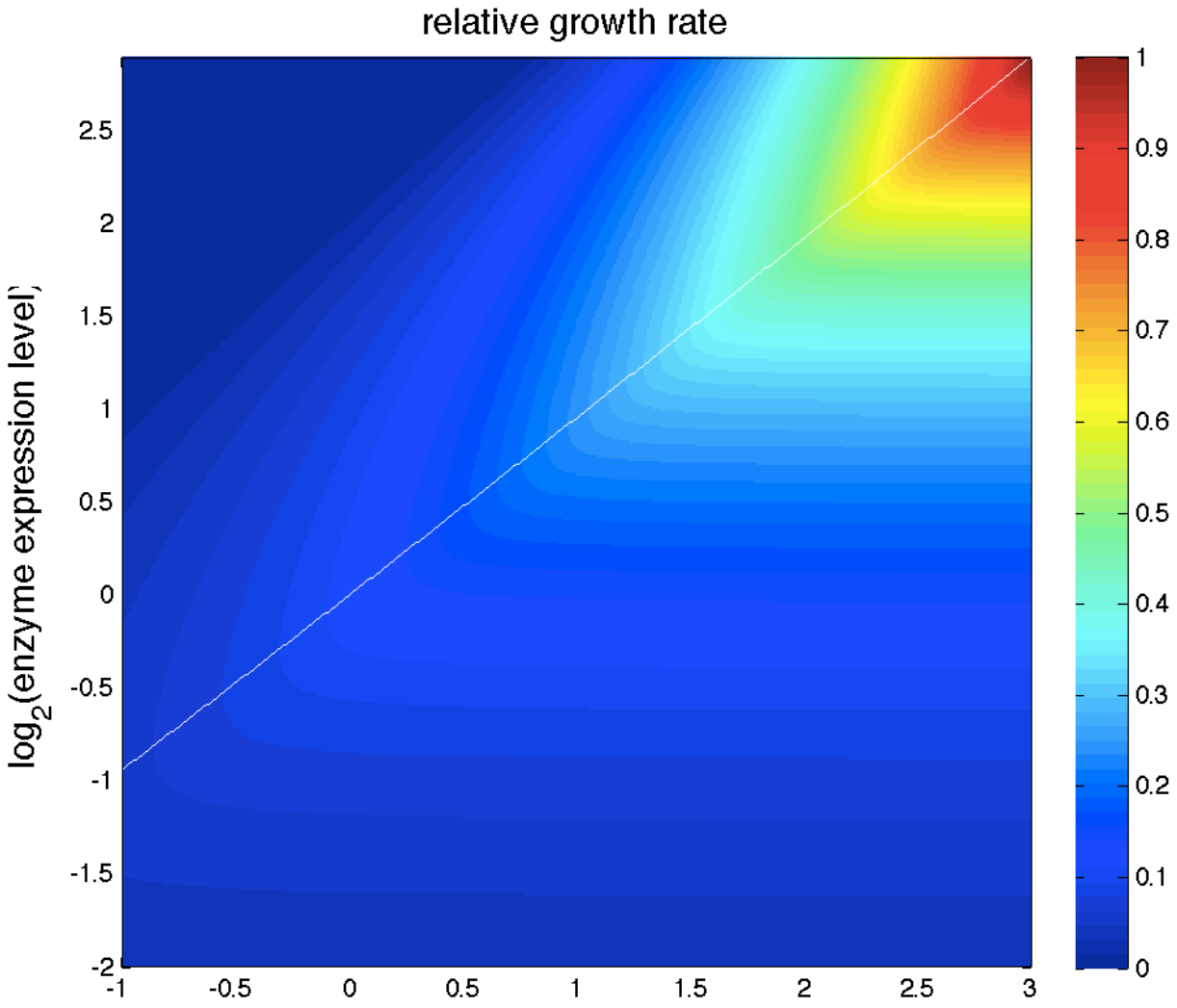
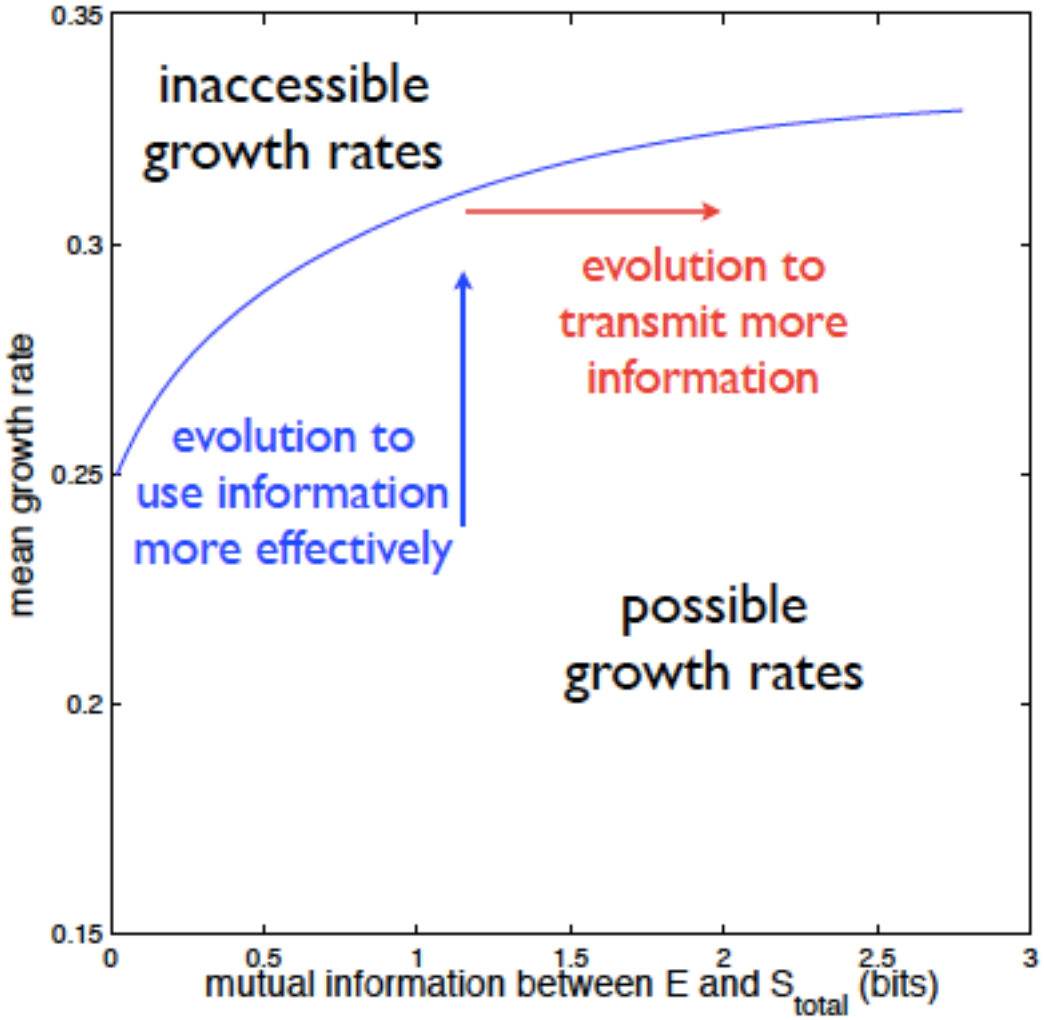
Rate Distortion arguments for bacterial growth (Bialek's book Biophysics, Chapter 6)



