BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 20 2010, pages 2635–2636
doi:10.1093/bioinformatics/btq469

Gene expression

Advance Access publication August 24, 2010

TFInfer: a tool for probabilistic inference of transcription factor activities

H. M. Shahzad Asif[1], Matthew D. Rolfe[2], Jeff Green[2], Neil D. Lawrence[3], Magnus Rattray[3] and Guido Sanguinetti[1,*]

[1]School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, [2]Department of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield S10 2TN and [3]School of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL, UK

Associate Editor: Martin Bishop

ABSTRACT

Summary: TFInfer is a novel open access, standalone tool for genome-wide inference of transcription factor activities from gene expression data. Based on an earlier MATLAB version, the software has now been extended in a number of ways. It has been significantly optimised in terms of performance, and it was given novel functionality, by allowing the user to model both time series and data from multiple independent conditions. With a full documentation and intuitive graphical user interface, together with an in-built data base of yeast and *Escherichia coli* transcription factors, the software does not require any mathematical or computational expertise to be used effectively.

Availability: http://homepages.inf.ed.ac.uk/gsanguin/TFInfer.html

Contact: gsanguin@staffmail.ed.ac.uk

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

Received on June 25, 2010; revised on August 10, 2010; accepted on August 11, 2010

1 INTRODUCTION

Transcription regulatory networks play a fundamental role in mediating external signals and coordinating the response of the cell to its changing environment. Recent technological advances in molecular biology, such as ChIP-on-chip and ChIP-seq, are uncovering an increasing amount of data about the static structure of these networks, providing us with information about interactions between promoters and specific transcription factors (TFs). However, despite these advances, intracellular concentrations of active TF proteins remain very challenging to measure directly in a dynamic fashion, thus limiting our ability to understand the dynamics of transcriptional regulation. To obviate these problems, several research groups have proposed statistical approaches that infer TF activity levels by combining connectivity data about the structure of the regulatory network with microarray data (e.g. Sabatti and James, 2006; Sanguinetti *et al*., 2006). In this note, we present a novel implementation of one of these methods (Sanguinetti *et al*., 2006) which makes it freely available to the academic community in an intuitive, user-friendly platform. The method employs a linear approximation (in log space) to the dynamics of transcription and is based on a state space model of the following form

$$y_n(t) = \sum_{m=1}^{q} X_{nm} b_{nm} c_m(t) + \mu_n + \epsilon_{nt}$$
$$c_m(t) = \gamma_m c_m(t-1) + \eta_{mt}. \quad (1)$$

Here, $y_n(t)$ is the mRNA log expression level for gene n at time t, $\mathbf{X}$ is a binary *connectivity matrix* (assumed known) encoding whether gene n is bound by TF m, b_{nm} encodes the regulatory strength with which TF m affects gene n and $c_m(t)$ is the (log) concentration of active TF m at time t; the other terms are used to model noise and biases. The model places Gaussian prior distributions over the concentrations $c_m(t)$ and strengths b_{nm} and uses a factorized variational approximation to infer posterior distributions given mRNA time course observations. Notice that the probabilistic nature of the model means that noise is treated in a natural and principled way, and estimates of the quantities of interest are always associated with a measure of the corresponding uncertainty. Since only the product of b_{nm} and $c_m(t)$ appears in the likelihood, there is a sign ambiguity in the inferred quantities [see online tutorial and Sanguinetti *et al*. (2006) for further discussion].

While the approach does rely on a simplified model of transcription, the model's results have been shown to capture important physiological effects which have led to the formulation and experimental validation of several hypotheses (Davidge *et al.*, 2009; McLean *et al.*, 2010; Partridge *et al.*, 2007). However, the model was until now only available as working code in MATLAB, requiring expert intervention to be used which resulted in significant bottlenecks in the analysis pipeline. We have now produced a new release which presents several significant advantages over the previous version:

- it is open source, and significantly more efficient computationally;
- it is fully documented and has an intuitive Graphical User Interface (GUI);
- it contains template connectivities for *Escherichia coli* and *Saccharomyces cerevisiae*;
- it has been given extra functionalities, handling both time-series data and data from several independent conditions, possibly with replicates.

*To whom correspondence should be addressed.

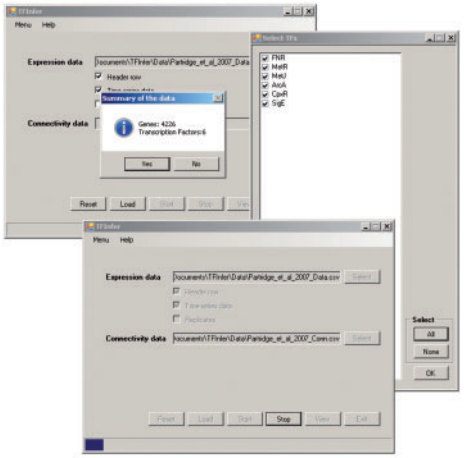

Fig. 1. Work flow of TFInfer.

