BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 21 2010, pages 2792–2793
doi:10.1093/bioinformatics/btq503

*Systems biology*

Advance Access publication September 1, 2010

# CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data

Elana J. Fertig[1,*], Jie Ding[2], Alexander V. Favorov[1,3], Giovanni Parmigiani[2] and Michael F. Ochs[1,*]

[1]Department of Oncology and Division of Oncology, Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, [2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA and [3]Research Institute for Genetics and Selection of Industrial Microorganisms, Moscow, 117545, Russia

Associate Editor: Jonathan Wren

## ABSTRACT

**Summary:** Coordinated Gene Activity in Pattern Sets (CoGAPS) provides an integrated package for isolating gene expression driven by a biological process, enhancing inference of biological processes from transcriptomic data. CoGAPS improves on other enrichment measurement methods by combining a Markov chain Monte Carlo (MCMC) matrix factorization algorithm (GAPS) with a threshold-independent statistic inferring activity on gene sets. The software is provided as open source C++ code built on top of JAGS software with an R interface.

**Availability:** The R package CoGAPS and the C++ package GAPS-JAGS are provided open source under the GNU Lesser Public License (GLPL) with a users manual containing installation and operating instructions. CoGAPS is available through Bioconductor and depends on the rjags package available through CRAN to interface CoGAPS with GAPS-JAGS.

**URL:** http://www.cancerbiostats.onc.jhmi.edu/cogaps.cfm

**Contact:** ejfertig@jhmi.edu; mfo@jhu.edu

**Supplementary Information:** Supplementary data is available at *Bioinformatics* online.

Received on February 25, 2010; revised on August 5, 2010; accepted on August 29, 2010

## 1 INTRODUCTION

Many biological processes (BPs) and phenotypes result from coordinated activity among sets of genes, so that inference from transcriptional measurements using gene sets is more powerful for inferring BPs than inference based on isolated genes. However, gene reuse in BPs is common, so genes are typically multiply regulated. Thus, inference on sets of genes should ideally begin by identifying the portion of each gene's behavior related to its use in a BP. We have developed Coordinated Gene Activity in Pattern Sets (CoGAPS), which infers biological activity by identifying overlapping, coregulated sets of genes and applying Z-score based statistics. CoGAPS can presently be used to isolate transcription factor (TF) or BP activity in datasets of thousands of genes and tens to thousands of samples.

*To whom correspondence should be addressed.

Several methods exist to infer activity of gene sets (GSs). Hypergeometric tests have been used to determine if genes in sets are differentially expressed across samples (Draghici *et al.*, 2003; Tavazoie *et al.*, 1999). These statistics have been extended to rank membership (e.g. Goeman and Buhlmann, 2007). However, these methods do not account for multiple regulation of genes. Matrix factorization techniques have been applied to infer overlapping patterns of coregulation in gene expression, including Non-negative Matrix Factorization (NMF; Lee and Seung, 1999), Bayesian Decomposition (BD; Ochs *et al.*, 1999) and Bayesian Factor Regression Modeling (BFRM; Carvalho *et al.*, 2008). Comparison of matrix factorization techniques on *Saccharomyces cerevisiae* transcriptomic data suggested that MCMC techniques more accurately find patterns that relate to BPs and phenotypes (Kossenkov and Ochs, 2009), inspiring our use of the GAPS MCMC matrix factorization in CoGAPS. Moreover, CoGAPS infers activity in specific gene sets related to the inferred BPs by applying the Z-score based statistic from Ochs *et al.* (2009) to patterns identified with GAPS.

CoGAPS is based on JAGS (Plummer, 2003) and includes an R interface. CoGAPS has been applied in the DESIDE algorithm to identify transcriptional responses to signaling through estimation of the activity of TFs (Ochs *et al.*, 2009). CoGAPS inferred expected decreased activity in the KIT pathway and unexpected activity in p53 and STAT3 pathways from microarrays generated from treated gastrointestinal stromal tumor cell lines and tumor sample data. We provide the data with CoGAPS and an R/Sweave document for this analysis in the Supplementary Material.

## 2 METHODS

CoGAPS takes as input preprocessed microarray measurements in a data matrix $\mathbf{D}$ of $N$ genes and $M$ conditions, an uncertainty matrix $\boldsymbol{\sigma}$, whose $ij$ entry is the standard deviation of the $i$-th gene and $j$-th sample of $\mathbf{D}$, and a list of gene sets $\mathcal{G}_k$, where $k$ indexes the sets. First, CoGAPS implements GAPS to infer common underlying patterns in gene expression across columns of $\mathbf{D}$ by factorization into a pattern matrix ($\mathbf{P}$) and a corresponding amplitude matrix ($\mathbf{A}$). GAPS seeks $\mathbf{P}$ and $\mathbf{A}$ matrices whose product is from the distribution for $\mathbf{D}$, which is assumed normal. That is,

$$\mathbf{D}_{ij} = (\mathbf{AP})_{ij} + \epsilon_{ij}, \quad (1)$$

where $\epsilon_{ij}$ is independent, normal noise with mean zero and variance $\boldsymbol{\sigma}^2_{ij}$. Estimates must be provided for $\boldsymbol{\sigma}$, which we have typically obtained from

 

sample covariance of replicates (Bidaut *et al.*, 2006; Ochs *et al.*, 2009). The rows of **P** form a set of non-orthogonal basis vectors that describe the patterns of coexpression behavior across the samples in the columns of **D**. The rows of **A** quantify the amount of the behavior of a gene that is explained by each of the patterns (the rows of **P**). The number of rows in **P** sets the number of patterns that GAPS will infer. GAPS presently constrains the entries in **A** and **P** to be non-negative. Even so, the **A** and **P** matrices are not mathematically uniquely determined independent of prior information.

