Regulatory transformation of gene expression for feature extraction in cancer

Rémy Nicolle[1,2], Mohamed Elati[1] and François Radvanyi[2]

[1] Institute of Systems and Synthetic Biology, CNRS UPS3509, University of Evry-Val-dEssonne EA4527, Genopole Campus 1, Genavenir 6, 5 rue Henri Desbruères, 91030 Evry, France
nicolle, elati@issb.genopole.fr

[2] Institut Curie, CNRS UMR 144, 26 rue dUlm, 75248 Paris, cedex 05, France
francois.radvanyi@curie.fr

Abstract

Classical approaches to analyze transcriptomic data usually produce average classification models that have very low reproducibility. In this work, genome wide gene expression is considered through the activity of large regulatory networks. We introduce a new measure of regulatory influence based on the variations of expression of genes in a large inferred regulatory network.

This methodology can be used to transform transcriptomic data into a smaller influence data set on which feature selection and classification models show similar predictive performance and increased stability and reproducibility, especially when comparing models trained on different datasets. The methodology was tested on two distinct bladder cancer data sets.

This work will be published in the 11th International Conference on Machine Learning and Application 2012.

Introduction

Gene expression profiling using microarray technologies usually results in data sets were the number of features is superior to the number of samples often by two numbers of magnitude. This problem of curse of dimensionality makes data processing and knowledge discovery an extremely difficult task. Feature selection and classification were widely used in large cohort of patients to produce Gene Expression Signatures (GES). Solely based on transcriptomic measurements, these methods aimed at extracting the smallest list of genes that is able to predict tumor phenotypes. Additionally to the predictive capabilities of GES, their size enables simple biological investigation which can provide information on the processes involved in tumorigenesis. In terms of predictability, GES hold acceptable performance [1]. However, they usually perform as well as a random list of genes [1] and GES of the same prognostic can be very different in terms of gene content [2]. This instability questions the biological relevance of the selected features and challenges the field for more reliable models.

Several methodologies attempted to overcome this problem by integrating prior knowledge for data analysis, often in the form of interaction networks. The basic idea is that if a set of biologically related genes has a homogeneous behavior, for instance co-regulated genes, they are more likely to reproduce it in other conditions than genes that were selected only because they have the highest variation (*e.g.* selected by t-test). For instance Chuang *et. al.* [3] searched in a protein interaction network for a sub-network in which the merged expression of genes is the most discriminant between two phenotypes. Yet transcriptomic data is a structured type of data where gene expression is governed by complex regulatory rules.

In this work we propose a computational system biology approach to analyze genome wide expression data through the influence of regulators on their targets. Inferring regulatory networks from gene expression data has been an active area of research to identify the machinery of the cell from experimental data [4][5]. The goal is to identify, for each gene expressed in a particular cellular context, the regulators (*e.g.* transcription factors) affecting its transcription. We distinguish three families of learning problem that have been recently studied. A first family of approach infers dynamical networks usually based on kinetic data to produce models such as dynamical Bayesian networks. Another way to consider the problem is to assume that some interactions are known and that the goal of learning is to build a classifier that defines if a pair of gene interacts or not [6]. Finally, unsupervised approaches use tools such as clustering, frequent item set mining or determine statistical dependencies between pairs of genes to infer networks from static data [7][8].

We propose to transform the data of gene expression variations into the activation, or repression, of regulatory programs by learning a large regulatory network in a reference data set and using it to decipher regulatory variations. We introduce a methodology to estimate the activity of sub regulatory networks in single samples and use this novel feature to a) reduce by approximately two orders of magnitude the number of features, b) increase the stability of predictive and feature selection models in transcriptomic data, c) enable simple interpretation of differential analysis with testable solutions.

Proposed approach

The initial goal of this work is to introduce a methodology to analyze genome wide expression data with a robust model in order to obtain more reliable comparisons and classifications in between data sets. In order to obtain a context specific regulatory network which describes the interactions between genes and transcription factors, a network inference method is applied to the gene expression data. We used LICORN [7], a data mining algorithm introduced by our team that can infer the targets of transcription factors from genome wide expression data. LICORN was shown to be suitable for cooperative regulation and to scale up to the complexity of mammalian transcriptional networks. Furthermore, LICORN was previously used by Birmelé *et. al.* [9] in yeast to infer a large regulatory network from gene expression data. From this inference step, the network specifies a set of activated and repressed genes for each regulator. Master regulators are defined as regulator genes that have a sufficiently large number of targets to estimate their influence.

As a mean to transform the transcriptomic data and to monitor the state of the global GRN, we propose to measure the influence of regulator genes on their targets in a given sample. The knowledge of the structure helps to analyze the data in a meaningful way. The inferred regulatory network is used to measure the influence of the master regulators on their target genes. This influence measure aims at representing the global impact of the activity of a master regulators on the transcriptome of a given condition. It is based on a t-test on the two sets of target gene of a given regulator in a given sample. Thus, when measured for each regulator in each sample, it produces a data set with the same number of samples but a reduced number of features representing the master regulators activity.

The computation of the influence for all regulators, results in an influence matrix of size m by k, containing the influence of k master regulators in the m samples, instead of a matrix of gene expression of size m by n with $k \ll n$. The minimum number of targets of a regulator for its influence to be computed is set to 10 activated and 10 repressed targets in order to obtain sufficient observations for the t-test as well as to filter insignificant regulators that could be have

been added by chance to the network. The number of samples is preserved since the influence is computed in each sample independently. Therefore, any method used in gene expression data, *i.e.* feature selection, classification or clustering can be applied in an identical way to the influence data.

