Systems biology

Analyzing gene perturbation screens with nested effects models in R and bioconductor

Holger Fröhlich¹, Tim Beißbarth¹, Achim Tresch², Dennis Kostka³, Juby Jacob⁴, Rainer Spang⁴,∗ and F. Markowetz⁵

¹German Cancer Research Center (DKFZ), INF 580, 69120 Heidelberg, Germany, ²Gene Center, Ludwig-Maximilian-Universität München, München, Germany, ³Genome Center and Department of Statistics, University of California Davis, Davis, CA 95616, USA, ⁴Computational Diagnostics Group, Institute of Functional Genomics, University of Regensburg, 93053 Regensburg and ⁵Lewis-Sigler Institute for Integrative Genomics and Department of Computer Science, Princeton University, Princeton NJ 08544, USA

Received on June 25, 2008; revised on August 12, 2008; accepted on August 17, 2008

Advance Access publication August 21, 2008

© The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

ABSTRACT

Summary: Nested effects models (NEMs) are a class of probabilistic models introduced to analyze the effects of gene perturbation screens visible in high-dimensional phenotypes like microarrays or cell morphology. NEMs reverse engineer upstream/downstream relations of cellular signaling cascades. NEMs take as input a set of candidate pathway genes and phenotypic profiles of perturbing these genes. NEMs return a pathway structure explaining the observed perturbation effects. Here, we describe the package nem, an open-source software to efficiently infer NEMs from data. Our software implements several search algorithms for model fitting and is applicable to a wide range of different data types and representations. The methods we present summarize the current state-of-the-art in NEMs.

Availability: Our software is written in the R language and freely available via the Bioconductor project at http://www.bioconductor.org.

Contact: rainer.spang@klinik.uni-regensburg.de

1 INTRODUCTION

The analysis of large-scale and high-dimensional phenotyping screens is moving to the center stage of computational systems biology as more and better experimental systems get established in model organisms. Nested effects models (NEM) are a class of models introduced to analyze the effects of gene perturbation screens visible in high-dimensional phenotypes like microarrays or cell morphology. NEMs achieve two goals: (i) to reveal clusters of genes with highly similar phenotypic profiles and (ii) to order (clusters of) genes according to subset relationships between phenotypes. These subset relationships show which genes contribute to global processes in the cell and which genes are only responsible for sub-processes. The NEM structure helps to understand signal flow and internal organization in a cell.

NEMs offer complementary information to traditional graphical models including correlation graphs, Bayesian networks and Gaussian graphical models (Markowetz and Spang, 2007). Thus, they are relevant for theoretical researchers developing methods in systems biology. In addition, a wide range of applications shows the broad impact of NEMs on both molecular biology and medicine: NEMs were successfully applied to data on immune response in Drosophila melanogaster (Markowetz et al., 2005), to the transcription factor network in Saccharomyces cerevisiae (Markowetz et al., 2007), and to the ER-α pathway in human breast cancer cells (Fröhlich et al., 2007, 2008).

2 NEM IMPLEMENTATION

NEMs are two-layered graph models. The first layer consists of a directed graph containing the genes that were experimentally perturbed. The second layer consists of the effects observed in high-dimensional phenotypes. Each node in the second layer is considered to be a specific reporter for the activity of a single gene in the first layer. Current NEM formulations differ in the constraints they pose on the NEM graph in the first layer and on the probabilistic model they assume for effect nodes in the second layer. All current types of NEMs are implemented in the R package nem, which is available from the Bioconductor project (Gentleman et al., 2004; R Development Core Team, 2007).

NEM formulations and inference: A first NEM formulation restricts the NEM graph to be transitively closed. The probabilistic model for effects is either Bernoulli (Markowetz et al., 2005, 2007) or a mixture distribution (Fröhlich et al., 2007, 2008). A second NEM formulation (Tresch and Markowetz, 2008) relaxes the constraints on the NEM graph and allows graphs that are not transitively closed. For each model formulation, the user can choose between different search methods for model inference. Exhaustive enumeration (Markowetz et al., 2005) is feasible for up to eight perturbed genes. For bigger pathways the package provides greedy search heuristics and divide-and-conquer like approaches that divide the graph into smaller units, use exhaustive enumeration for each subgraph and then reassemble the complete model. The division into subgraphs can either be into all pairs or triples of nodes
The core function of the package is `nem()`. Its output is a list with components containing the highest scoring NEM graph, the marginal likelihoods of all scored models, as well as the estimated positions of effect reporters in the NEM graph. In this example model, search is done by exhaustive enumeration. Specifying inference as ‘pairwise’ or ‘triple’ uses the search heuristics of Markowetz et al. (2007), while ‘nem.greedy’ and ‘ModuleNetwork’ employ the methods of Fröhlich et al. (2007, 2008). These methods extend model search to hundreds of perturbed genes.

The function `nem()` is applicable to a wide range of data representations. Instead of discretized data, the user can supply it with log-ratios or P-values for seeing an effect. For the example dataset a matrix of precomputed log-ratios (`BoutrosRNAiLogFC`) and P-value densities (`BoutrosRNAiDens`) is contained in the package. All data representations can be used in a MAP estimate (type="CONTmLLMAP") or in a model marginalizing over effect positions (type="CONTmLLBayes"). Additional feature selection to select only informative effect reporters (selEGenes=TRUE) is implemented for all data types. An example application is visualized in Figure 1 by executing:

R> res2 <- nem(BoutrosRNAiDens, + type="CONTmLLBayes", selEGenes=TRUE)
R> plot(res2,D=BoutrosRNAiLogFC)

Funding: National Genome Research Network (NGFN) of the German Federal Ministry of Education and Research (BMBF) through the platforms SMP Bioinformatics (01GR0450 to H.F., A.T. and T.B.) and EP-S19T04 to H.F., A.T. and T.B. National Institutes of Health (grant R01 GM071966); National Science Foundation (NSF) (grant IIS-0513552) (Princeton University); National Institute of General Medical Sciences (NIGMS) Center of Excellence (grant P50 GM071508); NSF (grant DBI-0546275).

Conflict of Interest: none declared.

REFERENCES


