BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 18 2004, pages 3638–3640
doi:10.1093/bioinformatics/bth395

SARGE: a tool for creation of putative genetic networks

O. J. Shaw[1], C. Harwood[2], L. J. Steggles[1] and A. Wipat[1,*]

[1]School of Computing Science, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, UK and [2]School of Cell and Molecular Biosciences, Medical School, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK

Received on January 22, 2004; revised and accepted on July 1, 2004
Advance Access publication July 9, 2004

ABSTRACT

Summary: SARGE is a tool for creating, visualizing and manipulating a putative genetic network from time series microarray data. The tool assigns potential edges through time-lagged correlation, incorporates a clustering mechanism, an interactive visual graph representation and employs simulated annealing for network optimization.

Availability: The application is available as a .jar file from http://www.bioinformatics.cs.ncl.ac.uk/sarge/index.html

Contact: anil.wipat@ncl.ac.uk

INTRODUCTION

Microarray technology now allows researchers to obtain genome-wide expression levels under varying conditions or times (Schena *et al*., 1995). The next challenge is to interpret the data in order to determine how the products of the genes modify each others expression.

The ultimate goal of most of the studies is to construct and model the complete transcriptional regulatory network of an organism. Time series data promise to be particularly valuable in this respect as they describe transcriptional flux throughout a developmental process, or in response to environmental stimuli. Such datasets are now available for a number of organisms (Cho *et al*., 1998). Analysis of time series transcription data allows the serial regulation of transcriptional activators and repressors to be inferred by reference to gene or gene cluster expression profiles.

AIMS, ASSUMPTIONS AND LIMITATIONS

SARGE (Simulated Annealing to Realize GEnetic networks) aims to provide a system to efficiently generate, visualize and manipulate a putative genetic network based on time series microarray data. SARGE is implemented in Java and is thus platform independent.

The central features of SARGE are:

- A clustering algorithm.
- Linkage of the clusters using time-lagged correlation.
- Optimization of the genetic network via simulated annealing.
- An interactive display of the network (Fig. 1).
- Configurable features.

SARGE seeks to establish transcriptional network by first clustering genes and then calculating possible links between clusters using correlation analysis. The resulting network is optimized using simulated annealing by applying some simple assumptions about gene expression. Our approach extends and improves that described by Chen *et al*. (2001). In particular, SARGE employs correlation-based clustering and link assignment (see Program function section). SARGE provides an interactive interface for browsing and modifying the output. In common with Chen *et al*. (2001), we assume that a gene or co-regulated cluster of genes cannot act as a repressor and activator simultaneously. That is, a gene or a cluster has an effect either as an activator, an inhibitor or has no regulatory effect on other genes or clusters. It is also assumed that if the expression patterns of genes are similar then they can be considered part of the same regulatory cluster. If two clusters are correlated with a time lag then an activation or inhibition link can be assigned between them.

We appreciate that such assumptions are simplistic and do not completely reflect our understanding of regulatory network function. For example, 22% of *Escherichia coli* regulators are assumed to possess a dual regulatory role (Perez-Rueda and Collado-Vides, 2000). However, their application to the network optimization problem allows an approximate network structure to be efficiently generated and visualized from a large experimental dataset. Thus, SARGE aims to give a first approximation of a regulatory network, or alternatively suggest some novel links in a network. The addition of features for user-driven editing of the derived graph allows the user to

*To whom correspondence should be addressed.

