BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 24 2007, pages 3388–3390
doi:10.1093/bioinformatics/btm454

Structural bioinformatics

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

D. Fusco[1,*], B. Bassetti[1,2], P. Jona[3] and M. Cosentino Lagomarsino[1,2,4]

[1]Università degli Studi di Milano, Dip. Fisica, Via Celoria 16, 20133 Milano, [2]I.N.F.N., Milano, [3]Politecnico di Milano, Dip. Fisica, Pza Leonardo Da Vinci 32, 20133 Milano, Italy and [4]UMR 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 (Kashtan *et al.*, 2004).

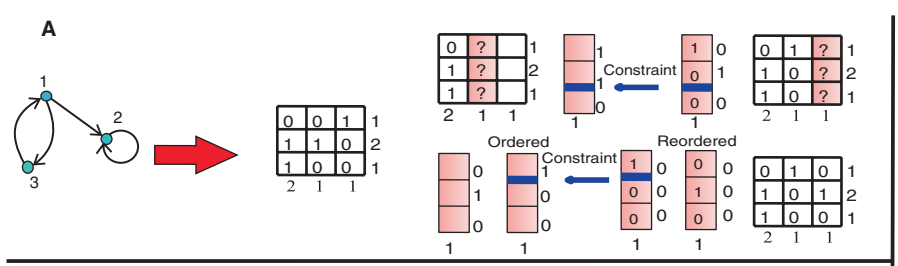
*To whom correspondence should be addressed.

2 ALGORITHM

A directed network can be conveniently represented as a zero-one adjacency matrix where element $a_{(i,j)}$ 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 $\mathcal{O}(M^2c^2)$ where M is the length of a column and c



B

				free diagonal		costrained diagonal	
	fig.	name	empirical value	rand ± sd	zscore	rand ± sd	zscore
E. Coli		FFL	42	19.2 ± 4.8	4.75	7.81 ± 3.13	10.93
		SIM	4777	5090 ± 40	7.83	4810 ± 5	6.6
		TGC	202	560 ± 41	8.73	193 ± 6	1.5
		M_{core}	0	16.3 ± 3.6	4.53	0.22 ± 0.71	0.31
S. cerevisiae		FFL	95	19.4 ± 4.4	17.2	17.1 ± 4.2	18.5
		SIM	5019	5175 ± 15	10.4	5112 ± 9	10.3
		TGC	279	453 ± 18	9.67	375 ± 10	9.6
		M_{core}	2	8.4 ± 3.4	1.88	0.41 ± 0.71	1.45

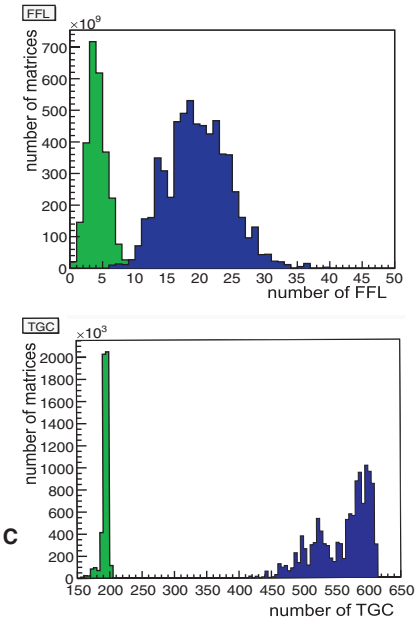


Fig. 1. (**A**) Description of the algorithm. Left: the graph is translated into an adjacent matrix, which is filled column by column. Right/top: the procedure for generating a matrix is showed. The second column (in pink) must be generated. In order to perform this operation, the updated row sums and the constraint are calculated. The free positions are extracted and the algorithm considers the next (and last) column. Right/bottom: the column is then reordered by the residual row sums. One constraint is found and the last column filled. The matrix is now complete. (**B**) Table summarizing the results for three triangular motifs and the feedback (measured by M_{core}, see Cosentino Lagomarsino *et al.*, 2006) of the *E.coli* and *S.cerevisiae* transcription networks. Note that the *E. Coli* dataset has no feedback. (**C**) Comparison of subgraph (FFL, TGC) distributions in *E.coli* randomizations for structured (light green) and unstructured (dark blue) diagonals. These distributions are systematically shifted with respect to each other.

the number of 1s contained in that column (Chen *et al.*, 2005). This is due to the fact that, the algorithm has to evaluate the probability of success for every position inside the column (Bekazova *et al.*, 2006). The probability of success in a given position can be well approximated using the corresponding row-sum if the in-degree distribution is sufficiently limited in range (see Supplementary Material and Supplementary Fig. 1). This last feature is typical of transcription networks. The computational cost for generating a column is then reduced to order $\mathcal{O}(Mc)$.

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-treelike components (Cosentino Lagomarsino *et al.*, 2006). With this algorithm, the feedback is measured by the size M_{core} 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

Babu,M. *et al*. (2004) Structure and evolution of gene regulatory networks. *Curr*. *Opin*. *Struct*. *Biol*., **14**, 14–283.

Bekazova,I. *et al*. (2006) Negative examples for sequential importance sampling of binary contingency tables. *Lecture Notes in Computer Science*, 4168, Springer, Berlin, pp.136–147.

Chen,S.X. and Liu,J.S. (1997) Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, **7**, 875–892.

Chen,Y. *et al*. (2005) Sequential Monte Carlo methods for statistical analysis of tables. *J*. *Am*. *Stat*. *Assoc*., **100**, 109–120.

Cosentino Lagomarsino,M. *et al*. (2006) Randomization and feedback properties of directed graphs inspired by gene networks. *q-bio.MN/0606039*.

Guelzim,N. *et al*. (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat*. *Genet*., **31**, 60–63.

Kashtan,N. *et al*. (2004) Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, **20**, 1746–1758.

King,O. D (2004) Comments on "Subgraphs in random networks". *Phys*. *Rev*., 058101–1,2,3:E**70**.

Maslov,S. and Sneppen,K. (2005) Computational architecture of the yeast regulatory network. *Phys*. *Biol*., **2**, 94.

Milo,R. *et al*. (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824.

Milo,R. *et al*. (2003) On the uniform generation of random graphs with prescribed degree sequences. *cond-mat/0312028*.

Milo,R. *et al*. (2004) Superfamilies of designed and evolved networks. *Science*, **303**, 1538.

Molloy,M. and Reed,B. (1995) A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, **6**, 161–179.

Rao,A. *et al*. (1996) A Markov chain Monte Carlo method for generating random zero-one matrices with given marginals. *Indian J*. *Stat*., **58**, 225.

Shen-Orr,S.S. *et al*. (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nat*. *Genet*., **31**, 64–68.