BIOINFORMATICS APPLICATIONS NOTE

Vol. 24 no. 21 2008, pages 2544–2545
doi:10.1093/bioinformatics/btn467

Gene expression

# STREAM: Static Thermodynamic REgulAtory Model of transcription

Denis C. Bauer* and Timothy L. Bailey*

Institute for Molecular Bioscience, University of Queensland, Brisbane, Qld. 4072, Australia

Received on June 27, 2008; revised on August 11, 2008; accepted on August 28, 2008

Advance Access publication September 6, 2008

Associate Editor: Joaquin Dopazo

**ABSTRACT**

**Motivation:** Understanding the transcriptional regulation of a gene in detail is a crucial step towards uncovering and ultimately utilizing the regulatory grammar of the genome. Modeling transcriptional regulation using thermodynamic equations has become an increasingly important approach towards this goal.

Here, we present STREAM, the first publicly available framework for modeling, visualizing and predicting the regulation of the transcription rate of a target gene. Given the concentrations of a set of transcription factors (TFs), the TF binding sites (TFBSs) in a regulatory DNA region, and the transcription rate of the target gene, STREAM will optimize its parameters to generate a model that best fits the input data. This trained model can then be used to (a) validate that the given set of TFs is able to regulate the target gene and (b) to predict the transcription rate under different conditions (e.g. different tissues, knockout/additional TFs or mutated/missing TFBSs).

**Availability:** The platform independent executable of STREAM, as well as a tutorial and the full documentation, are available at http://bioinformatics.org.au/stream/. STREAM requires Java version 5 or higher.

**Contact:** d.bauer@imb.uq.edu.au; t.bailey@imb.uq.edu.au

## 1 INTRODUCTION

Transcription of a gene can be induced by the binding of specific transcription factors (TFs) to so-called *cis*-regulatory modules (CRMs). The frequency and duration of the binding events are influenced by the concentrations of the TFs, the binding affinities and location of the TF binding sites (TFBSs) in the CRM and the properties of the TFs themselves (e.g. effectiveness, competitive interaction with other TFs). With the availability of an increasing number of detailed measurements of gene concentrations in different situations (e.g. tissues, developmental time points) as well as TF-DNA binding affinities, it has become possible to build mathematical models for transcriptional regulation. Building mathematical models to associate a specific occupation of a specific CRM with an observed transcriptional response promotes a better understanding of the transcriptional regulation and enables *in silico* hypothesis testing about postulated regulatory TFs or mechanisms.

An increasingly successful approach to mathematically simulating transcriptional regulation is using thermodynamic models that model the interaction of TFs and DNA using kinetic equations. Several such thermodynamic models have been proposed in the last years (Janssens *et al*., 2006; Segal *et al*., 2008; Zinzen *et al*., 2006). These models take the CRM sequence, a set of TFs along with their concentrations and predict the transcriptional response of the target gene as mediated by the CRM and the TFs. A training algorithm is used to optimize the model's internal parameters to minimize the difference between the observed and predicted transcriptional response.

## 2 APPROACH AND USAGE

Here we present STREAM, a Java-implemented framework to calculate and visualize transcriptional regulation using thermodynamic modeling approaches. STREAM currently uses the thermodynamic model introduced by Reinitz *et al*. (2003), but the framework is flexible and can be used in conjunction with other models implemented in Java. STREAM offers several optimization methods including gradient descent and simulated annealing for adjusting the internal parameters of the model to best fit the user's input data. To the best of our knowledge, STREAM is to data the only publicly available framework for modeling the regulation of the rate of transcription. STREAM has been tested extensively on the even-skipped gene (*eve*) in *Drosophila melanogaster* (Bauer and Bailey, 2008).