As noted in the Section 1, MCMC techniques recover BPs better than other factorization techniques. The Kossenkov and Ochs (2009) study found that inference of sparse matrices with atomic priors for MCMC inference (Sibisi and Skilling, 1997), such as in BD, has a particular advantage in retaining minimally varying patterns across samples, which define many BPs. These atomic priors also naturally enforce non-negativity and sparsity in the corresponding elements of **A** and **P**. Therefore, GAPS is implemented in GAPS-JAGS by incorporating the model in Equation (1) with an atomic prior in JAGS.

When running GAPS-JAGS, the user specifies a hyperparameter for the expected number of atoms ($\alpha_A$ and $\alpha_P$ for matrices **A** and **P**, respectively) and a parameter for the number of patterns ($n_p$) that provides the dimensionality required to reproduce **D**. The $\alpha$ parameters represent the sparsity of **A** and **P** and have default values of 1%. While the algorithm is insensitive to small changes in these parameters, order of magnitude changes will significantly alter the estimated **A** and **P** matrices. At 1%, GAPS was found to retain the sparsity of our previous successful MCMC studies, and we recommend this value for most applications. The appropriate number of patterns is data dependent and typically unknown. Dimensionality estimation prior to MCMC sampling can be obtained from new techniques (Leek, 2010) or by trying multiple matrix factorizations of different $n_p$ (Bidaut *et al.*, 2006). While there is no guarantee that the **A** and **P** matrices are uniquely identifiable, we have found in practice that a unique solution within uncertainty estimates from the sampling is typical for MCMC microarray analysis. However, we recommend multiple MCMC simulations to reduce the probability of finding a local maximum in the posterior distribution.

In order to infer activity of a BP, CoGAPS estimates the probability that genes in a set are overrepresented in a pattern from a statistic based on the average Z-score for all $A_{sp}$ for $s \in$ GS (Ochs *et al.*, 2009). This score can be used to rank sets and a frequentist interpretation is provided through permutation of the gene labels on a pattern-by-pattern basis.

## 3 IMPLEMENTATION

The software is run through the CoGAPS R package. The central R function for the CoGAPS algorithm inputs files containing **D** and $\boldsymbol{\sigma}$, as well as sparsity parameters $\alpha_A$ and $\alpha_P$, the number of patterns, and gene sets in a format specified in the Users Manual. This function also allows users to specify a folder and prefix for output files summarizing the statistics computed by CoGAPS. The Users Manual also describes additional runtime options.

CoGAPS first factors the matrix of microarray data through a C++ package called GAPS-JAGS (Plummer, 2003) as described in Section 2. This package is an extension of JAGS (version 1.0.3) that includes a module for GAPS, and this is required by the CoGAPS R package. With $N \gg M$ and typical data size of thousands of genes by hundreds of samples, the computational cost is $O(N \log N)$ and memory requirements are moderate (see Supplemental Material for specifics). The MCMC chain for **A** and **P** are output as temporary files, and the summarized estimates of the statistics for **A** and **P** are retained in output files. Optionally, CoGAPS also plots the identified patterns and creates a heatmap for the corresponding **A** intensities.

CoGAPS computes Z-scores and *P*-values for each GS in each pattern. An 'activity' is also calculated that rescales the *p*-value estimates from $-1$ to $+1$, suitable for pictorial representation of TF activity. These statistics are output into three separate files.

We have developed open-source C++ software, GAPS-JAGS, with an R interface, CoGAPS, for inferring GS enrichment from transcriptomic data. When applying GSs defined by TFs, the DESIDE algorithm used this approach to infer the changes in cell signaling during treatment of gastrointestinal tumors. We note that any high-throughput biological data representable as a quantitative matrix of biomolecule measurements across samples is amenable to this approach, if these biomolecules can be linked in GSs.

*Funding*: National Library of Medicine (Grant LM009382); National Science Foundation (Grant 0342111).

*Conflicts of Interest*: none declared.

## REFERENCES

Bidaut,G. *et al.* (2006) Determination of strongly overlapping signaling activity from microarray data. *BMC Bioinformatics*, **7**, 99.

Carvalho,C. *et al.* (2008) High-dimensional sparse factor modelling: applications in gene expression genomics. *J. Am. Stat. Assoc.*, **103**, 1438–1456.

Draghici,S. *et al.* (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.*, **31**, 3775–3781.

Goeman,J. and Buhlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Kossenkov,A. and Ochs,M. (2009) Matrix factorization for recovery of biological processes from microarray data. *Methods Enzymol.*, **467**, 59–77.

Leek,J. (2010) Asymptotic conditional singular value decomposition for high-dimensional genomic data. *Biometrics*, in press.

Lee,D. and Seung,H. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.

Ochs,M. *et al.* (1999) A new method for spectral decomposition using a bilinear bayesian approach. *J. Magn. Reson.*, **137**, 161–176.

Ochs,M. *et al.* (2009) Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. *Cancer Res.*, **69**, 9125–9132.

Plummer,M. (2003) JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In Hornik,K. *et al.* (eds) *Proceedings of the 3rd Internation Workshop on Distributed Statistical Computing*, Vienna, Austria.

Sibisi,S. and Skilling,J. (1997) Prior distributions on measure space. *J. Royal Stat. Soc. B*, **59**, 217–235.

Tavazoie,S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.