Humboldt-Universität zu Berlin
Department of Computer Science
Knowledge Management in Bioinformatics
Exposé Bachelor Thesis

Evaluation of Transcription Factor Activity in Gene Regulatory Networks

by Christopher Schiefer

28th of November 2014

Introduction

Comprehending gene regulation has been a major endeavour in the field of bioinformatics for many years and still poses a difficult problem with a lot of uncertainties. In spite of improvements in the technologies used, e.g. high-throughput methods like ChIP-sequencing, and progress in the analysis of the driving regulators, large parts of many gene regulatory networks still remain to be unveiled (Röttger et al., 2012). This is especially true for the human genome which is most interesting for us as detailed knowledge on regulatory networks helps us to further elucidate physiological processes. Naturally a diverse set of methods for broadening our understanding of gene regulation has been published over the years. I therefore propose to investigate and discuss a recently presented approach for estimating the activity of transcription factors (TFs) (Schacht et al., 2014) as it promises further progress in this area.

Goals

My first aim is to implement the method described by Schacht et al. (2014) and reconstruct each step necessary for this analysis of regulatory interactions as closely as possible to the original paper's specifications with the only exception being the underlying gene regulatory network. The network is an integral part of the published method, since the method aims to quantify the effects TFs have on the expression of networked genes. For this I will use a regulatory network suggested by Thomas et al. (2014). The data forming this network should be highly reliable as it was created by combining text mining with expert curation. Furthermore it is publicly available, in contrast to the data from the MetaCore database used by Schacht et al.

After applying this method to the given network, I am interested in how well it succeeds in explaining gene regulation. Therefore I will compare it with a tool called ISMARA which also aims on elucidating regulatory interactions by identifying the most influential regulators. Analysing how many regulators both methods agree on may as a first step indicate its soundness. Additionally I will apply cross validation and compare results to literature findings to further evaluate the method in its capabilities to explain regulation.

Approach

Method

Schacht et al. describe a method to measure TF activity by combining microarray data and an underlying regulatory network. Depending on the regulatory relationships in the network, in this approach the activity of a TF is calculated by measuring the expression of its affected genes. This has the advantage of including post-transcriptional modifications that might substantially impact

TFs regulating their target genes (Tootle, 2005). In contrast, calculating TF activity only by considering the expression of the associated gene might result in an incomplete picture.

For this task Schacht et al. present a regulation model which tries to minimize the difference between predicted and measured gene expression values. The predicted expression depends on several factors: the edge strength in the regulation network, the estimated effect of a TF on the genes in the samples and an additional coefficient serving as an optimization parameter. Since Schacht et al. base their network on combined data from several databases, the edge strength expresses how many of their sources include this regulatory relationship. In turn this has no impact on my implementation with a network from only a single source, thereby rendering this factor redundant.

Following their approach I will apply the Gurobi Optimizer 5.5[1] for optimizing the model, i.e. determining the coefficients in a way that minimizes the difference between predicted and measured gene expression.

Microarray Data

Analysis by Schacht et al. is based on a data set provided by the National Cancer Institute called NCI-60 panel. This is the largest source of cell-based anticancer testing data in a public database as it comprises 60 cancer cell lines of various origins (Shoemaker, 2006). Different microarray platforms were used for measuring gene expression, which must be taken into account during preprocessing. This includes combining and subsequently normalizing the data sets from five used platforms (Affymetrix HG-U95, HG-U133, HG-U133 Plus 2.0, GH Exon 1.0 ST and Agilent WHG).

Schacht et al. use the approach described by Reinhold et al. For this a data set of normalized gene expression values comprising cell lines from these five platforms is available at CellMiner[2] (Reinhold et al., 2012). As a result a z-score is stated for each gene, which stands for the distance to the general mean of all genes as measured in standard deviations. Z-score transformation of genes' expression intensities is especially suited for making microarray experiment results comparable in an automated fashion (Cheadle, 2003). Also following Schacht et al. one cell line, SF 539, is excluded for analysis as it lacks precision with a large number of entries being undefined.

