BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 12 2010, pages 1576–1577
doi:10.1093/bioinformatics/btq168

Systems biology
Advance Access publication April 21, 2010

NetRaVE: constructing dependency networks using sparse linear regression

A. Phatak[1], H. Kiiveri[1,*], L.H. Clemmensen[2] and W.J. Wilson[3]
[1]CSIRO Mathematical & Information Sciences, Private Bag 5, Wembley, WA, Australia, [2]Department of Informatics and Mathematical Modelling, Technical University of Denmark, Building 321, DK-2800 Lyngby, Denmark and [3]CSIRO Mathematical & Information Sciences, Locked Bag 17, North Ryde, NSW, Australia

Associate Editor: Trey Ideker

ABSTRACT

Summary: NetRaVE is a small suite of R functions for generating dependency networks using sparse regression methods. Such networks provide an alternative to interpreting 'top *n* lists' of genes arising out of an analysis of microarray data, and they provide a means of organizing and visualizing the resulting information in a manner that may suggest relationships between genes.

Availability: NetRaVE is freely available for academic use and has been tested in R 2.10.1 under Windows XP, Linux and Mac OS X.

Contact: harri.kiiveri@csiro.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 7, 2009; revised on March 24, 2010; accepted on April 13, 2010

1 INTRODUCTION

The advent of microarray technology has spawned a plethora of statistical and machine learning methods to analyze the resulting expression data. Their ultimate aim is to guide the researcher to some understanding of how the observed patterns of gene expression are related to the phenotype being studied and to suggest hypotheses that merit further exploration.

In this note, we outline NetRaVE, a two-step approach to analyzing microarray data that combines a fast, sparse variable selection method, known as GeneRaVE (Kiiveri, 2008), with dependency networks, an alternative to Bayesian networks developed by Heckerman *et al.* (2000). NetRaVE is a descriptive tool, whose use is facilitated by a small suite of R (R Development Core Team, 2008) functions that interface with the existing libraries to visualize the resulting networks. Its output is a local dependency network that encapsulates the results of the data analysis in a form that is easy to visualize and provides users with additional information about local relationships between gene transcripts. These local relationships can then be used to generate hypotheses that can lead to further experimentation.

2 APPROACH

The basic procedure for generating local dependency networks is straightforward and consists of two steps. First, we use GeneRaVE to select genes that 'best' discriminate between subtypes present in the data. In the example described below, discrimination is between two groups of individuals—smokers and never smokers—and we identify a small subset of genes that discriminates between these two groups by using the sparse logistic regression. Second, we use these genes as 'seed' genes for building the local dependency networks. To do so, the individual seed genes become the response variables in separate sparse linear regressions in which all other genes are potential explanatory variables. Again we use GeneRaVE, so that the resulting regression model is sparse and contains only a few other explanatory genes. After this first round of regressions, the network is defined by edges between the response (seed) genes and their selected regressor (explanatory or predictor) genes.

To expand the network beyond the first round, we simply use the explanatory genes from the first round as response variables in regressions using all other genes as potential regressors. Though it is possible to repeat this procedure several times, the resulting large network will be difficult to visualize and interpret. In practice, it is more useful to construct only *local networks*, which are typically obtained after two or three rounds. Note that dependency networks can be constructed around *any* genes of interest within a dataset. Furthermore, the R function in NetRaVE that is used for constructing networks is written so that the user can easily write his/her own method for generating conditional probability distributions using alternative regression methods. Further details about NetRaVE and dependency networks appear in the Supplementary Material, along with a larger version of Figure 1 and the R script used to generate it.

3 EXAMPLE

We illustrate NetRaVE on the dataset collected by Spira *et al.* (2004). It consists of 22 283 gene expression measurements (Affymetrix HG-U133A array) of epithelial cells taken from the tracheas of 57 individuals in two groups: smokers (34) and never smokers (23). The raw data were pre-processed by RMA background correction, followed by quantile normalization and finally summarization by robustly fitting a multi-chip linear model (Bolstad *et al.*, 2003). GeneRaVE selects only three genes separating the two classes—*CEACAM5*, *ALDH3A1* and *CYP1B1*. Plug-in classification was perfect, and 10-fold cross-validation yielded only 2 and 3 misclassifications out of 34 smokers and 23 never smokers, respectively. Then, prior to constructing a network, we combined data from probesets mapped to the same gene by calculating their median value for each individual.

*To whom correspondence should be addressed.

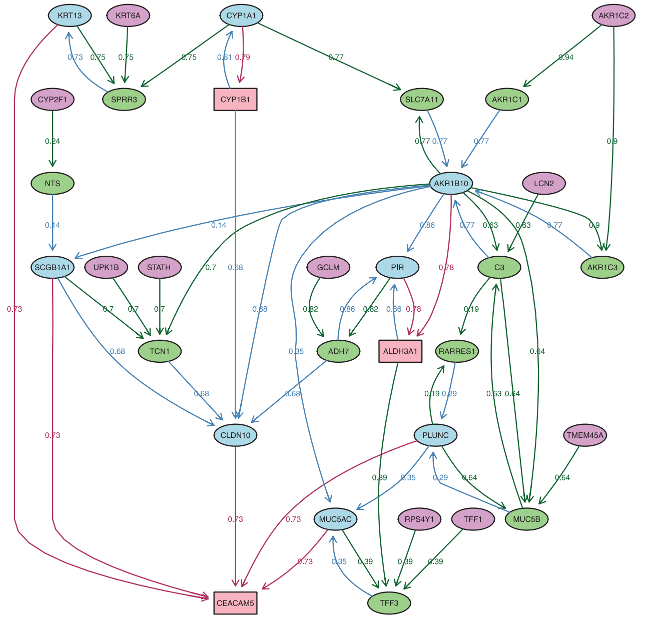

Fig. 1. Dependency network generated from the data of Spira *et al.* (2004) to a depth of three around the genes *CEACAM5*, *ALDH3A1* and *CYP1B1*. See text for explanation of colours and annotations.

