Systems biology

Identifying differentially expressed subnetworks with MMG
Josselin Noirel1,*, Guido Sanguinetti2 and Phillip C. Wright1

1Biological and Environmental Systems Group, Department of Chemical and Process Engineering, University 10 of Sheffield, Mappin Street, Sheffield, S1 3JD and 2Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Road, Sheffield, S1 4DP, UK

Received and revised on May 26, 2008; accepted on September 17, 2008
Advance Access publication September 25, 2008
Associate Editor: Trey Ideker

ABSTRACT
Background: Mixture model on graphs (MMG) is a probabilistic model that integrates network topology with (gene, protein) expression data to predict the regulation state of genes and proteins. It is remarkably robust to missing data, a feature particularly important for its use in quantitative proteomics. A new implementation in C and interfaced with R makes MMG extremely fast and easy to use and to extend.
Availability: The original implementation (Matlab) is still available from http://www.dcs.shef.ac.uk/~guido/; the new implementation is available from http://wrightlab.group.shef.ac.uk/people_noirel.htm, from CRAN, and has been submitted to BioConductor, http://www.bioconductor.org/.
Contact: j.noirel@sheffield.ac.uk

1 INTRODUCTION
Detecting differential expression is a task as old as molecular biology itself. An idea that is rapidly emerging in the bioinformatics community is the use of network or pathway information to aid the task, with as many as six articles published independently in the last year (Li and Li, 2008; Noirel et al., 2008; Rapaport et al., 2007; Sanguinetti et al., 2008; Wei and Li, 2007; Wei and Pan, 2008). The advantages of taking a network perspective are 2-fold: first, the network information could capture subtler changes than a gene-by-gene threshold-based method. For example, regulatory proteins for which a small change triggers larger downstream changes are captured by network-based methods. Second, by taking a global perspective, they allow the identification of biologically meaningful subnetworks which coordinate the cellular response to the conditions under study. Both these constitute a real boon when analyzing high-throughput proteomic data: the large amount of missing data makes threshold-based methods virtually unviable. For instance, only 15% of the metabolic network of Nostoc sp. PCC 7120 was readily identified and quantified in previous studies (Noirel et al., 2008; Stensjö et al., 2007), as described by KEGG (Kanehisa and Goto, 2000).

We recently proposed mixture model on graphs (MMG), a fully probabilistic model that integrates structural information encoded by network topology into the analysis of high-throughput proteomic data (Sanguinetti et al., 2008). The data are modelled using a mixture distribution that represents the regulation status of the proteins: up-regulated, down-regulated or unchanged. The network structure is taken into account through prior conditional probabilities relating the probability of a node’s belonging to a given class to the classes of its neighbours. Specifically, the prior probability of belonging to a certain class c is proportional to the number of neighbours in the same class:

\[ p(c_j | c_1 \cdots c_N) \propto \alpha + \sum_{i \in N_j} w_{ij} \]  

choosing suitable prior distributions for the mixture parameters (Gan et al., 2007), one could then derive exact expressions for the conditional posteriors, which enabled an efficient Gibbs sampling strategy (for details see Sanguinetti et al., 2008, Supplementary Materials).

MMG was originally developed with metabolic networks and proteomic datasets in mind, but is general enough to address a wider range of problems, owing to (i) the possibility to weight the network’s edges in order to increase or to decrease specifically the trustworthiness of a particular connection (w in Equation 1), (ii) the robustness against missing data. We demonstrated MMG’s usefulness through extensive benchmarking on synthetic data and on proteomic data describing nitrogen fixation in a cyanobacterium (Stensjö et al., 2007).

2 IMPLEMENTATION
The high level of interest in network-based approaches to differential expression is unfortunately not matched by an availability of software tools. As far as we are aware, only the original Matlab implementation of MMG has been made publicly available. A new R implementation is now available. The benefits of this implementation is a better integration with already existing statistical software tools. As far as we are aware, only the original Matlab implementation of MMG has been made publicly available. The benefits of this implementation is a better integration with already existing statistical and graphical tools provided by R with an enhanced usability and user-friendliness, the documentation and of course its open access nature. Technically, most of the work is carried through by compiled code written in C. However, the ‘bells and whistles’ of R allows one to further process the numerical results, as well as facilitate the comparison of different setups or different datasets. Furthermore, R allows for good result visualization.

The user provides a file containing the adjacency lists for the metabolic network. The function MMG.compute runs a Gibbs sampler to estimate the posterior probabilities of each node to be
The MMG library is loaded within a typical R session using the library command. The Gibbs sampler is run through MMG.compute with the parameters $\sigma = 0.3$ and $\alpha = 1.0$ (Sanguinetti et al., 2008). The Gibbs sampler returns a list that is useful to process further the data. (B) The results of the Gibbs sampler can be utilized to select up-regulated nodes, for instance, in order to investigate nitrogen fixation in the cyanobacteria Nostoc sp. PCC 7120 (Stensjö et al., 2007); thereafter, this selection can be used to generate a subgraph of the metabolic network of interest and represented using the Graphviz tools. (C) The same analysis as in (B) can be applied to toy models in order to fine tune the role of the parameters $\sigma$ and $\alpha$ (Sanguinetti et al., 2008). (D–F). The outcome of the Gibbs sampler allows the user to check that the chains mix well (D) or to represent the posterior probabilities corresponding to the parameters $\lambda_-$ and $\lambda_+$ (Sanguinetti et al., 2008), whose distributions indicate what is the level of down- and up-regulation. ‘down-regulated’, ‘unchanged’ or ‘up-regulated’. The output of this function includes a printout of the statistics of the model.

The posterior probabilities can then be used by MMG.cut.graph to cut the graph and produce a list of nodes which behave consistently. We look for clusters of adjacent nodes that belong to a given class, with a high likelihood, and identify the different connected components using a standard depth-first search algorithm (the network is considered unweighted and undirected). Several methods are available to find nodes belonging to a given class.

Once the connected components are identified using the undirected version of the network, the directedness of the network may be restored and represented using the package RGraphViz. This is done using the function MMG.make.dot, which outputs a DOT file. The probabilities of being up-regulated, unchanged, and down-regulated can handily be mapped onto the RGB colour scheme. A typical workflow is presented in Figure 1, from the analysis of the data up to the production of the graphics.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to R. Sochon (University of Sheffield) and W. Venables (CSIRO, Australia).

Funding: EPSRC (grants EP/E036252/1 and GR/s84347/01 to J.N and P.C.W.); BiomodularH2 EU-FP6 project NEST (contract number 043340 to J.N. and P.C.W.).

Conflict of Interest: none declared.

REFERENCES