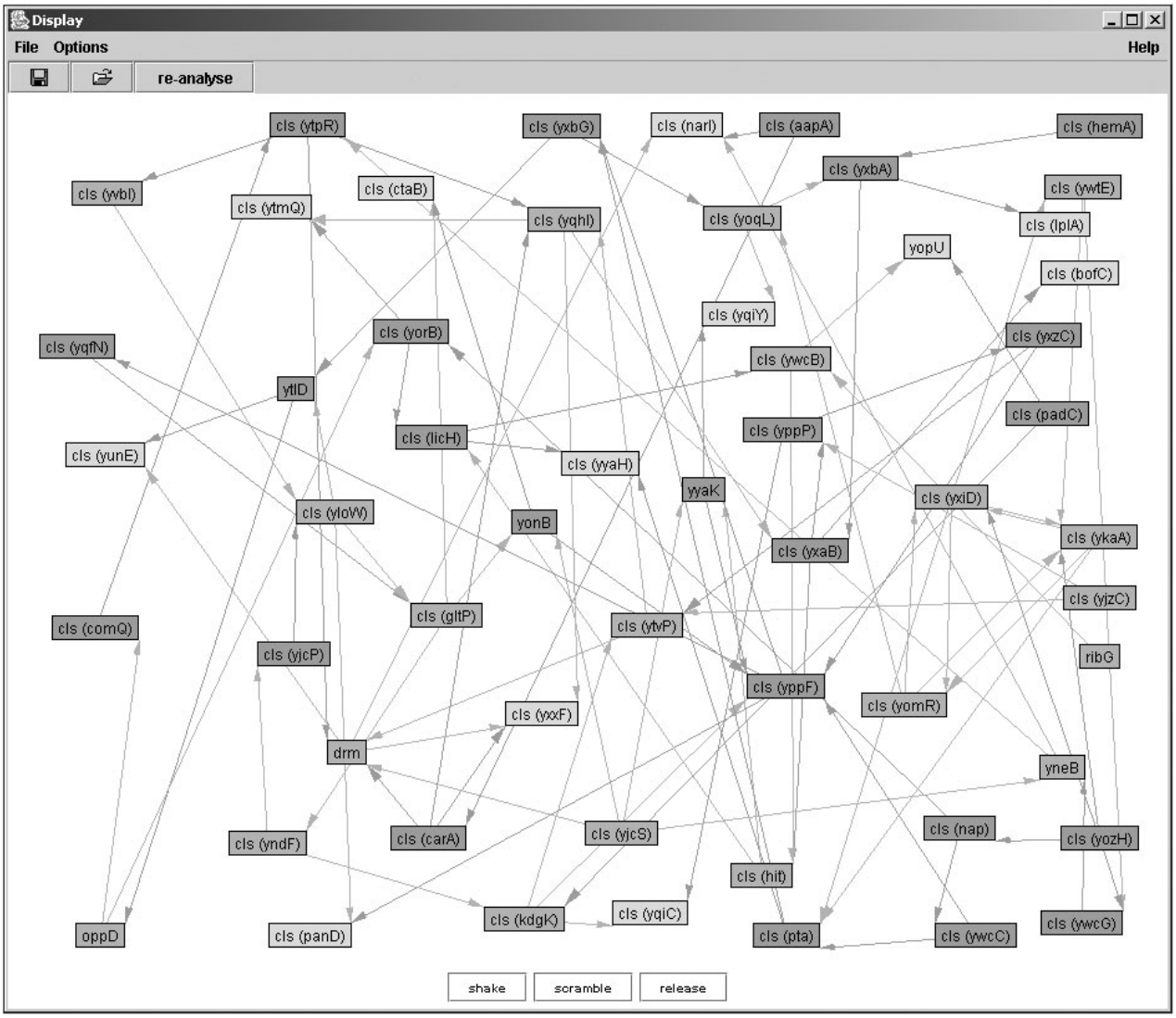

Fig. 1. Screen shot of a putative genetic network generated by SARGE from a subset of *Bacillus subtilis* time series microarray data (unpublished data).

correct any misplaced assumptions and to tailor the network according to their expert knowledge.

PROGRAM FUNCTION

When starting a new session the user is asked to specify a file containing the gene expression dataset, which can be presented in a number of standard text file configurations as described in the help feature of the program. SARGE then clusters the data on the basis of average link method of clustering. If the time-lagged correlation of two expression profiles is more significant they are prevented from being clustered. This prevents potential regulatory links being lost. Alternatively, the input data may be pre-clustered enabling custom clustering algorithms to be applied. In either case, gene/cluster function data may be included in the input dataset for display on the subsequent graph.

