BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 8 2011, pages 1174–1175
doi:10.1093/bioinformatics/btr086

Systems biology
Advance Access publication February 23, 2011

BiC: a web server for calculating bimodality of coexpression between gene and protein networks

George C. Linderman[1], Vishal N. Patel[2,3], Mark R. Chance[2,3] and Gurkan Bebek[2,4,*]

[1]Department of Biomedical Engineering, [2]Case Center for Proteomics and Bioinformatics, [3]Department of Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106 and [4]Genomic Medicine Institute, Cleveland Clinic, 9500 Euclid Avenue, Cleveland, OH 44195, USA

Associate Editor: Trey Ideker

ABSTRACT

Summary: Bimodal patterns of expression have recently been shown to be useful not only in prioritizing genes that distinguish phenotypes, but also in prioritizing network models that correlate with proteomic evidence. In particular, subgroups of strongly coexpressed gene pairs result in an increased variance of the correlation distribution. This variance, a measure of association between sets of genes (or proteins), can be summarized as the bimodality of coexpression (BiC). We developed an online tool to calculate the BiC for user-defined gene lists and associated mRNA expression data. BiC is a comprehensive application that provides researchers with the ability to analyze both publicly available and user-collected array data.

Availability: The freely available web service and the documentation can be accessed at http://gurkan.case.edu/software.

Contact: gurkan@case.edu

Received on December 7, 2010; revised on February 1, 2011; accepted on February 9, 2011

1 INTRODUCTION

Recently, a number of techniques have emerged to identify subsets of genes exhibiting bimodal patterns of gene expression, as these distinguishing features are highly suggestive of molecular switches (Hellwig *et al.*, 2010). While many of these approaches attempt to identify the modes of expression represented by a single gene with respect to two phenotypes (Bessarabova *et al.*, 2010; Wang *et al.*, 2009), the modes of expression among sets of genes can also be informative and can be evidenced by bimodal patterns of coexpression. Bimodal coexpression patterns arise among biologically related gene sets whose members have stronger mRNA correlations, $|\rho|$, with each other than with unrelated genes. As these highly related gene sets will have both large positive and large negative values of ρ, their correlation pattern has greater variance than the correlation distribution of unrelated gene sets (which is centered around $\rho = 0$), and Bebek *et al.* (2010) took advantage of this fact to calculate a property which they call the 'Bimodality of Coexpression' (BiC). Using a non-parametric approach, they have previously shown that the BiC can be useful in evaluating the strength of association between a hypothetical signaling network and an experimentally observed set of proteomic targets. In this context, a high degree of coexpression also suggests which genes in the candidate signaling network may act as controllers, or master modulators (Babur *et al.*, 2010), of the expression patterns seen in the proteomic experiment. We have developed an online interface to calculate the BiC for a user-defined set of genes using corresponding mRNA expression data to test models of interaction between these lists. The web interface developed accepts individual experiments in simple omnibus format (SOFT), the standard format for Gene Expression Omnibus (GEO) (Barrett *et al.*, 2009). As the BiC metric can use experimental omics data to evaluate candidate *in silico* network models, the newly developed web interface will be of great value in prioritizing network models for further biological evaluation.

2 IMPLEMENTATION

The starting point to calculate the bimodality of coexpression is two or more gene (or protein) lists provided by the user; as suggested by Bebek *et al.* (2010), one of these lists can be the set of genes posited by a model of a signaling network of interest, and the other list a set of proteomic targets. Through a simple and intuitive interface, BiC calculates the bimodality of coexpression (β) between these two lists by generating coexpression distributions from a given mRNA gene expression experiment, and a P-value based on two-group comparisons.

2.1 Algorithm

First, mRNA coexpression (Pearsons correlation coefficient) and standard t-statistics are calculated for all genes in the array. These two parameters are then used to compute the 'active' coexpression (Bebek *et al.*, 2010), based on user-specified case and control labels. Active coexpression is calculated as the product of a gene coexpression (measured by ρ) and its differential expression (measured by a t-statistic), thus making the analysis dependent on a two-group comparison. The active coexpression matrix relating a given gene list g_i to a target list g_t (g_i, $g_t \subset S$ and $g_i \neq g_t$, where S is the set of all genes on the microarray) is then transformed into vectors and its empirical cumulative distribution function (CDF) that we call $F_{i,t}$; the CDF for the active coexpression matrix relating g_i to the remainder of the genes on the array, S, is also calculated ($F_{i,S}$). It should be noted that, in Bebek *et al.* (2010), g_i represented the set of genes in a candidate signaling network and g_t represented an associated set of proteomic targets. The sample deviation is calculated as the difference of the two CDFs. In short, the bimodality of coexpression, β_i between g_i and g_t is the difference of the second

*To whom correspondence should be addressed.

