Bioinformatics Advance Access published March 16, 2012

GRiP: a computational tool to simulate transcription factor binding in prokaryotes

Nicolae Radu Zabet [1,2,*] and Boris Adryan [1,2]

[1]Cambridge Systems Biology Centre, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK; [2]Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK

Associate Editor: Dr. Trey Ideker

ABSTRACT

Motivation: Transcription factors (TFs) are proteins that regulate gene activity by binding to specific sites on the DNA. Understanding the way these molecules locate their target site is of great importance in understanding gene regulation. We developed a comprehensive computational model of this process and estimated the model parameters in (Zabet and Adryan, 2011).

Results: GRiP is a highly versatile implementation of this model and simulates the search process in a computationally efficient way. This program aims to provide researchers in the field with a flexible and highly customisable simulation framework. Its features include representation of DNA sequence, TFs and the interaction between TFs and the DNA (facilitated diffusion mechanism), or between various TFs (cooperative behaviour). The software will record both information on the dynamics associated with the search process (locations of molecules) and also steady state results (affinity landscape, occupancy-bias, collision hotspots).

Availability: http://logic.sysbiol.cam.ac.uk/grip

Supplementary information: Additional information and data is presented in the *Supplementary Material* which is available at *Bioinformatics* online.

Contact: n.r.zabet@gen.cam.ac.uk

1 INTRODUCTION

It is well established now that TF find their target site through *facilitated diffusion*, a combination between 1D random walk on the DNA and 3D diffusion in the cytoplasm (Berg *et al.*, 1981; Elf *et al.*, 2007). Once bound to the DNA, TFs perform three main types of movements: (i) sliding , (ii) hopping and (iii) jumping (Mirny *et al.*, 2009). The first two mechanisms, sliding and hopping, assume that the TF performs small movements on the DNA without releasing into the cytoplasm, whilst the third assumes a 3D diffusion in the cytoplasm before rebinding.

With few exceptions, most of the theoretical efforts have been invested into analytical solutions of the facilitated diffusion mechanism. If one wants to consider real DNA sequences and dynamic crowding on the DNA (mobile 'roadblocks'), then this rules out analytical solutions. Computational methods and, in particular, stochastic simulations overcome these limitations and provide a more accurate mechanistic representation of the underling biological process. In particular, these type of stochastic simulations can be used to answer question related to how TFs perform the search process. For example, one could investigate whether molecules prefer to hop or to slide and what is the contribution of these two alternative movements on the DNA to the overall 1D random walk in a crowded environment.

Building on the comprehensive model constructed in (Zabet and Adryan, 2011), we developed GRiP (Gene Regulation in Prokaryotes), a program that allows stochastic simulation of the search process of TFs for their target sites on the DNA.

The analysed systems can be large. For example, *E.coli* K-12 has a $4.6\ Mbp$ genome and there are $\sim 10^4$ DNA binding proteins (agents). To produce results within relative short time, previous software had to either rely on coarse grain models (Wunderlich and Mirny, 2008) or to consider small subsystems (Chu *et al.*, 2009). GRiP represents a new and efficient implementation of the TF search process, which considers a highly detailed model of 1D diffusion and, at the same time, it simulates at least ≈ 4 times faster than previous software (Chu *et al.*, 2009; Barnes and Chu, 2010). Consequently, by allowing genome-wide stochastic simulations of a highly detailed model of facilitated diffusion, GRiP can highlight possible biases in the results where the level of details was insufficient (coarse grain models) or the size of the analysed system was too small.

A few studies, such as Das and Kolomeisky (2010), addressed the problem of facilitated diffusion through simulations focusing on the 3D diffusion rather than the 1D case. The 3D diffusion is time and resource consuming, especially for simulations at the genome level. van Zon *et al.* (2006) showed that the model based on the zero-dimensional Chemical Master Equation can reliably represent the rate at which TFs associate non-specifically with the DNA, as long as the model takes into account that once a molecule unbinds from the DNA, it has a high probability of fast rebinding in close proximity. This suggests that there is no need to simulate the 3D diffusion explicitly, but rather have this replaced by a simple arrival rate and ensuring that the model incorporates the fast rebinding probability in the unbinding rate, a strategy which we also adopt.

