Structural bioinformatics

DIA-MCIS: an importance sampling network randomizer for network motif discovery and other topological observables in transcription networks

D. Fusco¹*, B. Bassetti¹,², P. Jona³ and M. Cosentino Lagomarsino¹,²,⁴

¹Università degli Studi di Milano, Dip. Fisica, Via Celoria 16, 20133 Milano, ²I.N.F.N., Milano, ³Politecnico di Milano, Dip. Fisica, Pza Leonardo Da Vinci 32, 20133 Milano, Italy and ⁴UMR 168/Institut Curie, 26 rue d’Ulm 75005 Paris, France

Received on July 24, 2007; revised on August 18, 2007; accepted on August 27, 2007

Advance Access publication September 27, 2007

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Transcription networks, and other directed networks can be characterized by some topological observables (e.g. network motifs), that require a suitable randomized network ensemble, typically with the same degree sequences of the original ones. The commonly used algorithms sometimes have long convergence times, and sampling problems. We present here an alternative, based on a variant of the importance sampling Monte Carlo developed by (Chen et al.).

Availability: The algorithm is available at http://wwwteor.mi.infn.it/~bassetti/downloads.html

Contact: diana.fusco@studenti.unimi.it and marco.cosentino@unimi.it

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Gene regulatory networks represent interactions between genes or proteins. For example, in a transcription network the nodes are genes, and the edges represent TF-promoter interactions (Babu et al., 2004). Considered their topology, one can study the deviation of the empirical topology from a ‘typical case’ statistics (Milo et al., 2004). To this end, one generates so called ‘randomized counterparts’ of the original dataset as a null model. That is, an ensemble of random networks with some invariant properties, such as the degree sequences, i.e. the number of outgoing and incoming links for each node. This approach has a wider application for networks of different kinds (Kashtan et al., 2004; Milo et al., 2004). Some algorithms to generate this uniformly distributed ensemble are commonly used (Chen et al., 2005; Milo et al., 2003). In particular, one Markov Chain Monte Carlo (MCMC) algorithm is based on swapping edges at random (Molloy et al., 1995). This generates an ergodic dynamics, with, however, large relaxation times. Another type of algorithm is the so called ‘stub-pairing’ algorithm (Milo et al., 2003) that consists in randomly linking ‘stubs’ made of nodes with prescribed degrees (Maslov et al., 2005; Rao et al., 1996). This technique may fall in metastable states, where no stubs can be connected (King, 2004). The algorithm developed in Chen et al. (2005) is free of these two problems. Based on importance sampling Monte Carlo, it generates matrices with an almost uniform probability, and subsequently adjusts the sample, assigning to every element a certain weight. Moreover, it is able to estimate the size of the sampled ensemble. Here, we present an implementation of this algorithm that works specifically on transcription networks, but may be applied in general, with two variants. The first variant is designed to improve speed and make the algorithm competitive to the existing ones, while sampling more efficiently. The second variant deals with ensembles of structured matrices, in particular with structured diagonal, as it is often done in transcription networks when dealing with self-regulations (Kashwan et al., 2004).

2 ALGORITHM

A directed network can be conveniently represented as a zero-one adjacency matrix where element $a_{ij}$ is 1 if node $j$ has a directed link to node $i$ (Fig. 1A). The null ensemble of degree-conserving graphs translates into a set of matrices having the same row and column sums of the empirical matrix. As the goal is the uniform distribution of the sample, the importance sampling weight for every element is $1/P(T)$, where $P(T)$ is the matrix probability. The algorithm generates the matrix column by column as illustrated in Figure 1A. One has to consider the row sums having subtracted the first column. When all the columns are filled, the total probability of having a certain matrix is the product of all the column probabilities, which can be computed knowing the constraints of each column (Chen et al., 1997). This number allows to weigh correctly the matrix sample. We introduced the following two variants.

2.1 Large matrices with compact indegree

Transcription networks typically have several hundreds of nodes. The computational cost for generating a column is of order $O(M^2c^2)$ where $M$ is the length of a column and $c$...
distributions in E.coli the E.coli shifted with respect to each other. Of TGC (Three Gene Chain), for the transcription networks Forward Loop), SIM (triangular Single Input Module) and of three triangular subgraphs (Fig. 1B and C). The FFL (Feed Forward Loop), SIM (triangular Single Input Module) and TGC (Three Gene Chain), for the transcription networks. Note that the E. Coli dataset has no feedback. (measured by Mcore, see Cosentino Lagomarsino et al., 2006) of the E.coli and Saccharomyces cerevisiae transcription networks. Note that the E. Coli dataset has no feedback. (B) Table summarizing the results for three triangular motifs and the feedback (measured by Mcore, see Cosentino Lagomarsino et al., 2006) of the E.coli and Saccharomyces cerevisiae (Guelzim et al., 2002) verifying the results that can be found in the literature (Kashtan et al., 2004), (Milo et al., 2004) and with the algorithm of (Chen et al., 2005) (Supplementary Fig. 1). In all cases, we find a quantitative difference between the subgraph distributions in the randomized ensembles with or without structured diagonal (Fig. 1B and C). In some instances, such as the FFL, this does not affect the status of motif. In other cases one can also find qualitative changes.

2.2 Structured diagonal
Self-regulatory interactions are often considered to have a particular status (Kashtan et al., 2004). They are represented in the matrix by 1 on the diagonal. In order to constrain the diagonal, we accounted for the fact that some positions are not available for the extraction (see Supplementary Material).

3 IMPLEMENTATION AND RESULTS
3.1 Triangular network motif
As an example of application, we have studied the occurrence of three triangular subgraphs (Fig. 1B and C). The FFL (Feed Forward Loop), SIM (triangular Single Input Module) and TGC (Three Gene Chain), for the transcription networks of Escherichia coli (Shen-Orr et al., 2002) and Saccharomyces cerevisiae (Guelzim et al., 2002) verifying the results that can be found in the literature (Kashtan et al., 2004). (Milo et al., 2004) and with the algorithm of (Chen et al., 2005) (Supplementary Fig. 1). In all cases, we find a quantitative difference between the subgraph distributions in the randomized ensembles with or without structured diagonal (Fig. 1B and C). In some instances, such as the FFL, this does not affect the status of motif. In other cases one can also find qualitative changes.

3.2 Feedback
We also evaluated (Fig. 1B) the feedback in the graph, using a simple decimation algorithm that removes the input- and output-tree-like components (Cosentino Lagomarsino et al., 2006). With this algorithm, the feedback is measured by the size Mcore of the decimated graph. As expected, the sample with structured diagonal is shifted towards smaller amounts of feedback. This can be explained considering the lower amount of available links to rearrange if the self-regulators are fixed.

4 CONCLUSIONS
In conclusion, we have implemented and tested a Monte Carlo importance sampling algorithm to randomize directed graphs conserving the degree sequence, and evaluate topological observables. The algorithm follows the design principles of...
Diaconis et al. but is more efficient without loss of uniformity on graphs with compact indegree such as the known transcription networks. Furthermore, we added a variant that works with constrained diagonal, as it is usually done in motif discovery (Kashtan et al., 2004). We implemented the algorithm as a C++ three-node motif and feedback finder (also available as linux and windows executable).

ACKNOWLEDGEMENTS
The authors would like to thank F. Bassetti, S. Holmes and P. Diaconis for helpful discussions.

Conflict of Interest: none declared.

REFERENCES