Phase space growth rate-mutual information



Data Processing Inequality and Applications to Bioinformatics

$$I(X,Z) \leq I(X,Y)$$

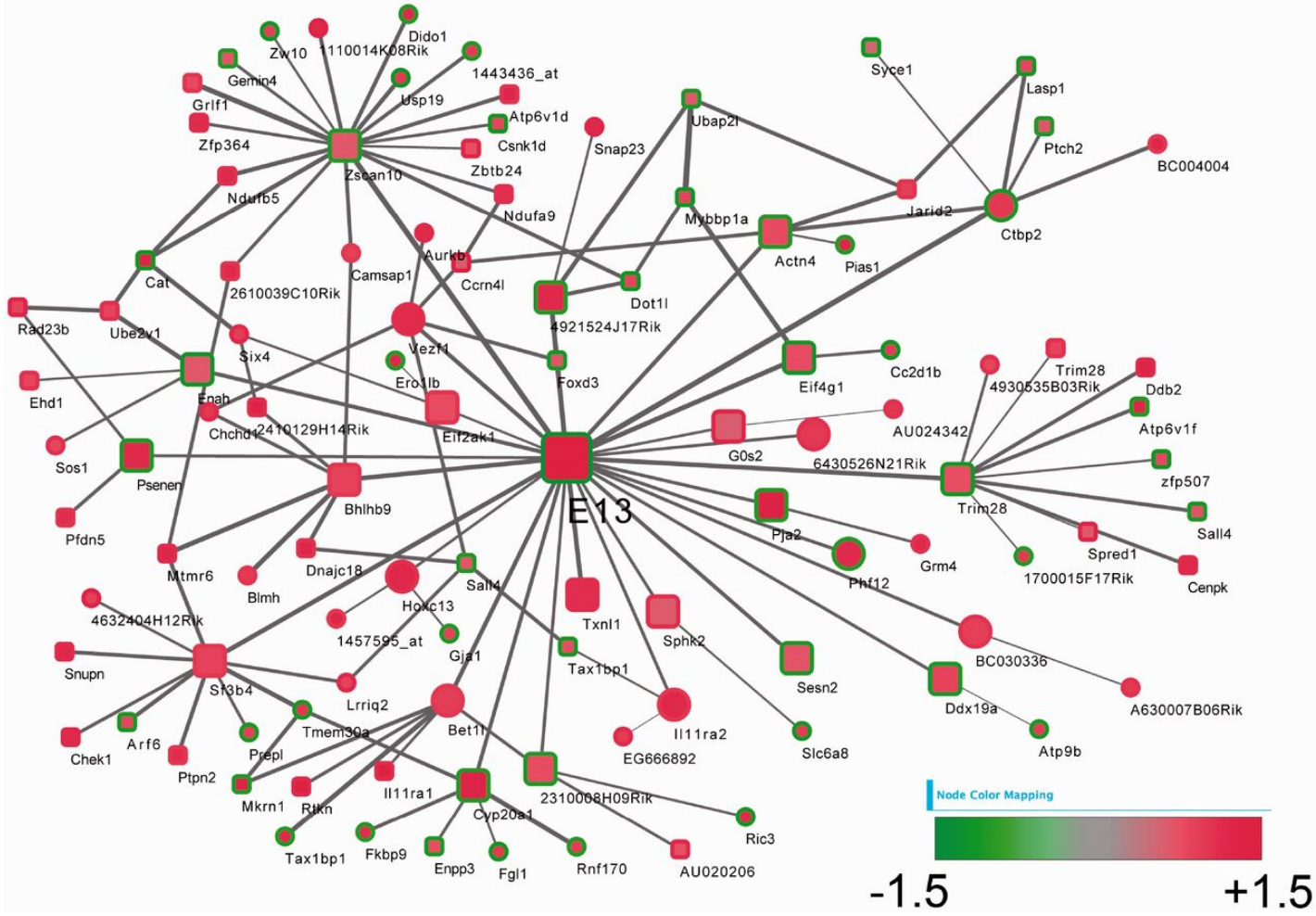
No miracles: if you process data, e.g. Z=f(Y) you cannot create extra information even though you might illustrate it much more clearly

ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context

Adam A Margolin[1,2], Ilya Nemenman[2], Katia Basso[3], Chris Wiggins[2,4], Gustavo Stolovitzky[5], Riccardo Dalla Favera[3] and Andrea Califano*[1,2]