Using NetRaVE and sparse linear regression, we can generate a dependency network as illustrated in Figure 1. The seed genes (*CEACAM5*, *ALDH3A1* and *CYP1B1*) are the light red nodes, and the directed edges pointing towards them from their regressors (light blue nodes) in the first round of regressions (depth 1) are annotated with the cross-validated R^2 of the sparse regression. For example, *ALDH3A1* has two regressors (*PIR* and *AKR1B10*), with a cross-validated R^2 of 0.78, whereas *CEACAM5* has five, including *MUC5AC*, *UPK1B*, *SCGB1A1*, *CLDN10* and *KRT13* (R^2=0.73). The regressors, or neighbours, of genes in depth 1 are coloured green, and, in turn, their neighbours are coloured purple. Note that genes that appear at early depths can also be explanatory of genes at later depths [e.g. the directed edge between *ALDH3A1* (depth 0) and *TFF3* (depth 2)].

4 DISCUSSION

There are a few options for generating and visualizing dependency networks that are discussed in the documentation accompanying NetRaVE and in the Supplementary Material. In our experience in analyzing this and other datasets, we have found many of the genes in a local network are also found in top *n* lists of differentially expressed genes obtained by, for example, modified *t*-tests. Here, 12 of the 32 genes in Figure 1 appear in the top-97 list of Spira *et al.* (2004). For example, they found that *ALDH3A1* (depth 0) and *CYP1B1* (depth 0) were highly up-regulated in smokers. However, the network imposes an additional structure on those lists that provides the researcher with a means of organizing and visualizing that information. The dependency network in Figure 1 shows other genes that were not comparatively highly up- or down-regulated, yet that are related to the effects of cigarette smoking. For example, *AKR1B10* is a potential diagnostic marker of non-small cell lung carcinoma (Penning, 2005), and together with other aldo–keto reductase genes (*AKR1C1*, *AKR1C2* and *AKR1C3*) participates in the metabolism of xenobiotics; the up-regulation of *PIR*, which appears at depth 1, represents one mechanism by which cigarette smoke induces apoptosis in the airway epithelium (Gelbman *et al.*, 2007); and many other such examples. In molecular toxicology research, it is well documented that the transcription factor aryl hydrocarbon receptor (*AhR*, also called the dioxin receptor) enhances gene transcription upon activation by polycyclic hydrocarbons such as those found in cigarette smoke (reviewed in Kitamura and Kasai, 2007). It is tempting to speculate that some gene linkages within the network in Figure 1 are due to coordinated gene expression in response to activation of *AhR* by cigarette smoke toxins, but the truth is inevitably more complex. Consider the effect of smoking on epithelial inflamation. Linkages between some genes such as *MUC5AC* (mucin super family member) and the *CEACAM* adhesion molecules could also be due to effects of neutrophils (Fischer and Voynow, 2007). Whether these mechanisms are behind the expression patterns observed in the dataset by NetRaVE or not, the ability to identify linkages between groups of genes provides an alternative to unstructured data exploration.

We emphasize here that, like all influence networks, local dependency networks should not be viewed as a rigorous means of uncovering true causal relationships among genes; instead, they should be viewed as a very useful tool for organizing complex information in a manner that is easy to visualize. The relationships contained in them can then be used, along with related information and background knowledge about the biological system being studied, to guide further experimentation.

Conflict of Interest: none declared.

REFERENCES

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Fischer,B.M. and Voynow,J.A. (2002) Neutrophil elastase induces MUC5AC gene expression in airway epithelium via a pathway involving reactive oxygen species. *Am. J. Respir. Cell Mol. Biol.*, **26**, 447–452.

Gelbman,B. *et al.* (2007) Upregulation of pirin expression by chronic cigarette smoking is associated with bronchial epithelial cell apoptosis. *Respir. Res.*, **8**, 10.

Heckerman,D. *et al.* (2000) Dependency networks for inference, collaborative filtering, and data visualization. *J. Mach. Learn. Res.*, **1**, 49–75.

Kiiveri,H.T. (2008) A general approach to simultaneous model fitting and variable elimination in response models for biological data with many more variables than observations. *BMC Bioinformatics*, **9**, 195.

Kitamura,M. and Kasai,A. (2007) Cigarette smoke as a trigger for the dioxin receptor-mediated signaling pathway. *Cancer Lett.*, **252**, 184–194.

Penning,T.M. (2005) AKR1B10: A new diagnostic marker of non-small cell lung carcinoma in smokers. *Clin. Cancer Res.*, **11**, 1687–1690.

R Development Core Team (2008) *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna, Austria.

Spira,A. *et al.* (2004) Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc. Natl Acad. Sci. USA*, **101**, 10143–10148.