After clustering, potential activation or inhibitory relationships are assigned between genes or gene clusters by examining the time-lagged correlation (Kato *et al.*, 2001). This technique uses the standard Pearson correlation, except that one expression profile is lagged or moved back by one or more time points. For example, cluster A is considered to be a potential activator of cluster B, if the correlation between cluster As data lagged one time unit and cluster Bs time data is significant and has a positive Pearson's correlation coefficient (r). Similarly, if r has a negative value then an inhibitory edge is assigned.

After identifying all potential relationships between nodes, the resulting graph is optimized to determine which of the

potential relationships should be kept. This involves assigning one of three values, activator, inhibitor or neutral to each node. If a node is assigned as an activator (respectively inhibitor) only activation (respectively inhibition) edges that originate from the node are considered valid. No edges can originate from a neutral node. SARGE aims to give each node an assignment which best satisfies the following assumptions:

- There is to be a minimum number of regulatory elements.
- There is to be a maximum number of clusters that are both activated and inhibited.
- Valid links are maximized.
- Invalid links are minimized.

The above can be formulated as an optimization problem by associating a cost with each set of assignments and searching for the lowest cost. It turns out that if there are n clusters there are 3^n possible sets of assignments. In fact, this problem has been shown to be NP-complete [non-deterministic polynomial time complete, see Chen *et al*. (2001) for a proof]. Thus the problem is not solvable for all but the smallest n, e.g. with 50 clusters in a genetic network we must explore 3^{50} possible states, an extremely large number.

An appropriate approximation technique is needed which will find a near optimal solution to the problem. SARGE uses the simulated annealing (Kirkpatrick *et al*., 1983) optimization technique to gain a near optimal approximation of the solution to the problem. Simulated annealing is a probabilistic method based on the physical annealing technique of cooling a substance in stages, allowing the substance to reach thermal equilibrium at each stage. The algorithm proceeds as follows. Initially, the nodes are assigned random regulatory values. The annealing algorithm randomly picks a node in the graph and assigns, at random one of the three regulatory roles. The cost value is then calculated for this new set of assignments. Based on the temperature the new cost is accepted or rejected using a probabilistic formula. This allows a new set of assignments with a higher cost to be accepted at high temperatures, preventing the graph becoming trapped in a local minima. This process is repeated, with the temperature being systematically lowered as the algorithm proceeds, until an equilibrium state is reached. This indicates that a near optimal solution has been obtained.

Once the analysis is complete, the optimized graph is displayed using a dynamic graph layout algorithm (Fig. 1). This dynamic approach allows the topology of the graph to be quickly visualized and the graph layout to be manually adjusted. Users can browse the network, nodes may also be moved around and pinned down to give the optimal layout. Clusters may be renamed and their contents maybe swapped with those of another. If the cluster contents are changed, the user can re-run the simulated annealing and linking parts of the algorithm. Repeating these steps iteratively will allow the user to combine their 'perceived knowledge' of the regulatory network together with information generated from the transcriptome data. The result should be a more biologically meaningful network structure than that generated in a completely automated fashion.

SARGE has been applied and evaluated against a number of gene expression sets, including *Saccharomyces cerevisiae* and *B.subtilis* gene expression data (see website for more details).

ACKNOWLEDGEMENTS

This work was partly funded by a BBRSC studentship to O.J.S. (02/B1/X/08298). We would like to thank the reviewers for their insightful comments.

REFERENCES

Chen,T., Filkov,V. and Skiena,S.S. (2001) Identifying gene regulatory networks from experimental data. *Parallel Comput.*, **27**, 141–162.

Cho,R., Campbell,M., Winzeler,E., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T., Gabrielian,A., Landsman,D., Lockhart,D. and Davis,R. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

Kato,M., Tsunoda,T. and Takagi,T. (2001) Lag analysis of genetic networks in the cell cycle of budding yeast. *Genome Inform.*, **12**, 266–267.

Kirkpatrick,S., Gelatt,C.D.,Jr and Vecchi,M.P. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Perez-Rueda,E. and Collado-Vides,J. (2000) The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 1838–1847.

Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitive monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.