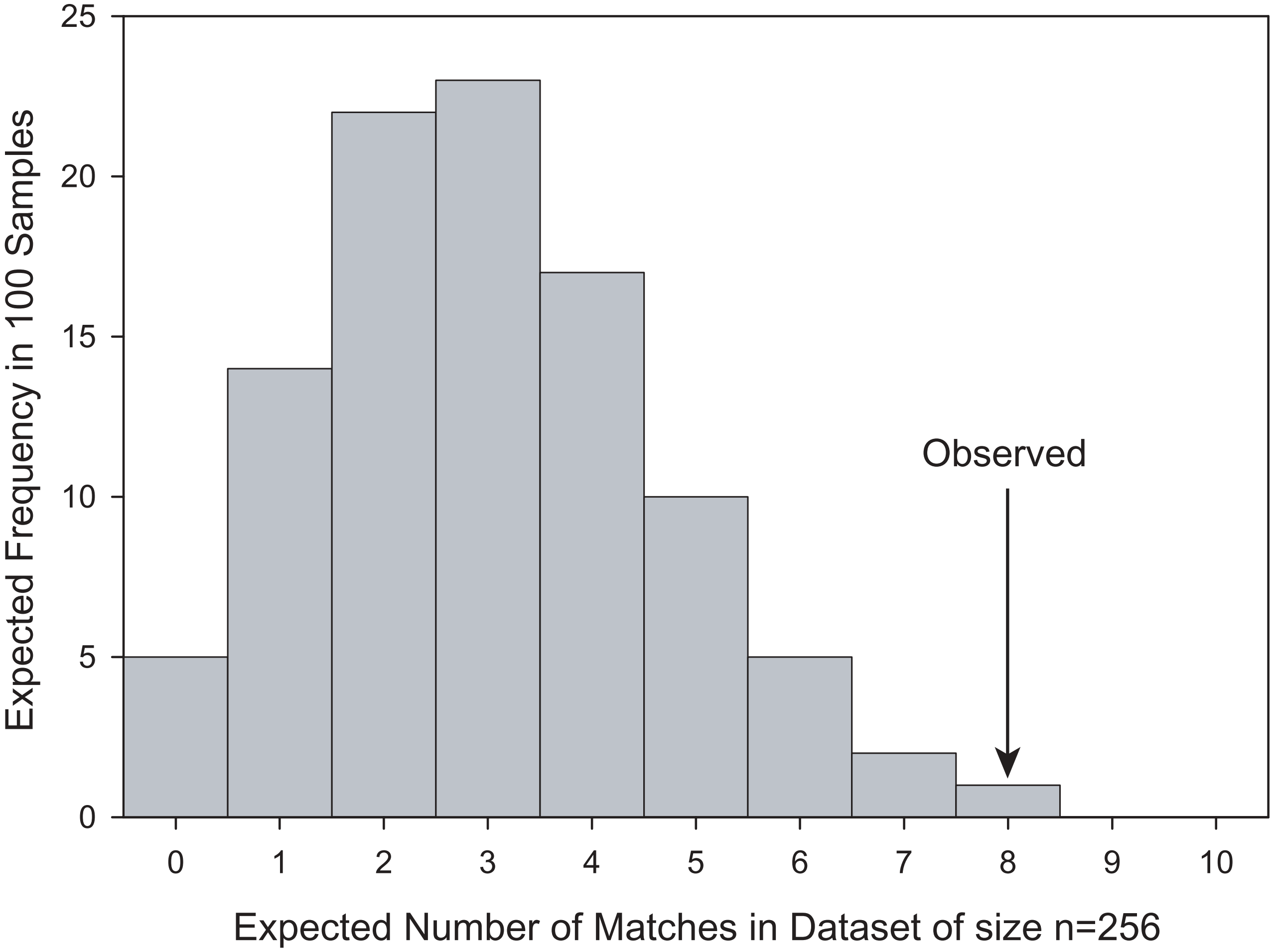

Figure 1
Binomial distribution of SPI-2 genes in a dataset The total filtered dataset consisted on 4290 unique elements. An SGL of 256 genes was generated using SAM and analyzed for 34 members of SPI-2. The arrow indicates the number of matches in the SGL, with $P(x > 8) = 0.004$. The binomial analysis required 5 seconds; Poisson analysis of the same datasets required 7 seconds. A 21,450 element dataset created by replicating the 4290 element dataset 5 times required 8 seconds for binomial analysis. The files used for this analysis are available at the LACK website or as supplementary data.

these terms being over-represented in the SGL ($p < 0.0005$), illustrating the danger of choosing terms directly from the SGL.

One possible approach that avoids this problem is to select a wordlist prior to generation of the SGL. This has the obvious limitation that it cannot be implemented for exploratory microarray experiments, only for specific hypothesis testing. Another approach is to generate the word list in a blinded fashion; for example, a researcher who did not generate the SGL could select search terms from the full dataset. We have also made a tool available, ALACK (automated LACK), which selects single search terms which are over-represented in the SGL in an automated, unbiased fashion (although multiple search terms are not supported in ALACK). While the benefits of LACK far outweigh the potential abuses, it is clear that care must be taken to avoid over-interpretation of significance.

Analysis of the significance of trends in expression data is critical for the generation of further hypotheses, as variation in the arrays and in annotations can confound visual assessment of significance. We have described a method for assessing the statistical significance of lexical trends in microarray data. Our approach is similar to, but distinct from, other lexical analysis methods [3,8–11], in that our method does not incorporate external information. Rather, LACK is designed to analyze the SGL generated from another analysis within the context of the microarray from which the data was generated. We expect that LACK can serve as a useful adjunct to other microarray analysis methods, including other global lexical trend identification methods.

