Inferring causal relationships between dynamic genomic phase clusters

ID - 234

Dutta Abhishek[1,2], Blanzieri Enrico[1]

[1]Department of Information and Communication Technology, University of Trento, Italy

[2]School of Informatics, University of Edinburgh, UK

Motivation

Identification of cyclic genes by biologists and the mining of clusters with their dynamic relationships with computational techniques is well explored. However, these tasks have been performed separately and on small chunks of data. Here we propose a novel approach of filtering and clustering the entire set of yeast genes that are further correlated with the 5 cell-cycle phases and a high level belief network is built on top to capture the regulatory dynamics with the help of a score metric.

Methods

The data set is of yeast consisting of 7744 genes whose expressions are measured across 14 time points[2]. We identify 3 logical steps. (1)Filtering--The missing values were imputed by an Elman network. Then the genes with no valid names or null value expressions were deleted across the entire data set and log normalised. Since we are interested in the dynamics of the network, the low variance genes to 10th and low entropy genes to 15th percentile were filtered out subsequently leaving a set of 5594 more workable set of gene profiles. (2)Clustering--The above interested genes were then clustered using self organising feature maps, characterized by the Kohonen learning rule. The input space has 14 dimensions and no. of neurons in competitive learning layer is fixed to 5, the rationale being to establish later a correlation with the cell cycle phases. We further cluster them using fuzzy k-means and Pearson correlation as the distance measure. (3)Bayesian Belief nets--The centroids of the above 5 cluster profiles were discretized into up and down regulations w.r.t their corresponding means giving us 5 vectors featuring 14 binary data points in time. These form the 5 discrete nodes of the Bayesian network. Now any combinations of these 5 nodes could be the possible regulators of the others and vice versa i.e. they could be the targets as well; and through all possible time lags. Here the prediction of the network structure with causality inference w.r.t data becomes NP hard in the no. of nodes. Hence we make first assumption that the underlying process is markovian i.e. regulators could affect their target only 1 time step ahead. The second assumption is that a target could be regulated by at most two regulators. After all such alignments between the gene nodes 1 to 5, the prior conditional probability distribution tables are computed based on the data matrices. Then the maximum likelihood estimates are learned for all such possible pairs and also for the respective joints by marginalising. Further we compute the score metric by summing over the regulator's ML estimates given the up and down regulations of the target and also with the bayesian score metric[1] over both SOM and K-means matrices for single and couple of possible regulators.

Results

As can be seen in the plots, the mean cluster profiles using SOM and K-means synchronise well w.r.t phases, exactly where our interest lies.These are ordered according to the peaking of the profile, i.e. SG1 peaks before SG2 and so on. These also bear a strong correlation with the Stanford cell cycle classification performed over a subset of genes and hence here on we also label them with their respective phases M, G1, S, G2 etc. The first table shows the scores of the Bayesian network over SOM over each target, with their corresponding regulators in that row. The second table is for the Joint(2 possible regulators) over K-means again numbered row wise. These scores match perfectly with the bayesian scoring metric within a bounded fixed ratio. W.r.t them the final causal inferences are shown between the gene nodes or phases of the yeast cell cycle. SOM and K-means unanimously agree that S/G2 phase genes are regulated jointly by M/G1 and S phase genes as can be verified from the above tables. They also advocate G1 to be the regulator for G2/M. A new framework was developed for obtaining a high level causal relationship between various phases of the cell cycle regulated gene profile clusters build from scratch.

Availability: http://dit.unitn.it

Image: http://homepages.inf.ed.ac.uk/s0566455/bits.html

Email: a.dutta-1@sms.ed.ac.uk