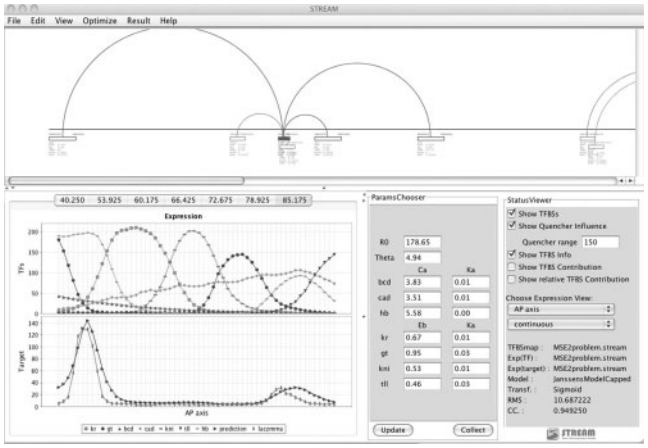
STREAM can be executed using a graphical user interface (GUI) as well as via the command line. The GUI is illustrated in Fig. 1. It offers the same functionality (e.g. multistart options of the optimization or automatic cross-validation evaluation) as the command-line tool, but in an intuitive and dialogue-based fashion. Both the command-line and the GUI version can save the current result and settings of the program into a file, which makes saving and modifying previous experiments simple.

## 3 METHODS

In order to generate a model for a particular target gene, the user must identify a set of putative TFs and a putative regulatory region. The suspected role of each TF, $x$, as an activator $x \in A$ or repressor $x \in B$ must be specified by the user. The program also requires measurements of the concentrations and binding preferences of those TFs, as well as the transcriptional output of the target gene. In the following section, we introduce the required input data, **D**, which contains the concentration and rate data, **C**, and a TFBS map, **S**.

The concentration and rate data, **C**, comprises a set of independent data points. Each data point is a pair $(\mathbf{V}, v_t)$, with $\mathbf{V} = ([a_1], \ldots, [a_n])$, a vector listing the protein concentrations of the putative regulatory TF proteins

*To whom correspondence should be addressed.



|  |  |  |
|---|---|---|
| R0 | 178.65 |  |
| Theta | 4.94 |  |
|  | Ca | Ka |
| bcd | 3.83 | 0.01 |
| cad | 3.51 | 0.01 |
| hb | 5.58 | 0.00 |
|  | Eb | Ka |
| kr | 0.67 | 0.01 |
| gt | 0.95 | 0.03 |
| kni | 0.53 | 0.01 |
| tll | 0.46 | 0.03 |

☑ Show TFBSs
☑ Show Quencher Influence
Quencher range 150
☑ Show TFBS Info
☐ Show TFBS Contribution
☐ Show relative TFBS Contribution

Choose Expression View:
AP axis
continuous

| | |
|---|---|
| TFBSmap : | MSE2problem.stream |
| Exp(TF) : | MSE2problem.stream |
| Exp(target) : | MSE2problem.stream |
| Model : | JanssensModelCapped |
| Transf. : | Sigmoid |
| RMS : | 10.687222 |
| CC. : | 0.949250 |

**Fig. 1.** Screen shot of the GUI of STREAM. The interface is exemplified on a model trained for the even-skipped gene (*eve*).

$(a_1,\ldots,a_n)$, and $v_t$ giving the corresponding observed transcription rate of the target gene.

The concentration and rate data, **C**, can be measured by various methods. To measure the TF concentrations, *in situ* antibody staining can be used to measure **V**. To obtain the corresponding $v_t$, one can proceed indirectly by measuring the mRNA levels of a reporter gene by staining against the mRNA of the reporter (Jaeger *et al*., 2004).

For visualization purposes, STREAM allows the user to label the data points with up to two 'features' (e.g. the 'condition' under which the measurements were made). Features may be 'continuous' (real numbers as in the example above), or 'categorical' such as cell types, tissue types or experimental treatments. STREAM utilizes the values of the features for visualization only, and the user may interactively select either feature as the $X$-axis for plots of the data.

