Discovering network relations in big time series with application to bioinformatics

Mariano Rubiolo[1,2], Diego Milone[2], Georgina Stegmayer[2]

[1]CIDISI, UTN-FRSF, Lavaise 610.
[2]sinc(*i*), FICH-UNL/CONICET, C. Universitaria - Pje El Pozo, R. N. 168
(3000) Santa Fe - Argentina
mrubiolo@santafe-conicet.gov.ar

Big Data concerns large-volume, complex and growing data sets, with multiple and autonomous sources. It is now rapidly expanding in all science and engineering domains [1]. Time series represent an important class of big data that can be obtained from several applications, such as medicine (electrocardiogram), environmental (daily temperature), financial (weekly sales totals, and prices of mutual funds and stocks) [2], as well as from many areas, such as social-networks and biology.

Bioinformatics seeks to provide tools and analyses that facilitate understanding of living systems, by analyzing and correlating biological information. In particular, as increasingly large amounts of genes information have become available in the last years, more efficient algorithms for dealing with such big data in genomics are required [3]. There is an increasing interest in this field for the discovery of the network of regulations among a group of genes, named Gene Regulation Networks (GRN) [4], by analyzing the genes expression profiles represented as time-series.

In [5] it has been proposed the GRNNminer method, which allows discovering the subyacent GRN among a group of genes, through the proper modeling of the temporal dynamics of the gene expression profiles with artificial neural networks. However, it implies building and training a pool of neural models for each possible gen-to-gen relationship, which derives in executing a very large set of experiments with $O(n^2)$ order, where n is the total of involved genes. This work presents a proposal for dramatically reducing such experiments number to $O((n/k)^2)$ when big time-series is involved for reconstructing a GRN from such data, by previously clustering genes profiles in k groups using self-organizing maps (SOM) [6]. This way, the GRNNminer can be applied over smaller sets of time-series, only those appearing in the same cluster.

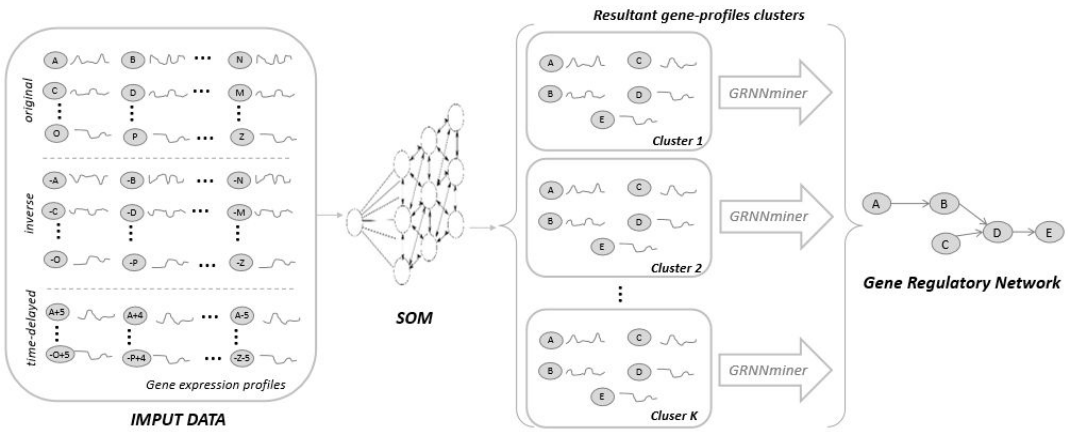

Fig. 1. Big time-series processing for hidden network reconstruction.

The proposal can be seen in Figure 1. An input dataset is formed by the original time series measures. The network discovery in this original dataset implies modeling all possible combinations between variables, which generates a quadratic complexity. In order to avoid processing the complete input data for finding hidden relations, the data is clustered. This allows more efficiently looking for networks of relations. Inverse and time-delayed versions of the original data according to a sliding window, are added to the original set, in order to facilitate grouping data with similar (or somehow related) temporal dynamics. Therefore, the new modified data set is used to train a SOM, which is able to group similar data in a single map-unit, or in nearby ones. It is expected that, with the modified versions of the original time series, those gene data measures that are related between them due to some sort of regulation, will be grouped together in the same neuron. After training, each map-unit in the SOM is evaluated in order to identify which original time-series data are clustered together. This way, regulation connections will be looked for only among data that have been clustered together.

Table 1. Results of clustering 10 target genes from a 1000 gene profiles data set.

SOM	0,5*p*	0,75*p*	*p*	1,25*p*	1,5*p*
map size	(21,17)	(26,21)	(30,24)	(33,27)	**(36,30)**
extra genes	17	13	1	3	**0**

The proposed approach has proven to significantly reduce the computational cost, because the network is going to be sought in small clusters instead of the whole original dataset. Table 1 shows the results of a study case using an artificial simulated data set with 1000 genes in which a well-known GRN [5] involving only 10 genes is present. The input data set was formed considering the original gene profiles of 200 time points, the inverse, and both delayed versions with a delay ranging from -5 to +5 time steps, which derive in 20000 time series. Five SOM were trained with such data, with a map size determined by Kohonen heuristic [6] ($p = 5\sqrt{q}$, p: map units, q: data samples). Four other variants of this size (0.5p, 0.75p, 1.25p and 1.5p) were considered as well, aiming evaluate the influence of different SOM sizes on the results. Each table row shows corresponding map-size, and the number of extra genes clustered together with the 10 genes of the real GRN.

In all cases, genes belonging to the real target GRN have been clustered together. Even though at some SOM models extra genes are clustered, the reduction of the experiments to be done is very important, ranging from $O(n^2) = O(1000^2)$ in the original dataset to $O(17^2)$ in the worst case (first column), and $O(10^2)$ in the best case (last column, in bold). This means a very significant reduction between 99.71% and 99.9 % of the original computational cost. Once the original dataset is divided into small clusters, the remaining stages of data mining can be easily executed in parallel.

References

1. Wu, X., Zhu, X., Wu, G., Ding, W.: "Data Mining with Big Data". IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO. 1, (2014).
2. Tak-chung Fu,: "A review on times series data mining", Engineering Applications of Artificial Intelligence, Vol. 24, Issue 1, pp. 164-181, (2011).
3. McDonald, E. and Titus Brown, C..: "khmer: Working with Big Data in Bioinformatics". ACM Computing Research Repository (CoRR), pp. 1-18, (2013).
4. Chang, Y.H., Gray, J.W., and Tomlin, C.J.: "Exact reconstruction of gene regulatory networks using compressive sensing", Bioinformatics, (2014).
5. Rubiolo, M., Milone, D., and Stegmayer, G.: "Mining gene regulatory networks by neural modeling of expression time-series.", IEEE Transactions on Computational Biology (2015).
6. T. Kohonen Self-Organizing Maps, Springer-Verlag, (2001).