BMC Bioinformatics 2006, **7**(Suppl 1):S7

Reconstruction of interaction networks

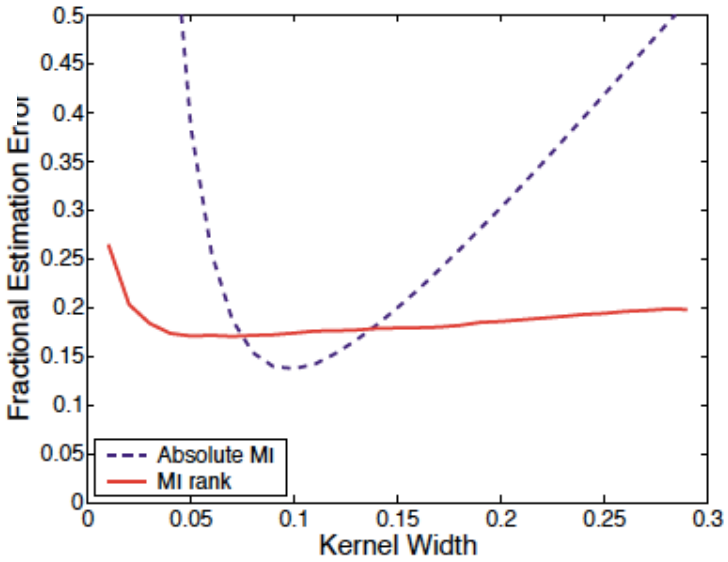


1st step: estimate ranks of MIs

We estimate MI using a computationally efficient Gaussian Kernel estimator [12]. Given a set of two-dimensional measurements, $\vec{z}_i \equiv \{x_i, y_i\}$, $i = 1 \ldots M$, the JPD is approximated as $f(\vec{z}) = 1/M \sum_i h^{-2} G\left(h^{-1}\left|\vec{z} - \vec{z}_i\right|\right)$, where $G(\ldots)$ is the bivariate standard normal density. With $f(x)$ and $f(y)$ being the marginals of $f(\vec{z})$, the MI is:

$$I\left(\{x_i\},\{y_i\}\right) = \frac{1}{M}\sum_i \log \frac{f(x_i, y_i)}{f(x_i) f(y_i)} \qquad (2)$$

Absolute estimates are very sensitive to smoothing, the ranks much less

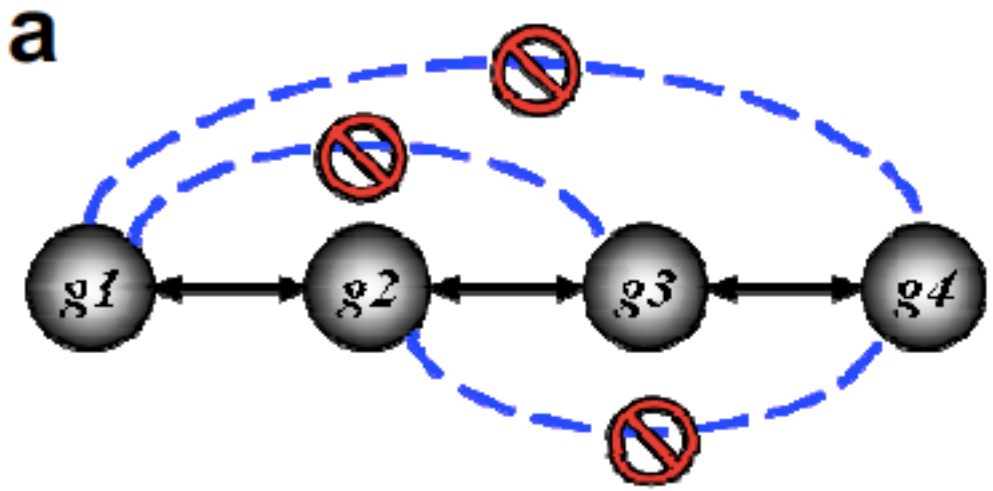


2nd step: threshold MIs

MI are positive and by chance they might take large values. Threshold set at a value such that the probability for that value to happen by chance is low

independent genes cannot be ruled out. To this extent, we randomly shuffle the expression of genes across the various microarray profiles, similar to [6], and evaluate the MI for such manifestly independent genes and assign a p-value, p, to an MI threshold, I_0, by empirically estimating the fraction of the estimates below I_0. This is done for different sample sizes M and for 10^5 gene pairs so that reliable estimates of $I_0(p)$ are produced up to $p = 10^{-4}$.

