SUPPLEMENTARY MATERIAL
Comparing association network algorithms for reverse engineering of large scale gene regulatory networks: synthetic vs real data

N. Soranzo†, G. Bianconi‡ and C. Altafini†*

† SISSA-ISAS, International School for Advanced Studies
via Beirut 2-4, 34014 Trieste, Italy
‡Abdus Salam International Center for Theoretical Physics
Strada Costiera 11, 34014 Trieste, Italy

March 23, 2007

The material of this Supplement is divided into 3 Sections:

1. **Synthetic data**: integrates the content of Section 3.1 of the paper.
2. **Influence of sparsity on the predictive power**: compares inference on 2 networks with different sparsity.
3. **Comparing B-spline and Gaussian Kernel in the computation of I**: evaluate how much the matrix I changes with the algorithm chosen.

1 Synthetic data

This Section integrates the results obtained in Section 3.1 of the paper. For both AUC(ROC) and AUC(PvsR), standard deviations (not shown) are around one order of magnitude smaller than the mean values, thus indicating that the repetitions are substantially faithful.

For the random and scale-free networks reconstructed in Fig. 1 of the paper, Fig. S1 reports the average runtimes (over the 10 repetitions) of the various algorithms: R_{C_2} is clearly one order of magnitude slower than the other methods. It must be remarked that for R, R_{C_1}, R_{C_2} we used MATLAB code, while for I, I_C, I_{DPI} C++ code was created (so faster than MATLAB) and $R_{C_{all}}$ was computed under R environment. Notice that I_C grows faster than the other methods with respect to the number of experiments.

The first row of Fig. S2 evaluates the algorithms on a 1000 gene scale-free network. The experiments are of knockout type, with steady-state measurements. It can be seen that all three parameters shown AUC(ROC), AUC(PvsR) and TP for fixed FP are similar to the equivalent ones on Fig. 1 of the paper (second row) for the same value of the ratio m/n.

2 Influence of sparsity on the predictive power

In this Section we show that a sparse network is easier to infer than a dense (or less sparse) one. For this scope we consider two classes of artificial networks of 1000 genes, both of scale-free topology, the first having an average node degree equal to 1.5 and the second equal to 3. The graphs for steady state knockout

*Corresponding author: altafini@sissa.it

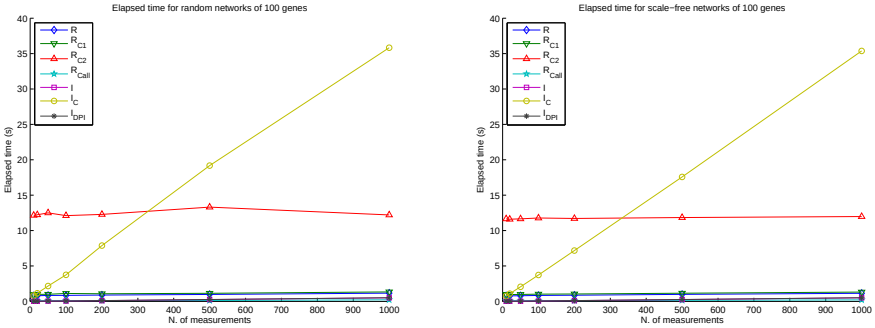

Figure S1: Runtime of the algorithms for the random (left) and scale-free (right) networks of 100 genes shown in Fig. 1 of the paper.