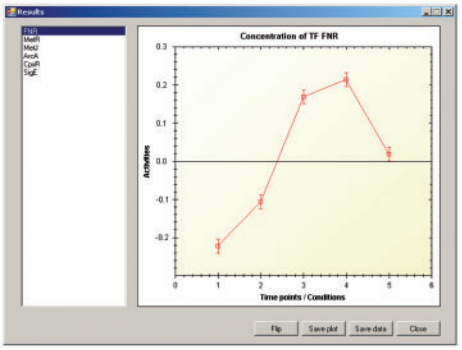

Fig. 2. Sample results obtained using TFInfer on data from Partridge *et al*. 2007.

2 SOFTWARE OVERVIEW

The model and GUI are implemented in C# which allows an efficient implementation of the variational Bayesian expectation maximization algorithm. For the numerical routines we used dnAnalytics (http://dnanalytics.codeplex.com/), a C# open source library for scientific computing. ZedGraph, an open source plotting tool, is used for displaying the results of the model in graphical format.

The work flow of TFInfer is shown in Figure 1 (overlapping tiles); the starting frame requires the user to browse for the expression data, specify its characteristics (time-series, replicates, etc.) and browse for the connectivity data. If template connectivity is selected, the user is asked to select either a file for yeast (based on available ChIP-on-chip data) or a file for *E.coli* (compiled manually from the Ecocyc data base, http://www.ecocyc.org/). Otherwise, the user can specify any binary connectivity matrix.

Once the data are selected, a summary of the data are displayed (number of genes and time points). If this is accepted, a list of all the TFs included in the connectivity matrix is displayed; the user can select a subset of TFs by clicking on the list of TFs names. Once this is completed, the optimisation starts; its progress (with respect to a maximum number of iterations, default 1500) is monitored through a progress bar at the bottom of the screen.

Once the run is complete, the user can visualise TF activity profiles by clicking the box next to the TF name. This displays a time series activity profile with associated error bars, and by clicking the save plot button the graph can be saved in a variety of formats. An example of the output of TFInfer is given in Figure 2, depicting the predicted activity of the FNR regulator in the switch from aerobic to microaerobic conditions, showing the overshoot in activity observed in Partridge *et al.* (2007).

2.1 Data files format and software requirements

Standard file format for TFInfer is comma separated file. This is a standard format supported by many spreadsheet applications including Microsoft Excel. Two types of input file are required; a csv file containing the logged gene expression data and a file specifying the connectivity matrix (which must be a binary matrix). Replicates are handled by uploading separate data files. For logged gene expression data, the file should contain a list of genes and the corresponding expression levels in different experimental conditions. Connectivity is specified in the form of grid where every entry (zero or one) specifies the connection between the corresponding TF and the gene; the first row of the file will contain the names of the TFs, and the first column the names of the genes. For *S.cerevisiae* and *E.coli*, this connectivity information is supplied as the part of the software; the gene names used are the systematic *b* names for *E.coli* and the open reading frame (ORF) identifiers for yeast. The software requires Microsoft .Net framework (version 2), which is freely downloadable (a link is provided on the TFInfer website). It runs on Windows platforms and on Linux/Mac via Mono.

3 CONCLUSION

Statistical methods for inferring TF activities are an important area of research in computational biology due to their ability to extract information which is not readily available through standard experimental practice. We believe that the time has arrived for these methods to become standard software used in biological laboratories to complement experimental work, much in the way that sequence alignment tools are now routinely used by experimentalists. By providing a simple yet powerful implementation of an already tried and tested method, we hope TFInfer will become accessible and useful to a wide community of scientists working on gene regulation.

Funding: University of Sheffield Director of Research devolved fund (to H.A., G.S. and J.G.). M.D.R. and J.G. thank the BBSRC for support through the SysMO initiative (www.sysmo.net) and are members of the SysMO-SUMO consortium.

Conflict of Interest: none declared.

REFERENCES

Davidge,K.S. *et al*. (2009) Carbon monoxide-releasing antibacterial molecules target respiration and global transcriptional regulators. *J. Biol. Chem.*, **284**, 4516–4524.

McLean,S. *et al*. (2010) Peroxynitrite toxicity in Escherichia coli K-12 elicits expression of oxidative stress responses, and protein nitration and nitrosylation. *J. Biol. Chem.*, **285**, 20724–20731.

Partridge,J.D. *et al*. (2007) Transition of Escherichia coli from aerobic to micro-aerobic conditions involves fast and slow reacting regulatory components. *J. Biol. Chem.*, **282**, 11230–11237.

Sabatti,C. and James,G. (2006) Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, **22**, 739–746.

Sanguinetti,G. *et al*. (2006) Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, **22**, 2775–2781.