3rd (crucial) step: pruning links using DPI

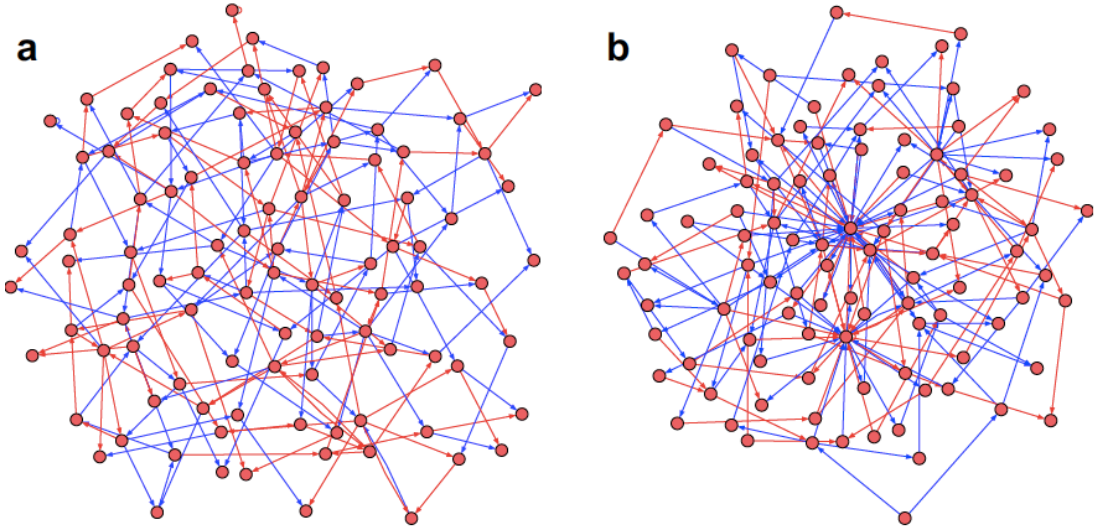


g_1 and g_3 might be strongly correlated and have high MI because of their indirect coupling via g_2. Their link is not necessarily eliminated by the threshold. Conversely, it is eliminated by DPI as

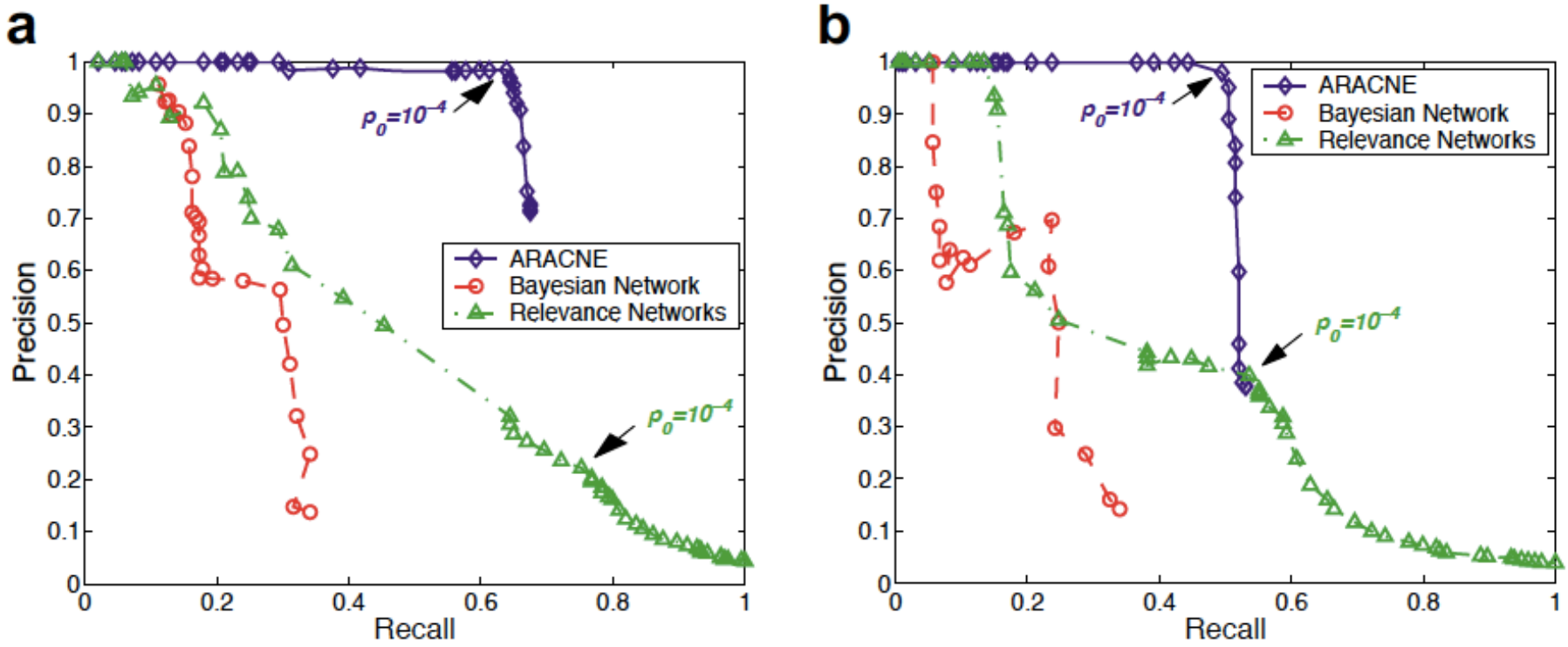
$$I(g_1, g_3) \leq \min\left[I(g_1, g_2), I(g_2, g_3)\right]$$