*to whom correspondence should be addressed

© The Author(s) 2012. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on October 1, 2015

2 DESCRIPTION

We implemented the target finding process as a hybrid model mixing agent based methods with event driven stochastic simulation algorithms (Gillespie, 1977). The software is implemented in Java 1.6, which ensures high portability.

In the simulator, each TF molecule is represented as an agent able to perform certain actions, while the DNA molecule is modelled as a string of base pairs (A, T, C, G). There is no measure of distance between the molecules, but the TF molecules can be either free in the cytoplasm or bound on the DNA at certain positions. The free TF molecules have only one action available, namely to bind to the DNA.

The cytoplasm is assumed to be a perfectly mixed reservoir from where the free TF molecules can find the DNA at exponentially distributed times. To simulate the 3D diffusion we use the Direct Method implementation of Gillespie Algorithm (Gillespie, 1977) which generates a statistically correct trajectory of the Master Equation.

The model considers volume exclusion, allowing only one TF to cover certain base pair at any specific time point. A bound molecule will occupy a number of consecutive base pairs on the DNA. The size on the DNA of each TF molecule is computed as the number of base pairs of the DNA binding motif added to the number of obstructed base pairs on the left side of the molecule and the number of obstructed base pairs on the right side.

A feature which was not considered by previous models (Chu *et al.*, 2009; Barnes and Chu, 2010) is TF orientation on the DNA. If TFs are not symmetric, the user can set TF molecules to have two orientations on the DNA, which can lead to different affinities depending on the molecule orientation. Whenever a TF binds to the DNA, the system selects a random orientation. This can be changed only after the TF molecule unbinds and rebinds to the DNA, including during hops.

The simulator supports the definition of multiple TF species, which are classified in two types: (i) non-cognate TFs and (ii) cognate TFs. The cognate ones are the TFs that are of interest and that we can follow, while the non-cognate ones' main purpose is to simulate the 'other' proteins on the DNA, which might interfere with the search process of the cognate TFs. For efficiency reasons, we pre-calculate the affinities of each TF species, both cognate and non-cognate, and store them in individual arrays. The non-cognate binding energy is randomly generated using a Gaussian distribution with the mean and variance provided as inputs for each non-cognate species.

For cognate TFs, there are several ways in which the binding energy can be computed, but this work is restricted to three well known ones: (i) mismatch energy (Gerland *et al.*, 2002), (ii) Position Frequency Matrix (PFM) and information theory (Stormo, 2000) and (iii) PFM and binding energy (Berg and von Hippel, 1987). In all scenarios, we assume that each position in the DNA binding motif is approximately independent and additive.

A bound TF molecule can perform, with user defined probabilities, one of the following actions: (i) slide left, (ii) slide right, (iii) hop to a position that is Gaussian distributed around original position with a user defined variance or (iv) unbind from the DNA. We assume reflecting boundaries. In the case the molecule unbinds, there is a certain probability that it will rebind fast near the original place.

Finally, the model allows cooperative behaviour between TF molecules and this can be either mediated by DNA (binding of one molecule to a certain site on the DNA can alter the affinity between another molecule and a different site) or represented as direct TF-TF interaction (two molecules bound to the DNA and in physical contact can have different affinities for their current positions compared to the case where they are not in contact); for more details see (Zabet and Adryan, 2011).

The simulation speed is sensitive to the number of agents in the system. This mainly comes from the fact that the events queue becomes larger with increasing number of molecules in the system and, consequently, higher queues require higher maintenance time. For 10^6 TF molecules and the genome of *E.coli* K-12 ($4.6\ Mbp$), we can simulate $\sim 4 \times 10^5$ events per second on a Mac Pro 2x2.26GHz quad-core Intel Xeon with 32GB memory running Mac OSX 10.6.8.

