

GENIUS: a new tool for gene networks visualization

Paolo Ciccarese^{a,b}, Stefano Mazzocchi^c, Fulvia Ferrazzi^{a,b}, Lucia Sacchi^{a,b}

^a Dipartimento di Informatica e Sistemistica, Università degli Studi di Pavia, Italy

^b Laboratory for Medical Informatics, Pavia, Italy

^c Massachusetts Institute of Technology, Digital Libraries Research Group, Cambridge, MA, USA

Abstract

GENIUS is a graphical software tool for visualizing genetic networks composed by a high number of genes. It accepts as input a matrix summarizing gene relationships inferred by means of reverse engineering methods. GENIUS offers two visualization modalities (the Agorà style and the TouchGraph style), that can be exploited in a complementary way. It is thus possible to obtain a clear and easily customizable view of one or more networks of interest which allows to simplify the exploration of the inferred relationships. GENIUS is in a constant development: the algorithms on which the network visualization is based are in fact very flexible, thus offering the possibility to extend the features provided by the tool.

Introduction

The problem of inferring gene regulatory networks is attracting an increasing interest in the bioinformatics community. Various methods for gene network reconstruction (often called “reverse engineering methods”) have been proposed in the literature [1]. Among these, methods based on Boolean networks, on Bayesian networks and on differential equations. In many cases such methods do not offer a tool for visualizing the inferred networks, or, if they provide one, this tool doesn’t allow the user to easily manipulate and explore the graph. The software presented here tries to overcome this problem, offering a clear and easily customizable visualization of gene networks.

Materials and Methods

GENIUS has been developed with the idea of giving a useful visual representation of genetic networks, even composed by a high number of genes, that is in the order of some hundreds. Reverse engineering methods usually identify, for each of the analyzed genes, those other genes which determine its expression level and can thus be considered as its “regulators”. It is then possible to represent the inferred connections between genes in matrix form. In the most general case, given n analyzed genes, a $(n \times n)$ matrix A can be constructed such that $a_{ij}=1$ if a connection between genes i and j exists (that is gene

i belongs to the set of “regulators” of gene j), while $a_{ij}=0$ if no connection is present. An example is shown in Figure 1.

	Gene_1	Gene_2	Gene_3	Gene_4	Gene_5	Gene_6
Gene_1	0	1	1	1	1	1
Gene_2	0	0	0	0	0	0
Gene_3	0	1	0	0	0	0
Gene_4	1	1	1	0	1	1
Gene_5	1	1	1	1	0	1
Gene_6	0	1	0	0	0	0

Figure 1 - Example of input data matrix for GENIUS. When a number 1 is present in position (i, j) this means that a connection between genes i and j exists. It is possible to think of this connection as the effect of a regulatory action of one gene on another. In the example shown, ‘Gene_1’ regulates ‘Gene_2’, ‘Gene_3’, ‘Gene_4’, ‘Gene_5’ and ‘Gene_6’.

GENIUS visualizes the network representing genes as nodes and interactions between them as edges. It is possible to choose among two different visualizations, that will both be briefly presented in the following part of this section. Such visualizations are complementary, in fact, the Agorà view clearly shows clusters of nodes, while, the spring embedding, TouchGraph view optimally positions genes using methods based on the connection strength [2].

Agorà style

This type of visualization was originally developed to represent virtual communities shapes [3]. Therefore, it was explicitly designed to visualize a huge amount of nodes and relationships between them. The algorithm used is based on the assumption that every individual can be treated exactly the same. Even if this is a strong assumption, it allows to simplify a lot the mathematical model underlying the network representation.

The simulation paradigm is the “private space”: every individual has the perception of a private space that surrounds him/her and his/her level of comfort is reduced if this space is violated by living entities (humans beings or animals) that are not recognized (this can be easily identified as an instinctive protection system). The mathematical model of this private space is based on the use of a repulsive force field and a very basic attractive force field. The repulsive force tends to infinity when the distance between objects is 0 and then rapidly decreases to 0 on a short distance, while the attractive force

starts from 0 and linearly increases toward infinity. Although in the model the relations between nodes are directed, the relationships showed by the original Agorà view were not oriented: the tool was in fact born to process e-mails of a virtual community and, regarding e-mails, relationships are often symmetric. In the case of genes, instead, it is often useful to show directions to gene relationships. Thus, the tool has been extended in order to explicitly visualize oriented links between genes: a connection between two genes is in fact represented by an arrow which starts from one of them (the one identified as the regulator) and finishes into the other (the regulated gene).