Besides the concentration data, the input data also contains the TFBS map, **S**, corresponding to the regulatory region of interest. The TFBS map, $\mathbf{S}=(\mathbf{s_1},\mathbf{s_2},\ldots,\mathbf{s_n})$, is a list, where each $s_i$ represents a TFBS as a triple $(a,l,s)$ giving the name, $a$, of the binding TF, the position, $l$, of the binding site and the log-odds score (natural logarithm), $s$, of the site. The log-odds score is proportional to the binding strength of the TFBS (Stormo, 2000). **S** can be constructed from experimentally verified binding sites, *in silico* predicted sites using a prediction algorithm such as FIMO (http://meme.sdsc.edu), or a combination of both. For more detailed information, see Bauer and Bailey (2008).

The objective of the optimization is to determine the set of model parameters, $\Theta$, that optimally explains the input data, **D**. $\Theta$ depends on the thermodynamic model. Currently, STREAM implements the model introduced by Reinitz *et al*. (2003). This model uses free parameters $\Theta=(\theta_0,R_0,\mathbf{W})$, where $\theta_0$ is an energy barrier, $R_0$ is the maximal transcription rate, and **W** contains a tuple, $(K_x,E_x)$, for each TF, $x$, where $K_x$ is the association constant of the TF to the DNA and $E_x$ is the effectiveness of the TF to activate transcription, if $x\in A$, or to repress, if $x\in B$. For more details see Reinitz *et al*. (2003) and Bauer and Bailey (2008).

Four different optimization methods are implemented: simulated annealing (SA), gradient descent (GD), genetic algorithm (GA) and limited-memory quasi-Newton unconstrained optimization (LBFGS) (for a comparison of the optimization methods see D.C.Bauer and T.L.Bailey, manuscript in preparation). All optimization methods seek to optimize the free parameters of the Reinitz model by minimizing the root mean-squared (RMS) error between the *known* transcription rate and the rate *predicted* by the Reinitz model, averaged over all input points, **D** (Bauer and Bailey, 2008).

## 4 FUTURE DEVELOPMENTS

Besides the Reinitz model, we plan to provide additional models with enhanced functionality to simulate interacting TFs in a more detailed way, e.g., by incorporating TF-TF cooperation. Furthermore, simplifying models by using discreet TFBSs might introduce artifacts, hence, changing the model approach to use continuous binding gradients deems beneficial. Future research on other CRMs will guide the development of new models and our framework can provide the environment to directly compare them.

In addition, we plan to extend the approach to optimize one model to fit more than one CRM. Being able to fit one model to the data of several CRMs, which are suspected to have the same regulatory TFs, increases the confidence in the produced model.

Finally, we plan to improve the functionality of the GUI in manipulating input data, e.g., by an interactive interface to vary the properties of the TFBS map and to directly observe the changes in the predicted transcription rate.

## ACKNOWLEDGEMENTS

The authors wish to acknowledge John Reinitz, Stefan Maetschke and Fabian Buske for their input on conceptual and implementation details for STREAM as well as Lachlan Dufton for providing the initial implementation of SA.

*Funding*: Australian Research Council Centre of Excellence in Bioinformatics (to T.L.B.); National Institutes of Health (RO-1 RR021692-01 to T.L.B.), UQ International Research Tuition Award (D.C.B.).

*Conflict of Interest*: none declared.

## REFERENCES

Bauer,D.C. and Bailey,T.L. (2008) Studying the functional conservation of cis-regulatory modules and their transcriptional output. *BMC Bioinformatics*, **9**, 220.

Jaeger,J. *et al*. (2004) Dynamical analysis of regulatory interactions in the gap gene system of Drosophila melanogaster. *Genetics*, **167**, 1721–1737.

Janssens,H. *et al*. (2006) Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. *Nat. Genet.*, **38**, 1159–1165.

Reinitz,J. *et al*. (2003) Transcriptional control in Drosophila. *Complexus*, **1**, 54–64.

Segal,E. *et al*. (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, doi:10.1038/nature06496.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Zinzen,R.P. *et al*. (2006) Computational models for neurogenic gene expression in the Drosophila embryo. *Curr. Biol.*, **16**, 1358–1365.