Results

In order to assess the relevance of using regulatory features to analyze transcriptomic data, a comparison was done using classification and feature selection on two bladder cancer transcriptomic data set. The accuracy of the classification and the reproducibility of the associated feature selection method was compared between models trained on the gene expression and the influence data set.

The shrunken centroid method [10] was used to produce a gene signature and a classifier. It uses soft-thresholding to reduce, for each feature gene, the class centroids towards the overall centroid. Therefore, only features which have non-zero centroids in at least one class after thresholding are selected. The classification is done by assigning to each of the sample the class with the nearest shrunken centroids.

Using a cross validation procedure, we show that the accuracy of prediction is very similar between classifier trained on gene expression or regulatory influences. Although the influence data set has a reduced number of features, it contains enough information to accurately classify samples. Furthermore, classification models were trained on influence data sets built from random regulatory models which perfomed nearly as well as random guessing showing the importance of context specific network.

Although classification accuracy is an important feature, it has been previously shown that good accuracy can be obtained with a classifier based on randomly selected features in gene expression[1]. Thus, the ability of a method to select the same features after perturbing the data set, defined as the stability, is a determinant factor. The stability was estimated by measuring the average overlap of gene (or regulatory features) content between all pairs of signatures trained during the corss validation procedure. The Kuncheva index [11] was used to measure the overlap because it uses a correction based on the probability of selecting randomly identical features. The Kuncheva index corrects the bias due to the fact that their are far less regulators in the influence data set then genes in the measured transcriptome. We find that while the classification accuracy is nearly identical between the two views of the data (transcriptome and influence), the stability of the feature selection method in the influence data set is much higher. This shows the interest of using the regulatory network and its influence on gene expression data set for stable and robust knowledge discovery.

One of the most important drawback of current models in genome biology, is their lack of reliability when used in a different data set than the one it was trained on. To show that transforming the gene expression data into regulatory features does increase the reliability of classical analysis, we performed a comparison between the models trained on the one data set and tested on another one. First, we tested how well a classifier learned in one data set could be applied on the other data set in terms of accuracy. We found that although the regulatory features do not significantly improve the classification accuracy, a very small number of regulatory features are necessary to classify samples in another data set. However, in gene expression, a high number of features is needed to obtain an acceptable classification of samples.

Next, we evaluated whether the features selected on the two different bladder cancer data sets were comparable. This is the foremost problem in current methodologies[2], while they are able to produce models that have acceptable performance in classifying samples of independent

data set, they are unable to find models with overlapping features. Our results indicate that we can obtain a much higher reproducibility within models trained on the influence data sets, independently of the number of selected features.

Conclusion

We developed a method to reduce the number of features of gene expression profiling to a much smaller data set representing the variations of global regulatory activity. The resulting influence data set was used for feature selection which was significantly more reproducible when selecting influent regulators. The influence was also used for tumor classification and compared to the gene expression matrix. It globally performed as well and even better with a very small number of features when applied to different data sets.

Furthermore, the selection of tumor class specific influent regulator generates testable hypothesis and models for understanding tumorigenesis and tumor specificities by indicating master regulators that are accountable for large scale perturbations in the transcriptome.

Future work should focus on understanding regulatory divergence between tumor sub-types and further investigation of regulatory influences. For instance, by integrating regulator cooperativity and protein-protein interactions between regulator genes to analyze the influence of co-regulators and the variation in cooperativity.

References

[1] Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert. The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures. *PLoS ONE*, 6:e28210, 2011.

[2] Cheng Fan, Daniel S Oh, Lodewyk Wessels, Britta Weigelt, Dimitry S A Nuyten, Andrew B Nobel, Laura J van't Veer, and Charles M Perou. Concordance among gene-expression-based predictors for breast cancer. *N. Engl. J. Med.*, 355:560–569, 2006.

[3] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3, 2007.

[4] Mohamed Elati and Céline Rouveirol. *Unsupervised Learning for Gene Regulation Network Inference from Expression Data: A Review*, pages 955–978. John Wiley & Sons, Inc., 2011.

[5] W.P. Lee and W.S. Tzou. Computational methods for discovering gene networks from expression data. *Briefings in bioinformatics*, 10(4):408–423, 2009.

[6] J. Qian, J. Lin, N.M. Luscombe, H. Yu, and M. Gerstein. Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, 19(15):1917–1926, 2003.

[7] Mohamed Elati, Pierre Neuvial, Monique Bolotin-Fukuhara, Emmanuel Barillot, Francois Radvanyi, and Celine Rouveirol. LICORN: learning co-operative regulation networks from gene expression data. *Bioinformatics*, 23:2407–2414, 2007.

[8] A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7:S7, 2006.

[9] Etienne Birmelé, Mohamed Elati, Céline Rouveirol, and Christophe Ambroise. Identification of functional modules based on transcriptional regulation structure. *BMC Proc*, 2 Suppl 4:S4, 2008.

[10] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS*, 99:6567–6572, 2002.

[11] LI Kuncheva. A stability index for feature selection. *Proc. IASTED, Artificial Intelligence and Applications*, 2007.