TouchGraph style

TouchGraph [4] provides a useful and easy way to visualize networks of interrelated information. Networks are rendered as interactive graphs, which lend themselves to a variety of transformations. TouchGraph allows the user to navigate through large networks, and to explore different ways of arranging the network components on the screen. The TouchGraph view is particularly useful to show relationships between nodes characterized by a certain maximum level of “locality”, defined as the number of edges in the minimum-length path connecting them in the graph.

In the following paragraph we will present how to exploit some of the features provided by GENIUS.

Example of network visualization

We used GENIUS to visualize results obtained using different methods for gene network reconstruction. The example presented here refers to the analysis of an expression data set already analyzed in [5] and relative to the response of human fibroblasts to serum addition. cDNA microarrays were used to measure the relative expression level of serum-stimulated fibroblasts with respect to non-treated ones, at 11 different time points ranging from 15 min to 24 hours.

Iyer et al. identified a subset of 517 genes whose expression level changed significantly in response to serum and performed a cluster analysis on these genes. This data set is publicly available from the web site <http://genome-www.stanford.edu/serum>, which complements the paper published by Iyer et al. We have retrieved it and used a modified version of the Reveal algorithm presented in [6] to analyze temporal profiles relative to the 517 genes. The Reveal algorithm looks for the set of the regulators of each gene x , defined as the minimal set of input genes that can univocally explain the behavior of the output gene x . The algorithm works on data discretized into 2 levels and it is based on the use of Entropy and Mutual Information scores: if for two genes x and y , their Mutual Information is equal to the Entropy of gene x , this means that y univocally determines x . In other words each expression value of y always corresponds to the same expression value of x . We have extended and implemented the algorithm in the case in which gene expression data are discretized into 3 levels. These levels are indicated as -1 , 0 and $+1$, respectively referring to under-expression, equal expression and over-expression of genes coming from serum stimulated cells with respect to expression values of the same genes measured using non-stimulated cells. At the beginning

of the analysis we have excluded the genes characterized by a flat profile, that is the genes whose discretized expression level was constant in each time point. We then have assembled together the genes having the same discretized profiles, identifying 179 groups.

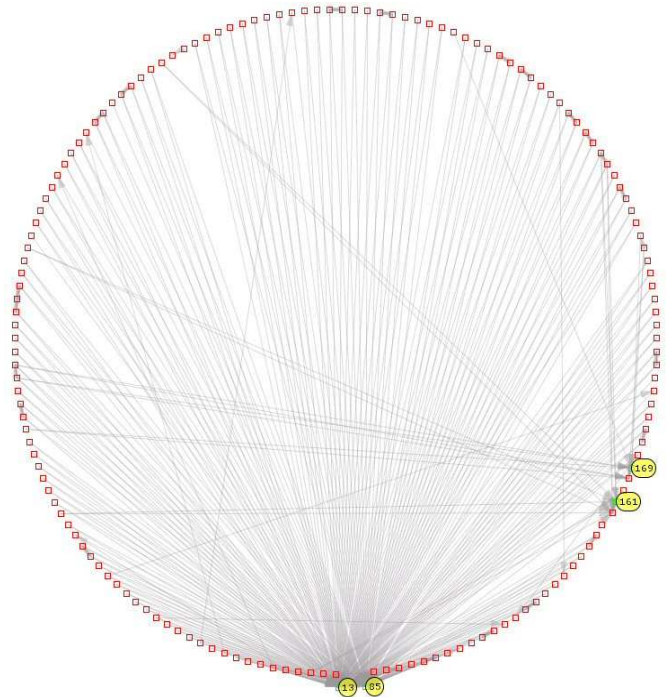


Figure 2 - Visualization offered by GENIUS Agorà view: entire network inferred from the serum response data. Nodes whose label is highlighted are the so called “hubs”, that is the most highly connected pseudo-genes (indicated with the numbers 13, 85, 161, 169).

