Procedia Computer Science

Volume 29, 2014, Pages 1993–2002

ICCS 2014. 14th International Conference on Computational Science

A Clustering Based Method Accelerating Gene Regulatory Network Reconstruction

Georgios N. Dimitrakopoulos[1,2], Ioannis A. Maraziotis[1], Kyriakos Sgarbas[2] and Anastasios Bezerianos[1,3]

[1]Department of Medical Physics, Medical School, University of Patras, Patras, Greece.
[2]Electrical and Computer Engineering Department, University of Patras, Patras, Greece.
[3]SINAPSE, National University of Singapore, Singapore.
geodimitrak@upatras.gr, imaraziotis@gmail.com, sgarbas@upatras.gr, tassos.bezerianos@nus.edu.sg

Abstract

One important direction of Systems Biology is to infer Gene Regulatory Networks and many methods have been developed recently, but they cannot be applied effectively in full scale data. In this work we propose a framework based on clustering to handle the large dimensionality of the data, aiming to improve accuracy of inferred network while reducing time complexity. We explored the efficiency of this framework employing the newly proposed metric Maximal Information Coefficient (MIC), which showed superior performance in comparison to other well established methods. Utilizing both benchmark and real life datasets, we showed that our method is able to deliver accurate results in fractions of time required by other state of the art methods. Our method provides as output interactions among groups of highly correlated genes, which in an application on an aging experiment were able to reveal aging related pathways.

Keywords: Gene Regulatory Network, Clustering, Maximal Information Coefficient

1 Introduction

A basic goal of Systems Biology is to model the relations among genes and their products in order to increase the knowledge about the functional organization of cells. Specifically, these relations are modeled as networks of genes or proteins and diseases are studied through the observation of perturbations of those relations across different experimental conditions. The ultimate goal of this analysis is to determine genes with key role in the network that can be potential drug targets (Fernald, et al., 2011).

Recent molecular high-throughput techniques, like microarrays and sequencing, produce a huge amount of data, usually consisting of thousands of genes from few tens of samples. This poses a hard challenge to the computational methods, to cope with these large scale data and deliver accurate

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2014

doi:10.1016/j.procs.2014.05.183

prediction, taking into account the complexity of the relations among genes (a group of genes can co-regulate a third gene). Another obstacle to accurate network reconstruction is the noise embedded in microarray experimental data.

A plethora of Gene Regulatory Network (GRNs) reconstruction methods have been developed recently, based on different assumptions on modeling the underlying network, such as Regression methods, Information Theory, Boolean Networks, Bayesian Networks and Ordinary or Stochastic Differential Equation models (Hecker, et al., 2009; Emmert-Streib, et al., 2012). Methods from the first two categories have been applied in a broad set of problems showing promising results. However, a main problem with most of these approaches is that a high correlation value does not necessarily mean causal relation, therefore, considering the large number of variables, there is a high rate of false positive predictions.

In this work, the main goal is to provide a general methodology to reconstruct efficiently GRN. We present a method based on clustering to achieve two goals; increasing accuracy and reduction in time complexity. Clustering will guide an existing GRN inference method to avoid testing a certain set of genes that have expression profiles with high similarity degree as possible regulators of a certain gene. Although performance in terms of time is not critical in biological applications, it is desirable that a method is executed fast, to allow researchers experimenting with different parameters and multiple setups. Additionally, detection of small groups of interacting genes is of great biological interest, since it helps understanding the organization of the network and it can lead to detection of pathways related with a certain disease or condition (Langfelder & Horvath, 2008). Moreover, we introduce the usage of Maximal Information Coefficient (MIC) metric (Reshef, et al., 2011) for GRN reconstruction and show that it can accurately predict relationships among genes. First, the efficiency of our method was explored on benchmark datasets provided by Dialogue on Reverse Engineering Assessment and Methods (DREAM) project (Marbach, et al., 2012) and subsequently it was applied on two experimental datasets, studying ovarian cancer and aging, showing robust results.

2 Methods

We developed a general clustering framework that can be combined with any similarity metric or GRN method, with aim to reduce the search space and total execution time, while maintaining or improving performance in accuracy. The outcome of our method is a large number of groups containing relatively few genes with high similarity in their expression profiles and strong interactions among groups. Next, we present selective state of the art GRN regression methods and similarity metrics, including the recently proposed MIC. This metric resulted in high performance in benchmark sets and an experimental dataset studying ovarian cancer, so we applied it in combination with our clustering scheme to study aging.

2.1 GRN Methods

Initial approaches to create a network used a metric to assess the relation between every possible pair of genes and constructed a similarity matrix, based on the fact that similar gene expression shows similar functionality and interaction in molecular level. Furthermore GRNs are sparse, hence most methods, after computing a similarity matrix for each gene pair, use a threshold to eliminate weak interactions.

A simple and fast metric to assess similarity is Pearson Correlation Coefficient (PCC) or absolute value of PCC. This metric is able to estimate a regulatory network effectively, with the advantage that it can characterize interactions as activation or inhibition (Song, et al., 2012), however, is able to capture only linear relations. Weighted correlation network analysis (WGCNA) is a framework relying on PCC to infer and analyze gene networks in order to provide modules of highly correlated genes (Langfelder & Horvath, 2008). Many well established methods rely on Mutual Information (MI), a metric that can detect linear and non linear relations, such as Relevance Networks (Butte & Kohane, 2000), ARACNE (Margolin, et al., 2006), CLR (Faith, et al., 2007) and MRNET (Meyer, et al., 2007). Relevance Networks (RN) computes the MI and then uses a threshold to discard low MI values. Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE), uses the same process and in addition uses the Data Processing Inequality criterion, which considers every triplet of genes and removes the weakest interaction as immediate. Context Likelihood Relatedness (CLR) applies an adaptive background correction step, taking into account the network context, by calculating the distribution of all MI values involving either the regulator or the target. Minimum Redundancy/Maximum Relevance Networks (MRNET) uses maximum relevance/minimum redundancy (MRMR) strategy, under which the interactions are ranked and then each interaction is added to the network if shows maximum relevance (direct interactions with high MI value) and at the same time minimum redundancy (has small MI values with already selected interactions, which eliminates indirect interactions).