3 DISCUSSION

GRiP is a highly versatile program which comes with both command-line interface and graphical user interface. Furthermore, being written in Java, the software can be run on any machine where the Java Runtime Environment 1.6 (or higher) is installed.

The program takes as input a *parameters file*, which can specify, among many other parameters, three additional data files, namely: (i) the DNA sequence file (from a FASTA file), (ii) TF file (a csv file with TF-specific characteristics) and, optionally, (iii) TF cooperativity file (a csv file). Note that, if either the DNA sequence file or the TF file are not provided, then the simulator can randomly generate that data (DNA sequence or TF species).

Once started, the simulation runs until the time in the cell reaches a predefined stop time, or until all target sites are reached (if the stop time is set to 0). As output, the simulator can print information on: (i) the position of TF molecules on the DNA (or proportion of bound molecules to the DNA), (ii) computed affinity landscapes for each TF species, (iii) measured occupancy bias for each TF species, (iv) statistical information related to TF species (such as residence time, sliding lengths, actual sliding lengths, binding events, etc.), (v) simulation speed, (vi) stored sliding lengths for each species and (vii) statistics on collisions (total number, total number per species, hot spots on DNA).

GRiP can simulate $1\ s$ of *E.coli* K-12 and lacI using biologically plausible parameters between $1\ h$ and $4\ h$ (depending on the simulation parameters, the machine on which the simulation is run and even on the interface of the application, GUI or command line), which means that one can simulate up to $10\ min$ of a bacterial cell within a month; for details see *Supplementary Material*.

Funding: This work was supported by the Medical Research Council [G1002110 to N.R.Z.] and the Royal Society [B.A.].

REFERENCES

Barnes, D. J. and Chu, D. F. (2010). An efficient model for investigating specific site binding of transcription factors. In *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on*, pages 1–4, Chengdu, China. IEEE Xplore.

Berg, O. G. and von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins statistical-mechanical theory and application to operators and promoters. *Journal of Molecular Biology*, **193**(4), 723–750.

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on October 1, 2015

Berg, O. G., Winter, R. B., and von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. models and theory. *Biochemistry*, **20**(24), 6929–6948.

Chu, D., Zabet, N. R., and Mitavskiy, B. (2009). Models of transcription factor binding: Sensitivity of activation functions to model assumptions. *Journal of Theoretical Biology*, **257**(3), 419–429.

Das, R. K. and Kolomeisky, A. B. (2010). Facilitated search of proteins on DNA: correlations are important. *Physical Chemistry Chemical Physics*, **12**, 2999–3004.

Elf, J., Li, G.-W., and Xie, X. S. (2007). Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*, **316**, 1191–1194.

Gerland, U., Moroz, J. D., and Hwa, T. (2002). Physical constraints and functional characteristics of transcription factor-DNA interaction. *PNAS*, **99**(19), 12015–12020.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, **81**, 2340–2361.

Mirny, L., Slutsky, M., Wunderlich, Z., Tafvizi, A., Leith, J., and Kosmrlj, A. (2009). How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *Journal of Physics A: Mathematical and Theoretical*, **42**, 434013.

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, **16**(1), 16–23.

van Zon, J. S., Morelli, M. J., Tanase-Nicola, S., and ten Wolde, P. R. (2006). Diffusion of transcription factors can drastically enhance the noise in gene expression. *Biophysical Journal*, **91**, 4350–4367.

Wunderlich, Z. and Mirny, L. A. (2008). Spatial effects on the speed and reliability of protein-DNA search. *Nucleic Acids Research*, **36**(11), 3570–3578.

Zabet, N. R. and Adryan, B. (2011). A comprehensive computational model of facilitated diffusion in prokaryotes. Manuscript submitted for publication.

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on October 1, 2015