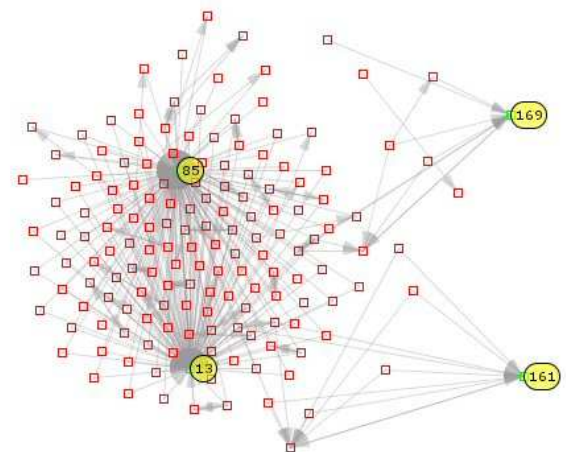


Figure 3 - Visualization offered by GENIUS Agorà view: entire network inferred from the serum response data as displayed after the automatic process which rearranges nodes according to their force of attraction/repulsion. It is clear

how the “hubs” (numbers 13, 85, 161, 169) are the nodes towards which the others have been attracted.

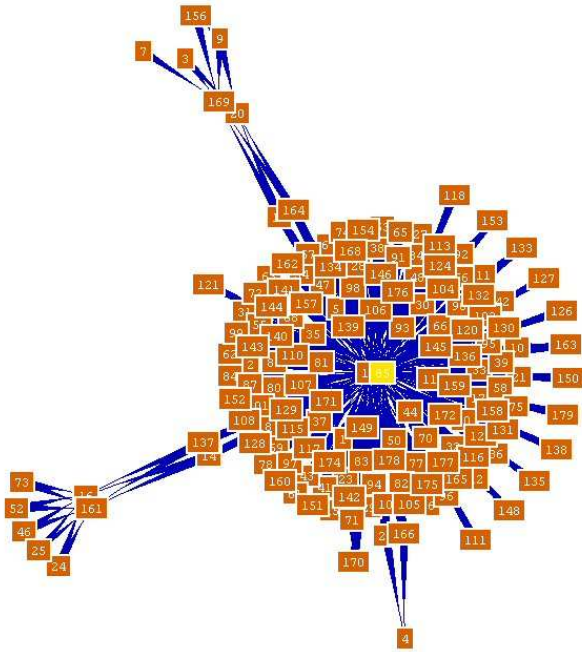


Figure 4 - Visualization offered by GENIUS TouchGraph view: entire network inferred from the serum response data. Three hubs are clearly detectable also here (numbers 85, 161, 169) while the other (number 13) is hidden (by number 85). However, even this hub can be put into evidence by means of the “zoom” function.

These groups are here called “pseudo-genes”, as each one of them can correspond to more than one gene. The modified Reveal algorithm has thus been applied on these pseudo-genes. Due to their high number, the analysis was limited to pair wise relationships, as in classical cluster analysis. The output provided by the algorithm can be visualized as a network in which nodes representing pseudo-genes are connected by edges starting from the node referring to the regulator pseudo-gene and pointing to the node corresponding to the regulated pseudo-gene.

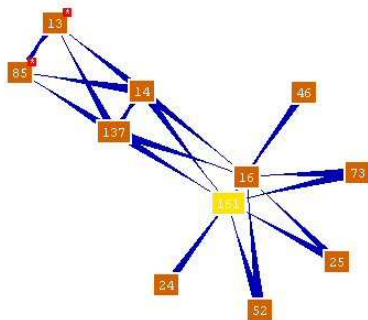


Figure 5 - Visualization offered by GENIUS TouchGraph view: by means of the “locality” function it is possible to

highlight only the node of interest (in this case number 161) and the ones connected to it at a certain maximum distance (in this example equal to 2). The distance between two nodes is calculated as the number of edges between them.

