Bioinformatics Advance Access published January 17, 2014

netClass: An R-package for network based, integrative biomarker signature discovery

Yupeng Cun [1]* and Holger Fröhlich [1]

[1]Bonn-Aachen International Center for IT (B-IT), University of Bonn, Dahlmannstr. 2, 53113 Bonn, Germany

Associate Editor: Dr. Janet Kelso

ABSTRACT

In the last 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 expression) 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

1 INTRODUCTION

One of the major goals of personalized medicine is to identify molecular biomarkers that reliably predict a patient's response to therapy in order 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 employed, 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 last 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 reprodicible on a different data set (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 (GO) 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. While 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 e.g. *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: 1) Average expression profile of pathways (Guo *et al.*, 2005); 2) Pathway activity classification (Lee *et al.*, 2008); 3) Classification based on differential expression of hub genes and correlated partners (Taylor *et al.*, 2009); 4) Filtering of genes according to a modified Google PageRank algorithm (Winter *et al.*, 2012); 5) 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 have 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 (Cun and Fröhlich, 2013).

*to whom correspondence should be addressed

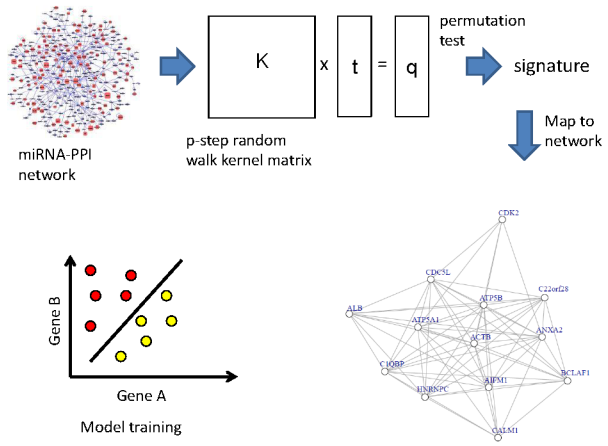

Fig. 1. Workflow of **stSVM**: Marginal statistics for features in each -omics dataset are computed and smoothed over the structure of a joined miRNA-PPI network. After re-ranking a permutation test selects the most relevant features and trains a SVM model. The obtained signature can be visualized as a network.

The basic idea behind *stSVM* is to smooth a feature-wise marginal statistic (like the commonly used t-statistic) over the structure of a joint protein-protein and miRNA-target gene interaction graph. For this purpose a random walk kernel is employed (Gao *et al.*, 2009). A permutation test is used to select features in a highly consistent manner, and then these features are employed for subsequent SVM training. In our paper we demonstrated the utility of this approach on four datasets from different tumor entities and specifically showed that integration of miRNA and mRNA expression could enhance the prediction power for prostate cancer prognosis (Cun and Fröhlich, 2013).

2.2 Integration of *igraph*

netClass facilitaties the post-hoc analysis of obtained feature sets by integrating the R-package *igraph* (Csardi and Nepusz, 2006). Algorithms incorporating network structures return the connected sub-graph(s) between selected features. This enables the full functionality of graph algorithms and plotting routines (Figure 1). In this context specifically Steiner tree methods as e.g. implemented in our package *SteinerNet* may provide a useful additional tool (Sadeghi and Fröhlich, 2013).

2.3 Example Usage

To illustrate the use of *netClass* we show an example for running **stSVM** on a small sample dataset. First we get the sample data $expr$ with gene expression matrix $genes$, miRNA expression matrix $miRNA$ and class labels y. The adjacency matrix for the network is given in $ad.matrix$. We then train **stSVM** on the whole dataset and plot the sub-graph induced by selected features:

```
>library(netClass)
>data(expr)
>data(EN2SY)
>data(ad.matrix)
>dk<-calc.diffusionKernelp(L=ad.matrix,p=2,a=1)
>st=train.stsvm(x=cbind(expr$genes,expr$miRNA),
y=expr$y, Gsub=ad.matrix, dk=dk, EN2SY=EN2SY)
>plot(st$trained$graph)
```

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: YC was supported by the state of NRW via the B-IT research school. HF is a member of the excellence cluster ImmunoSensation.

REFERENCES

Binder, H. and Schumacher, M. (2009). Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC Bioinformatics*, **10**, 18.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.

Chapelle, O. and Vapnik, V. (1999). Model selection for support vector machines. In *NIPS*, pages 230–236.

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, **20**, 273 – 297.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, **Complex Systems**, 1695.

Cun, Y. and Fröhlich, H. (2012a). Biomarker gene signature discovery integrating network knowledge. *Biology*, **1**(1), 5–17.

Cun, Y. and Fröhlich, H. (2012b). Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. *BMC Bioinformatics*, **13:69**.

Cun, Y. and Fröhlich, H. (2013). Network and data integration for biomarker signature discovery via network smoothed t-statistics. *PloS one*, **8**(9), e73074.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**(2), 171–178.

Gao, C., Dang, X., Chen, Y., and Wilkins, D. (2009). Graph ranking for exploratory gene data analysis. *BMC Bioinformatics*, **10 Suppl 11**, S19.

Gönen, M. (2009). Statistical aspects of gene signatures and molecular targets. *Gastrointestinal cancer research: GCR*, **3**(2 Supplement 1), S19.

Guo, Z., Zhang, T., Li, X., Wang, Q., Xu, J., Yu, H., Zhu, J., Wang, H., Wang, C., Topol, E. J., Wang, Q., and Rao, S. (2005). Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, **6**, 58.

Johannes, M., Brase, J., Fröhlich, H., Sültmann, H., and Beissbarth, T. (2010). Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics*, **26**(17), 2136 – 2144.

Johannes, M., Fröhlich, H., Sültmann, H., and Beißbarth, T. (2011). pathclass: an r-package for integration of pathway knowledge into support vector machines for biomarker discovery. *Bioinformatics*, **27**(10), 1442–1443.

Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., and Lee, D. (2008). Inferring pathway activity toward precise disease classification. *PLoS Comput Biol*, **4**(11), e1000217.

Sadeghi, A. and Fröhlich, H. (2013). Steiner tree methods for optimal sub-network identification: an empirical study. *BMC Bioinformatics*, **14**, 144.

Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J. L. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol*, **27**(2), 199–204.

Winter, C., Kristiansen, G., Kersting, S., Roy, J., Aust, D., Knösel, T., Rümmele, P., Jahnke, B., Hentrich, V., Rückert, F., *et al.* (2012). Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Computational Biology*, **8**(5), e1002511.