Methods

LACK Programming

LACK was written and tested using ActiveState Perl Build 631 on an Athlon 1.4 GHz system running Windows 2000 Pro. The graphical user interface uses Perl/Tk. The source has also been successfully tested on RedHat Linux 7.3.

Microarray Experiments

Our Salmonella enterica serovar Typhimurium microarray is described elsewhere [13]. Typhimurium SL1344 was grown in unagitated cultures of 100 ml M9 minimal medium in 125 ml Erlenmeyer flasks to an OD_{600} of 0.2. The culture was divided in half and transferred to 50 ml Erlenmeyer flasks. 2,2'-dipyridyl was added to 500 μM to one of the half cultures. Aliquots (10 ml) were removed at 10 min, 30 min, and 60 min post-addition and directly added to 1/10 volume of 95% ethanol / 5% phenol stop-solution. RNA was prepared by a modified Qiagen RNeasy prep protocol http://falkow.stanford.edu and hybridized using standard reverse transcription and cDNA-labeling procedures http://www.microarrays.org.

Authors' contributions

CK performed programming of LACK, microarray experiments, and preparation of the manuscript. SF provided funding and supervision for the experiments described herein. Both authors have read and approved the final manuscript.

Additional material

Additional File 1

LACK Windows Executable. Additional file descriptions text (including details of how to view the file, if it is in a non-standard format).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-12-S1.zip]

Additional File 2

LACK Perl source code. Additional file descriptions text (including details of how to view the file, if it is in a non-standard format).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-12-S2.pl]

Additional File 3

LACK Manual. Additional file descriptions text (including details of how to view the file, if it is in a non-standard format).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-12-S3.pdf]

Additional File 4

Sample Data. Three sample test files, identical to those used for the described analysis, are available. The files are full.txt (full dataset), analyzed.txt (analyzed dataset, SGL), and words.txt (34 members of SPI-2).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-12-S4.zip]

Acknowledgements

We thank D. Scott Merrell for suggesting improvements to the algorithm and Kaman Chan for reading the manuscript. CK thanks the Howard Hughes Medical Institute and the Stanford Graduate Fellowship Program for funding. This work was funded in part by the Stanford Digestive Disease Center grant DK56339.

References

1. Quackenbush J **Computational analysis of microarray data** *Nat Rev Genet* 2001, **2:**418-427
2. Kaminski N and Friedman N **Practical approaches to analyzing results of microarray experiments** *Am J Respir Cell Mol Biol* 2002, **27:**125-132
3. Altman RB and Raychaudhuri S **Whole-genome expression analysis: challenges beyond clustering** *Curr Opin Struct Biol* 2001, **11:**340-347
4. Tusher VG, Tibshirani R and Chu G **Significance analysis of microarrays applied to the ionizing radiation response** *Proc Natl Acad Sci U S A* 2001, **98:**5116-5121
5. Eisen MB, Spellman PT, Brown PO and Botstein D **Cluster analysis and display of genome-wide expression patterns** *Proc Natl Acad Sci U S A* 1998, **95:**14863-14868
6. Liu X, Brutlag DL and Liu JS **BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes** *Pac Symp Biocomput* 2001, 127-138
7. Zimmer DP, Soupene E, Lee HL, Wendisch VF, Khodursky AB, Peter BJ, Bender RA and Kustu S **Nitrogen regulatory protein C-controlled genes of Escherichia coli: scavenging as a defense against nitrogen limitation** *Proc Natl Acad Sci U S A* 2000, **97:**14674-14679

8. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ and Church GM **Systematic determination of genetic network architecture** *Nat Genet* 1999, **22:**281-285
9. Masys DR, Welsh JB, Lynn Fink J, Gribskov M, Klacansky I and Corbeil J **Use of keyword hierarchies to interpret gene expression patterns** *Bioinformatics* 2001, **17:**319-326
10. Blaschke C, Oliveros JC and Valencia A **Mining functional information associated with expression arrays** *Funct Integr Genomics* 2001, **1:**256-268
11. Jenssen TK, Laegreid A, Komorowski J and Hovig E **A literature network of human genes for high-throughput analysis of gene expression** *Nat Genet* 2001, **28:**21-28
12. Raychaudhuri S, Schutze H and Altman RB **Using text analysis to identify functionally coherent gene groups** *Genome Res* 2002, **12:**1582-1590
13. Chan K, Baker S, Kim CC, Detweiler CS, Dougan G and Falkow S **Genomic Comparison of Salmonella enterica Serovars and Salmonella bongori by Use of an S. enterica Serovar Typhimurium DNA Microarray** *J Bacteriol* 2003, **185:**553-563
14. McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F, Hou S, Layman D, Leonard S, Nguyen C, Scott K, Holmes A, Grewal N, Mulvaney E, Ryan E, Sun H, Florea L, Miller W, Stoneking T, Nhan M, Waterston R and Wilson RK **Complete genome sequence of Salmonella enterica serovar Typhimurium LT2** *Nature* 2001, **413:**852-856
15. Hensel M, Shea JE, Waterman SR, Mundy R, Nikolaus T, Banks G, Vazquez-Torres A, Gleeson C, Fang FC and Holden DW **Genes encoding putative effector proteins of the type III secretion system of Salmonella pathogenicity island 2 are required for bacterial virulence and proliferation in macrophages** *Mol Microbiol* 1998, **30:**163-174
16. Zaharik ML, Vallance BA, Puente JL, Gros P and Finlay BB **Host-pathogen interactions: Host resistance factor Nramp1 up-regulates the expression of Salmonella pathogenicity island-2 virulence genes** *Proc Natl Acad Sci U S A* 2002, **99:**15705-15710