Figures 2 and 3 show the entire network inferred from the data, visualized using Agorà. In Figure 2 it is presented how Agorà initially displays the network in a circular form. The nodes whose labels are highlighted correspond to the most connected pseudo-genes. These highly connected nodes are the so-called “hubs” and their presence in genetic networks has been pointed out in various studies, such as the one presented in [7]. The hubs can be more easily detected by resorting to the automatic arrangement offered by Agorà. One of the main advantages of the Agorà algorithm is in fact the ability to automatically arrange the genes in the network according to their force of attraction/repulsion. The result of this automatic process is visible in Figure 3: here it is clear how the hubs have functioned as “attractors” of the other nodes. If the user would like to inspect one of the hubs and its neighbours at a deeper level, he could manually manipulate the network, moving nodes so as to isolate the genes belonging to the desired group. Another possibility is to resort to the visualization provided by TouchGraph. Figure 4 shows how TouchGraph depicts the entire network. Some of the hubs are clearly detectable, while others cannot be easily seen. However it is possible to search for one of the genes of interest and then use the “zoom” and “locality” functions provided by TouchGraph. The former simply allows to enlarge a desired portion of the graph, while the latter allows to highlight the nodes at a certain maximum “distance” from the node of interest (the distance between two nodes is measured as the number of edges connecting them in the graph). Figure 5 shows an example of visualization obtained choosing a level of locality equal to 2.

Conclusion and future developments

The example discussed in the previous paragraph has pointed out the clearness and easiness of manipulation/customization of the visualizations provided by GENIUS. However this example has not highlighted all the potentials offered by the tool.

First of all, the input matrix accepted by GENIUS must not necessarily be composed of elements representing an on/off relationship between genes (all 1s and 0s entries). In fact GENIUS can accept as an input also a matrix whose elements give information about the strength of the connection between them. The elements a_{ij} can for example be equal either to the correlation calculated between the data vectors for genes i and j , or to their mutual information. In these cases, the network visualization provided by GENIUS will explicitly take into account the differences in the pair wise connections between genes, producing a visualization in which couples of genes for which the corresponding a_{ij} is higher are more closely connected. An example is visible in Figures 6 and 7 (obtained using Agorà), which refer to a very simple case with only four genes. In the network shown in Figure 6 all genes have only one connection and all connections between genes have an

equal weight, while in Figure 7 each gene still has one connection but these connections are characterized by different weights, corresponding to the correlations between couples of genes, measured on the basis of their expression vectors. In this figure it is clearly visible how the arrangement of the nodes has changed, such that the higher the correlation between two genes, the shorter the distance between the nodes representing them.

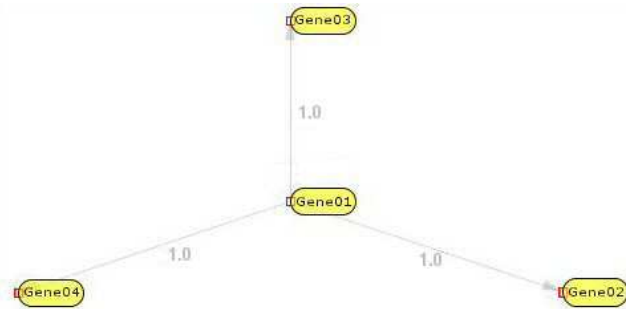


Figure 6 - Example of a very simple network composed by 4 genes. The edges between nodes have all the same length because each gene is connected with only another gene and the weights associated with the pair wise connections between them are all equal to 1.

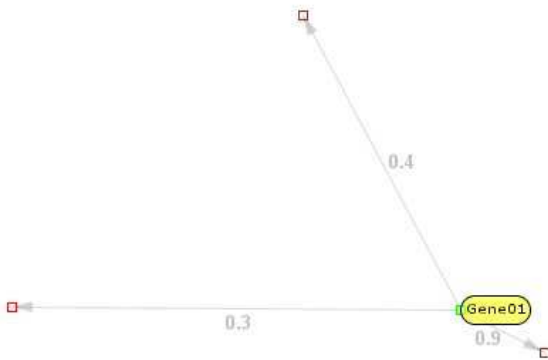


Figure 7 - Example of a network involving the same 4 genes presented in Figure 6. In this case the lengths of the edges between nodes are different because each gene is still connected with only another one but the weights associated with the pair wise connections vary. In particular in this case the weights have been set equal to the correlations measured between the expression vectors of the corresponding genes.

Moreover, another interesting feature offered by GENIUS is the possibility to display together two or more networks inferred using different datasets which involve partially overlapping groups of genes. As depicted in Figure 8, it is possible to obtain a view of the networks which clearly highlights the genes in common. In addition to these features, many others

are going to be added. GENIUS is in fact in a constant development and we are working on various extensions of the functionalities it offers. First of all an annotation tool could be added, using functional categories based on Gene Ontology. In particular, it would be interesting to allow the user to choose according to which taxonomy genes must be annotated (Biological process, Molecular Function or Cellular Component) and at which level of specificity/coverage, similarly to what is implemented in the annotation tool provided by David [8].