Testing the algorithm

We benchmark the three algorithms using synthetic transcriptional networks proposed by Mendes et al. [16] as a platform for comparison of reverse engineering algorithms. These networks consist of 100 genes and 200 interactions organized either in an Erdös-Rényi (random network) [24] or a scale-free [25] topology (Figure 3). In the former, each vertex of a graph is equally likely to be connected to any other vertex; in the latter, the distribution of the number of connections, k, associated with each vertex follows a power law, $p(k) \sim k^{-\gamma}$ with $\gamma > 0$, and large interactions hubs are present. Many real biological networks appear to exhibit such structure [26].



Precision Recall Curves



rics, precision and recall. Recall, $N_{TP}/(N_{TP} + N_{FN})$, indicates the fraction of true interactions correctly inferred by the algorithm, while precision, $N_{TP}/(N_{TP} + N_{FP})$, measures the fraction of true interactions among all inferred ones.

Performance in reconstruction of artificial networks

Erdös-Rényi Topology	ARACNE		Relevance Networks		*DPI Sensitivity*	*DPI Precision*	Bayesian Networks	
Num samples	N_{TP}	N_{FP}	N_{TP}	N_{FP}			N_{TP}	N_{FP}
1000	128.00	1.33	143.33	462.67	99.71%	96.78%	50.00	32.33
750	124.33	2.67	139.33	411.00	99.35%	96.46%	45.33	31.00
500	119.00	1.67	130.67	311.33	99.46%	96.37%	41.00	29.00
250	101.00	4.67	110.00	182.33	97.44%	95.18%	24.67	25.33
125	81.00	4.67	84.67	95.00	95.09%	96.10%	5.33	19.00

Scale-Free Topology	ARACNE		Relevance Networks		*DPI Sensitivity*	*DPI Precision*	Bayesian Networks	
Num samples	N_{TP}	N_{FP}	N_{TP}	N_{FP}			N_{TP}	N_{FP}
1000	97.67	2.33	113.33	234.00	99.00%	93.67%	38.67	17.00
750	90.67	3.33	103.00	200.00	98.33%	94.10%	33.33	15.33
500	80.33	5.33	91.67	154.67	96.55%	92.95%	27.00	13.33
250	63.33	7.67	70.00	80.00	90.42%	91.56%	9.00	9.67
125	46.33	3.67	48.00	49.67	92.62%	96.50%	4.00	6.00

Applications to biological expression data

In the original paper, the algorithm was applied to human B cells expression profiles with very good results. Since then, it has been applied to all sorts of cell types and organisms (almost 600 citations) so it's not perfect but no algorithm can perfectly "solve the problem" and ARACNE is very popular, useful and elegant in its "simple" exploitation of DPI