

An integrative reverse engineering algorithm to infer gene networks from multiple sources of data

Vincenzo Belcastro¹, Mukesh Bansal^{1,2,*}, Diego di Bernardo^{1,2}

1. Telethon Institute of Genetics and Medicine, Via P. Castellino, Naples, Italy
2. European School of Molecular Medicine, Naples, Italy

*E-mail: belcastro@tigem.it

Introduction

Information about gene regulatory interactions in different species has already been produced using microarray technology and a variety of other experimental approaches. This information should be properly merged to yield predictions of gene interactions. Here we propose a framework based on a Bayesian version of Mutual Information (MI) to tackle this problem.

Bayesian approach to Mutual Information

We consider two random variables (or genes) $i \in \{1, \dots, r\}$ and $j \in \{1, \dots, s\}$ and an i.i.d. random process with samples (or expression observation) $(i, j) \in \{1, \dots, r\} \times \{1, \dots, s\}$ draw with joint probability π_{ij} . An important measure of the stochastic dependence of i and j is the mutual information

$$I(\pi) = \sum_{i=1}^r \sum_{j=1}^s \pi_{ij} \log \frac{\pi_{ij}}{\pi_{i+} \pi_{+j}}, \quad (1)$$

where $\pi_{i+} = \sum_j \pi_{ij}$ and $\pi_{+j} = \sum_i \pi_{ij}$ are marginal probabilities.

Once we have a sample set (set of microarrays) with n_{ij} outcomes of pair (i, j) the frequency $\hat{\pi}_{ij} := \frac{n_{ij}}{n}$ can be used as a first estimate of the unknown probabilities. This leads to a point (frequency) estimate $I(\hat{\pi}) = \sum_{ij} \frac{n_{ij}}{n} \log \frac{n_{ij}n}{n_{i+}n_{+j}}$ for the mutual information.

In order to include prior knowledge in the estimation of mean and variance of the mutual information, we apply a Bayesian approach instead of the frequentist one ([2]) assuming a prior probability density $p(\pi)$ for the unknown probabilities π_{ij} in order to compute the posterior $p(\pi | \mathbf{n}) \propto p(\pi) \prod_{ij} \pi_{ij}^{n_{ij}}$ (assuming a multinomial distribution among the n_{ij} 's). This allows to compute the posterior probability density of the mutual information:

$$p(I | \mathbf{n}) = \int \delta(I(\pi) - I) p(\pi | \mathbf{n}) d^s \pi. \quad (2)$$

Depending on the distribution one chooses for the prior $p(\pi)$, the posterior density of the mutual information may follow a Dirichlet and its expected value can be computed as:

$$E[I] = \sum_{ij} \frac{n_{ij}}{n} \log \frac{n_{ij}n}{n_{i+}n_{+j}} + \frac{(r-1)(s-1)}{2n} + O(n^{-2}). \quad (3)$$

Where $n_{ij} = n'_{ij} + n''_{ij}$; n'_{ij} counts the number of samples (i, j) and n''_{ij} are pseudo-counts that take into account the prior information. In ([2]) is also derived a formula for the variance.

A uniform prior distribution among observations leads to a Dirichlet distribution. However, when information about the sign of the regulation is known, we show numerically that a non uniform prior should be applied in order to get a higher performance.

A new algorithm for network inference

We developed an algorithm, based on the formula (3), able to merge different sources of information for the purpose of gene regulatory network inference. We applied our algorithm on simulated and real dataset in [1]. Results are show in Figure (1) where we varied the weight assigned to prior information (solid lines) from 0 (no prior) to 1 (high confidence in the prior); dashed lines represent performances where expressions data have been shuffled removing all the information they were carrying. A too high confidence in the prior forces the algorithm to consider only the information carried by the prior discarding that contained into the expressions.

A good choice for the weight of the prior is needed to properly merge expression data with other sources of data. To this end, we computed a *synergy score* in order to measure the synergistic increase in performance due to the use of prior information and expression data. The synergy score is computed using differences of the areas under the PPV-Sensitivity curves for the expression data and randomized data, We show that for a weight of 0.2 for the prior (red curve) the synergy score is 2, thus showing a good compromise between prior information and expression data.

