Gene expression

netClass: an R-package for network based, integrative biomarker signature discovery
Yupeng Cun* and Holger Fröhlich
Bonn-Aachen International Center for IT (B-IT), University of Bonn, Dahlmannstr. 2, 53113 Bonn, Germany

ABSTRACT
In the past years, there has been a growing interest in methods that incorporate network information into classification algorithms for biomarker signature discovery in personalized medicine. The general hope is that this way the typical low reproducibility of signatures, together with the difficulty to link them to biological knowledge, can be addressed. Complementary to these efforts, there is an increasing interest in integrating different data entities (e.g. gene and miRNA expressions) into comprehensive models. To our knowledge, R-package netClass is the first software that addresses both, network and data integration. Besides several published approaches for network integration, it specifically contains our recently published stSVM method, which allows for additional integration of gene and miRNA expression data into one predictive classifier.

Availability: netClass is available on http://sourceforge.net/p/netclassr and CRAN (http://cran.r-project.org).
Contact: yupeng.cun@gmail.com

Received on September 17, 2013; revised on December 3, 2013; accepted on January 11, 2014

1 INTRODUCTION
One of the major goals of personalized medicine is to identify molecular biomarkers that reliably predict a patient’s response to therapy to avoid ineffective treatment and to reduce drug side effects and associated costs. For that purpose, prognostic and diagnostic biomarker signatures have been derived from omics data in numerous publications for various disease entities.

To construct biomarker signatures typically machine learning algorithms are used, such as SVMs (Cortes and Vapnik, 1995) and RandomForests (Breiman, 2001). The challenge is the extreme high dimensionality of omics data coupled with a relatively small sample size, which imposes a major need for careful feature selection. However, during the past years it has become more and more clear that classical feature selection methods, such as t-test based filtering, frequently lead to signatures that are neither reproducible on a different dataset (Ein-Dor et al., 2005) nor biologically interpretable (Gönen, 2009). Hence, there has been a growing interest to incorporate prior information on protein–protein interactions, pathways or Gene Ontology annotation into feature selection algorithms (see Cun and Fröhlich, 2012a for an extensive review). It has been shown that such approaches can at least increase the feature selection stability and facilitate the biological interpretation of signatures (Cun and Fröhlich, 2012b).

In this article, we present our R-package netClass, which implements five network-based gene selection methods. Although there is a rich literature on general data integration, netClass is, to our knowledge, the first software that allows for integrating miRNA and mRNA expression data together with protein–protein interactions and miRNA-target gene information (Cun and Fröhlich, 2013) into one predictive model. netClass thus complements the functionality of our earlier software package pathClass (Johannes et al., 2011). It is worth emphasizing that netClass focuses on classification algorithms only. A software package that is more tailored to Cox regression is CoxBoost (Binder and Schumacher, 2009).

2 PACKAGE OVERVIEW
netClass currently implements five network-based gene selection methods, which have turned out to be successful in the literature: (i) Average expression profile of pathways (Guo et al., 2005); (ii) Pathway activity classification (Lee et al., 2008); (iii) Classification based on differential expression of hub genes and correlated partners (Taylor et al., 2009); (iv) Filtering of genes according to a modified Google PageRank algorithm (Winter et al., 2012); (v) Kernel-based smoothing of t-statistics over a network structure (Cun and Fröhlich, 2013). Specifically, the latter approach also allows for integrating miRNA and mRNA expression data. Neither of the five above-mentioned methods has been implemented in pathClass, which mainly focuses on the SVM-RFE algorithm and variants thereof (Johannes et al., 2010). Hence, netClass and pathClass complement each other.

Pathway activity classification is the only non-SVM based classification approach in netClass: it uses logistic regression (Lee et al., 2008). All the other algorithms internally use (linear) SVM classification. netClass enables to tune the soft margin parameter automatically in a computationally efficient manner using the span rule, which provides a theoretical upper bound on the leave-one-out cross-validation error and can be calculated from training data only (Chapelle and Vapnik, 1999). Furthermore, to evaluate the prediction performance of classification algorithms, in netClass feature selection and soft margin parameter tuning are embedded into a repeated k-fold cross-validation scheme. Cross-validation can be performed via user-friendly interface functions and allows for parallel computing.

2.1 Data and network integration via kernel-based smoothing of t-statistics
A specific feature of netClass is the implementation of our recently proposed stSVM algorithm, which allows for joint integration of network information, together with miRNA and mRNA expression data
3 CONCLUSION

netClass is an R-package that allows for network and data integration for biomarker signature discovery. It includes several published approaches for incorporating network information into gene selection. Moreover, netClass contains our recently published stSVM algorithm, which allows for additional integration of miRNA and mRNA expression data. All implemented methods can perform repeated cross-validation to estimate the prediction performance. Moreover, integration of igraph facilitates the follow-up analysis of selected features via graph algorithms and plotting functions. In summary, we believe that netClass provides a useful tool for biomarker signature discovery in personalized medicine.

Funding: Y.C. was supported by the state of NRW via the B-IT research school. H.F. is a member of the excellence cluster ImmunoSensation.

Conflict of Interest: none declared.

REFERENCES