Background Regulatory Network

The network used in this thesis will differ from the one used in the original paper. Instead of compiling data from the commercial database MetaCore and other sources, my approach is to build on a network which was created by text-mining biological literature and manually curating the most promising findings. Thomas et al. describe this procedure in detail; the data is publicly

[1] http://www.gurobi.com/

[2] http://discover.nci.nih.gov/cellminer/

available at the FastForward DNA database[3]. Regulatory relationships from three other databases (TRANSFAC, TRRD, OregAnno), which follow a similar approach by only integrating manually curated relations, complement the data found by the text-mining procedure.

This approach promises a robust network as the regulatory relationships were discovered in low-throughput experiments rather than with high-scale methods as ChIP-chip. With the former mentioned methods being more reliable (Furey, 2012), these findings are generally considered to be of higher quality. The downside being a far smaller amount of curated regulatory relationships with currently 807 in the FastForward database and 359 TFs involved. The underlying network used by Schacht et al. in contrast comprises 1120 TFs.

Evaluation

The resulting model will be evaluated using several approaches.

First I will compare the results with the ones from ISMARA (Integrated System for Motif Activity Response Analysis)[4]. It also aims on elucidating gene regulation but follows a different approach with a greater focus on promoters and their motifs (Balwierz et al., 2014). ISMARA calculates the most influential regulators from gene expression data and allows me to quantitatively compare the results. Counting how strong the both methods agree in identifying the most active regulators will indicate the degree to which their results match.

ISMARA is used by uploading gene expression data files to its web site and it then preprocesses and evaluates this data automatically. As the data formats from the five mentioned microarray platforms differ and the format used for the Agilent WHG platform is not applicable to ISMARA, only four of them may be used. For these the cell line SF 539 again has to be removed.

As a second approach, cross validation can show the model's success in explaining variance in gene expression. For this purpose, a fraction of the data is left out of the model-building process and afterwards the resulting model is tested on this data. This indicates the model's performance on independent data and its resilience against overfitting.

Thirdly, the FastForward DNA database created by Thomas et al. includes the kind of effect, i.e. activation or inhibition, TFs have on genes. This data again is considered to be of high reliability as it originates from low-throughput experiments. Therefore I can compare these findings with the effects in the model and use this degree of agreement as an additional factor for evaluation.

[3] http://fastforward.sys-bio.net/

[4] https://ismara.unibas.ch/fcgi/mara

References

Balwierz, P. J., Pachkov, M., Arnold, P., Gruber, A. J., Zavolan, M., & van Nimwegen, E. (2014): *ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs.* Genome research, 24(5), 869-884.

Cheadle, C., Vawter, M. P., Freed, W. J., & Becker, K. G. (2003): *Analysis of microarray data using Z score transformation.* The Journal of molecular diagnostics, 5(2), 73-81.

Furey, T. S. (2012): *ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions.* Nature Reviews Genetics, 13(12), 840-852.

Lemon, B., & Tjian, R. (2000): *Orchestrated response: a symphony of transcription factors for gene control.* Genes & development, 14(20), 2551-2569.

Orphanides, G. & Reinberg, D. (2002): *A Unified Theory of Gene Expression.* Cell, 108(4), 439-451.

Reinhold, W. C., Sunshine, M., Liu, H., Varma, S., Kohn, K. W., Morris, J., Doroshow, J. & Pommier, Y. (2012): *CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set*. Cancer Research, 72(14), 3499-3511.

Rottger, R., Ruckert, U., Taubert, J., & Baumbach, J. (2012): *How little do we actually know? On the size of gene regulatory networks.* Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 9(5), 1293-1300.

Schacht, T., Oswald, M., Eils, R., Eichmüller, S. B., & König, R. (2014): *Estimating the activity of transcription factors by the effect on their target genes*. Bioinformatics, 30(17), i401-i407.

Shoemaker, R. H. (2006): *The NCI60 human tumour cell line anticancer drug screen.* Nature Reviews Cancer, 6(10), 813-823.

Thomas, P., Durek, P., Solt, I., Klinger, B., Witzel, F., Schulthess, P., Mayer, Y., Tikk, D., Blüthgen, N. & Leser, U. (2014): *Computer-assisted curation of a human regulatory core network from the biological literature.* Bioinformatics (under review).

Tootle, T. L., & Rebay, I. (2005): *Post-translational modifications influence transcription factor activity: A view from the ETS superfamily*. Bioessays, 27(3), 285-298.