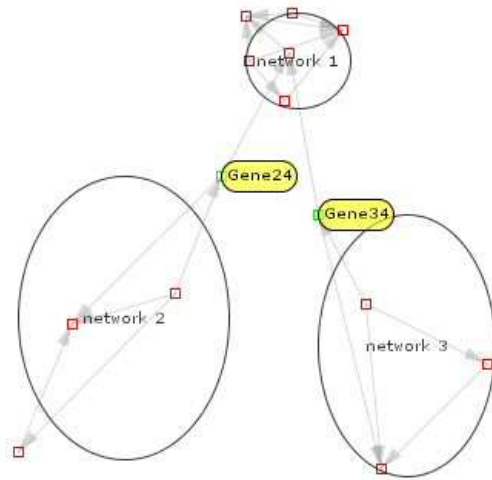


Figure 8 -. Example of a visualization of three different networks which involve partially overlapping sets of genes. In particular, 'Gene 24' belongs to both network 1 and network 2, while 'Gene34' is part of both network 1 and network 3.

Another interesting extension of GENIUS would be the ability to visualize the dynamic of a gene network, that is how the network evolves during time. In functional genomics the study of gene expression time series is in fact gaining increasing importance and in these cases it is interesting to analyze not only how the expression level of gene x at time t+1 is affected by the levels of other genes at the same time point (instantaneous relationships), but also how the level of gene x at time t+1 is influenced by the levels of genes at time t (delayed relationships).

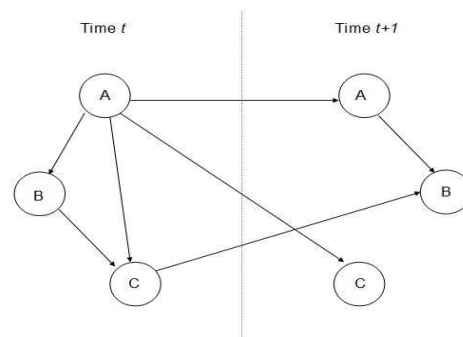


Figure 9 - Example of a dynamic network representing both instantaneous and delayed relationships between 3 genes (A, B, C). In this example the expression level of gene B at time $t+1$ depends on the expression level of gene A at the same time point but also on the level of gene C at time t .

e-mail: paolo@aim.unipv.it

Figure 9 shows an example of a network in which both types of gene relationships are represented. Currently available network visualization tools are not able to offer a clear representation of the dynamic of a network, while the mathematical models underlying the visualization provided by GENIUS are instead very flexible and could be extended to the case in which also delayed relationships between genes have to be represented.

Acknowledgments

We thank Professor Riccardo Bellazzi for his support.

References

- [1] De Jong H. Modeling and simulation of genetic regulatory systems: A literature Review. *Journal of Computational Biology* 2002; 9 (1): 67-103
- [2] Quinn Jr., N. R.; Breuer, M. A.: A Force Directed Component Placement Procedure for Printed Circuit Boards, *IEEE Trans. on Circuits and Systems*, CAS-26(6), pp. 377-388, 1979.
- [3] Mazzocchi S. Apache Agorà 1.2. www.apache.org/~stefano/agora/
- [4] TouchGraph. Software available at: <http://www.touchgraph.com/index.html>
- [5] Iyer V. R. et al. (1999): The transcriptional program in the response of human fibroblasts to serum. *Science*: 283: 83-87
- [6] Liang S, Fuhrman S, Somogyi R. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symp. Biocomp.* 1998: 98 (3): 18-29.
- [7] Shaw S. Evidence of Scale-Free Topology and Dynamics in Gene Regulatory Networks. In *Proceedings of the ISCA 12th International Conference on Intelligent and Adaptive Systems and Software Engineering 2003*: 37-40.
- [8] Glynn Dennis Jr., Brad T. Sherman, Douglas A. Hosack, Jun Yang, Michael W. Baseler, H. Clifford Lane, Richard A. Lempicki. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology* 2003: 4(5): 3.

Address for correspondence

Paolo Ciccarese.

Dipartimento di Informatica e Sistemistica, Università degli Studi di Pavia, via Ferrata 1, 27100 Pavia, Italy