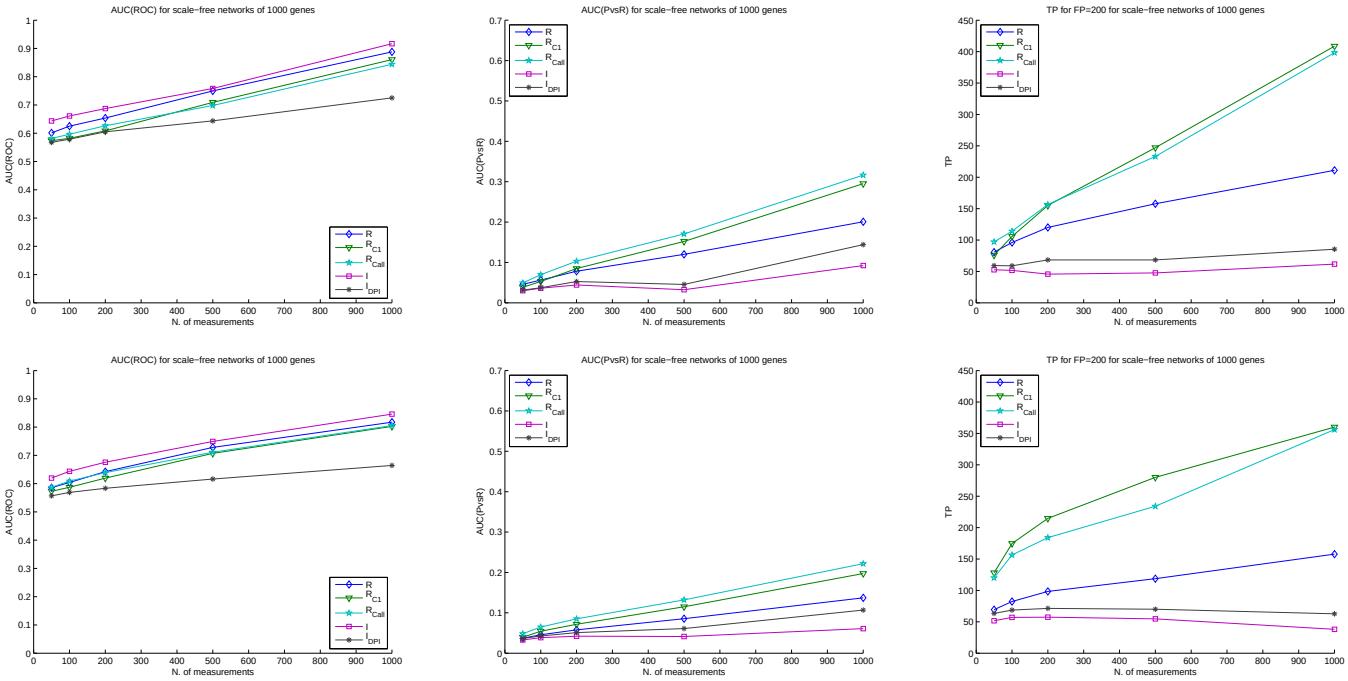

Figure S2: Evaluating the reconstructions via R, R_{C_1}, $R_{C_{\rm all}}$, I and I_{DPI} on 1000 gene artificial networks of scale-free type, for increasing numbers of measurements, all for knockout experiments and all at steady state. First row: average node degree 1.5. Second row: average node degree 3. Left column: AUC(ROC). Central column: AUC(PvsR). Right column: number of TP for a number of FP equal to 200. Values shown are means over 3 repetitions.

experiments are shown in Fig. S2. They clearly show that on the sparser network the inference algorithms are more incisive, having better performances for all the three metrics considered.

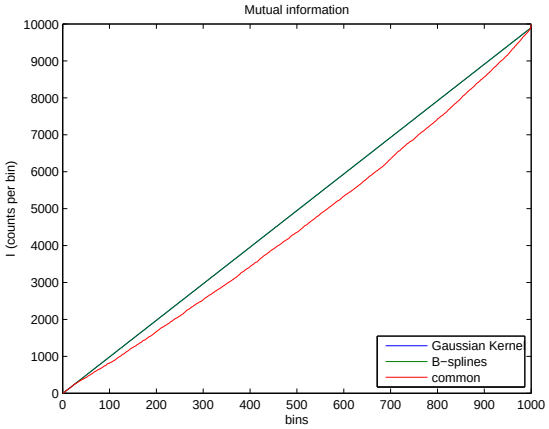

Figure S3: Comparison of I computed via Gaussian Kernel method from [2] and B-spline method used in the paper with 4 bins and spline order 2 for a network of 100 genes and 200 experiments. The elements of the two matrices are sorted and the sorted values divided in 1000 bins. The figure shows the cumulative counts of the values of the sorted elements (y-axis) up to the i-th bin (x-axis). The counts for the two algorithms overlap (by construction), while the number of edges in common differs for less than 10% of the total.

3 Comparing B-spline and Gaussian Kernel in the computation of I

In [2] the computation of MIs is carried out using a Gaussian Kernel method. This is known to be computationally more intense than binning into an histogram, even when the B-spline approach is used [1]. In order to evaluate how much the choice of the algorithm can influence the reconstruction, we compared two MI matrices computed using a Gaussian Kernel estimator (with the routines provided in [3]) and the B-spline approach. A typical result is shown in Fig. S3 for a rather conservative choice of number of bins ($q = 4$) and spline order 2. It can be seen that the two ordering of edges weights always differ for less than 10%.

References

[1] C. O. Daub, R. Steuer, J. Selbig, and S. Kloska. Estimating mutual information using B-spline functions – an improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, 5(1):118, 2004.

[2] A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006.

[3] A. A. Margolin, K. Wang, W. K. Lim, M. Kustagi, I. Nemenman, and A. Califano. Reverse engineering cellular networks. *Nature Protocols*, 1(2):663–672, 2006.