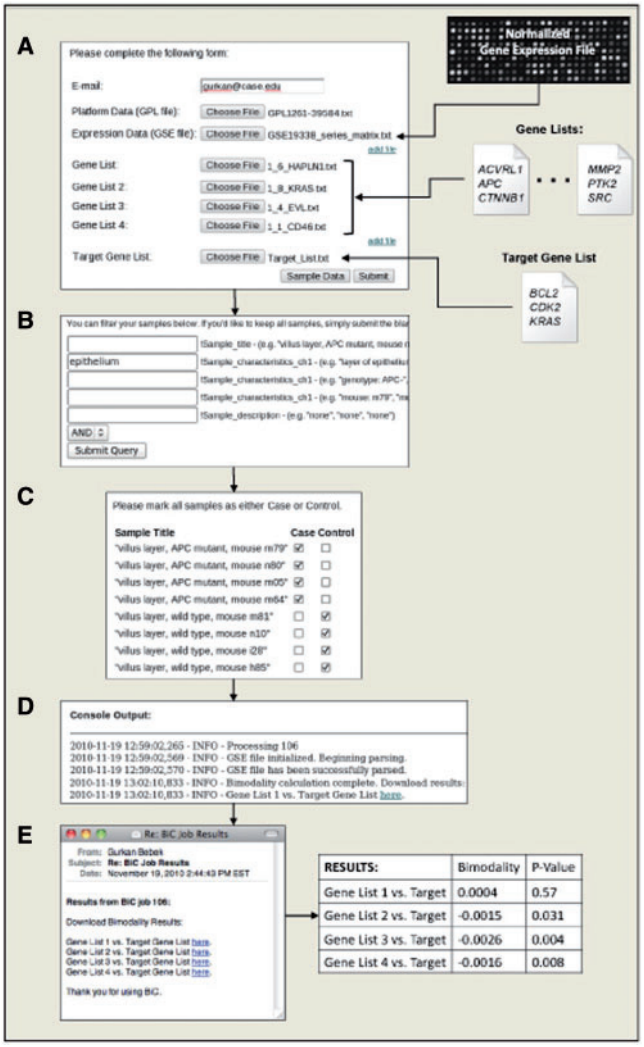

Fig. 1. Workflow of BiC is depicted. (**A**) mRNA gene expression data and gene lists are uploaded. (**B**) The user may filter the uploaded samples by utilizing sample annotations. (**C**) Next, case and control samples are selected. (**D**) BiC processes the job while providing console output. At this stage, the user can leave the website or monitor the progress. (**E**) The BiC web interface provides bimodality of coexpression values and corresponding *P*-values, indicating significant associations with the target gene list. The results are also emailed to the user.

moments of the two empirical distributions, where negative values of β_i represent more correlation than expected (with significant values of β_i on the order of -1×10^{-4}). We calculated the significance of β_i by generating β_{rand} from randomly selected set of candidate targets (1000 such sets of cardinality equal to that of g_t). Then, for the null hypothesis that the coexpression pattern between g_i and g_t is random, the *P*-value represents the probability of attaining at least a value of $|\beta_i|$ via stochastic, i.e. non-biological, generation of random targets. g_i and g_t cannot be equal, and increasing overlap between g_i and g_t will increase β_i. When multiple candidate lists, g_i, are being tested, β_i allows one to prioritize the hypothesized lists (or networks) by the strength of their mRNA coexpression with an experimentally defined target set, g_t.

2.2 Web server

The first step of the BiC web interface requires the user to upload gene expression data in the SOFT format as defined by GEO (Barrett *et al.*, 2007) (Fig. 1A).This allows the user to submit both user prepared files, as well as files downloaded directly from the GEO repository. The uploaded arrays are required to be prenormalized (e.g. via Robust Multiarray Averaging). The user must then upload one or more gene lists (i.e. networks), g_i and a single target list, g_t against which the bimodality of each gene list will be calculated. In the second step, the user can filter the samples. This allows the user to remove samples that are not relevant to the analysis without editing the data files (Fig. 1B). The filtering is done using basic Boolean expressions with respect to sample characteristics in the data files, e.g. sample_characteristics_ch fields in SOFT-formatted sample file. During the third step, the user is asked to label samples as either case or control (Fig. 1C). Finally, the job is queued and the user is presented with the console output of the job progress (Fig. 1D). The results are then displayed in the console output and emailed to the user (Fig. 1E). BiC was primarily developed in Python and the Django framework. To increase speed and handle memory more effectively, the more resource-intensive processes were implemented in C. BiC also implements a queuing system, to handle both large jobs and high traffic.

3 CONCLUSION

We have developed a tool for the analysis of mRNA gene expression data in the context of two user-defined gene lists. The web application uses the mRNA correlation between the gene lists to calculate a parameter called the bimodality of coexpression, or BiC. This new tool linking user-defined gene lists or networks with global experimental measurements provides a way forward in evaluating the functional value of candidate networks, pathways or gene lists. BiC accepts experimental measurements in the widely used GEO SOFT format utilized by publicly available datasets, and it is freely available to the academic community and simple to use.

Funding: National Institute of Health (grants P30-CA043703, UL1-RR024989 and P01-DE019759).

Conflict of Interest: none declared.

REFERENCES

Babur,O. *et al*. (2010) Discovering modulators of gene expression. *Nucleic Acids Res.*, **38**, 5648–5656.

Barrett,T. *et al*. (2007) NCBI geo: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.

Barrett,T. *et al*. (2009) NCBI geo: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.

Bebek,G. *et al*. (2010) Petals: proteomic evaluation and topological analysis of a mutated locus' signaling. *BMC Bioinformatics*, **11**, 596.

Bessarabova,M. *et al*. (2010) Bimodal gene expression patterns in breast cancer. *BMC Genomics*, **11** (Suppl. 1), S8.

Hellwig,B. *et al*. (2010) Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes. *BMC Bioinformatics*, **11**, 276.

Wang,J. *et al*. (2009) The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. *Cancer Inform.*, **7**, 199–216.