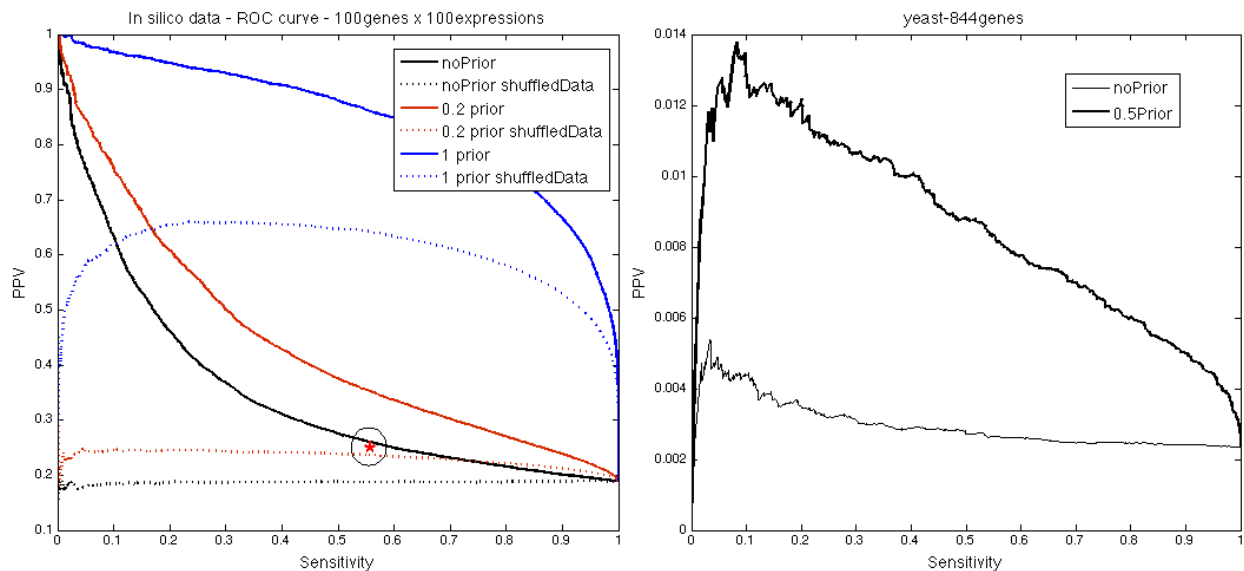


Figure 1: The left plot emphasize the difference between the performance without prior information (black curve) and the one with prior (red curve) compared with random performances without and with information respectively (dashed lines). Red star underline ARACNe ([4]) performance. The right plot show the performance on a real dataset (yeast) with (bold curve) and without prior information obtained from global CHIP-chip data in [3].

Conclusions

The algorithm we presented here is able to integrate information coming from expression data and prior knowledge. This represent the first method based on Mutual Information able to merge difference sources of information.

References

- [1] Mukesh Bansal, Vincenzo Belcastro, Alberto Ambesi-Impiombato, and Diego di Bernardo. How to infer gene networks from expression profiles. *Mol Syst Biol*, 3:78, 2007. Comparative Study.
- [2] Marcus Hutter. Distribution of mutual information. *Advanced in Neuronal Information Processing Systems*, 18:339–406, 2004.
- [3] Tong Ihn Lee, Nicola J Rinaldi, Francois Robert, Duncan T Odom, Ziv Bar-Joseph, Georg K Gerber, Nancy M Hannett, Christopher T Harbison, Craig M Thompson, Itamar Simon, Julia Zeitlinger, Ezra G Jennings, Heather L Murray, D Benjamin Gordon, Bing Ren, John J Wyrick, Jean-Bosco Tagne, Thomas L Volkert, Ernest Fraenkel, David K Gifford, and Richard A Young. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594):799–804, Oct 2002.
- [4] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7, 2006.