Another important category of GRN inference algorithms is based on Regression methods, which are used to predict one output variable based on one or more input variables. Various classes of regression methods have been used in the past to predict gene networks such as Artificial Neural Networks (ENFRN) (Maraziotis, et al., 2010), Support Vector Machines (SIRENE) (Mordelet & Vert, 2008) and Random Forests (GENIE3) (Huynh-Thu, et al., 2010). Evolutionary Neuro-Fuzzy Recurrent Network (ENFRN) algorithm accepts as input gene expression as a training and a testing dataset and determines the best potential regulators and target genes. Interesting properties are that it explores the combined effect of two or more regulators to a target and that it can predict the type of the relation as activation or inhibition. Supervised Inference of Regulatory Networks (SIRENE) uses SVM to solve a classification problem for each Transcription Factor (TF), if a gene is target of it a or not. The operation of SIRENE is in a little different direction, since it requires as input examples of known TFs and their targets and can predict new targets for these TFs, but does not reconstruct the whole network, neither can find new regulators. Random Forest is an ensemble method, which uses a large number of classification or regression trees to infer the final result. Gene Network Inference with Ensemble of Trees (GENIE3) trains a Random Forest for each gene as target, considering all other genes as regulators. Next, it uses the Variable Importance metric of the Random Forest to evaluate the rank of the potential regulators for each gene. We will use GENIE3 in comparisons to our method, because it displayed the best performance according to DREAM 5 contest, setting the number of trees equal to 1000.

2.2 Maximal Information Coefficient

We use Maximal Information Coefficient (MIC) in our framework to estimate the pairwise similarity of gene expression profiles. MIC is a recently proposed metric relying on mutual information, which can detect a wide range of associations, showing solid results. The idea behind MIC is that if a relationship exists between two variables, then a grid can be drawn on the scatterplot of the two variables for partitioning the data points and encapsulating this relationship. Thus, it partitions the space into different grids up to a maximal grid resolution *B*, Mutual Information (MI) is computed for each grid and MIC is set equal to the maximum normalized value of MI. Formally, for two variables *X* and *Y*, MIC is defined by the equation:

$$MIC(X,Y) = \max_{|X||Y|<B} \frac{MI(X,Y)}{\log(\min(|X||Y|))}$$

where |X| and |Y| are the number of bins for each variable and *B* the maximal resolution. Number of grids to be searched is controlled by two parameters, *alpha* and *c*. *B* is function of the number of samples *s*, $B = s^{\text{alpha}}$ and *c* is the maximum allowed difference between |X| and |Y|. Interesting properties are symmetry, i.e. *MIC*(*a*,*b*) = *MIC*(*b*,*a*), so we can skip half of the computations and that it ranges in [0,1], making easier the interpretation of the results, in contrast to mutual information which is positive but without an upper limit.

There are only few works discussing ability of MIC to capture relationships among genes. In (Song, et al., 2012), various metrics, including MIC, are examined resulting in the conclusion that in most cases metrics agree, with the authors proposing a variation of correlation for GRN problems. In a similar study, it is shown that each metric can perform well depending on the data and type of relationship to be identified, with MIC being an appropriate selection for large data and a broad type of interactions, such as non linear and non monotonic (de Siqueira Santos, et al., 2013).

2.3 Proposed Framework

The proposed method is based on the key idea that a set of genes, which have a high degree of similarity in terms of their expression profiles, will have similar relation (i.e. target - regulator or vice versa) with a certain gene. Therefore, we employ a clustering scheme to avoid computing the weight of interactions between all possible pairs. The increased time complexity of most GRN methods (i.e. GENIE3, SIRENE) combined with the large dimensionality of gene expression datasets makes most of these methods practically inapplicable in full scale gene expression experiments. Hence, in the majority of the cases, genes having small variation across the experimental conditions or small absolute value (considered as not expressed) are discarded. In our framework, instead of applying a GRN reconstruction method to the whole dataset, which is very consuming in time and memory, we will apply it only on a subset which will be provided by clustering.

Initially, we cluster the data in a very large number of groups, so each group contains few genes, but with very similar profiles. An important property of the clustering algorithm is to provide a medoid for each cluster, which is an actual point of the dataset, in contrast to centroid, which is a mean value of points belonging to the cluster. Thus, we are going to use only the medoid of each cluster as representative of all genes in the group. We build a "medoid network" of reduced size, by evaluating the weight of interactions among all medoids. At this point, we have a hierarchical network, where each node is a group of genes, however, for evaluation and comparison with other methods, we need to provide a network with genes as nodes. Therefore, for each gene pair, we allocate as weight of their interaction the weight of interaction of their medoids. The rationale behind this choice is that based on the compactness of the clustering, we need to examine one profile per cluster. Another advantage of considering the medoid instead of each gene in a cluster is that the effect of the noise is reduced, as one can regard each gene as a noisy measure of the expression profile of the medoid. Figure 1 summarizes the workflow of our proposed method.

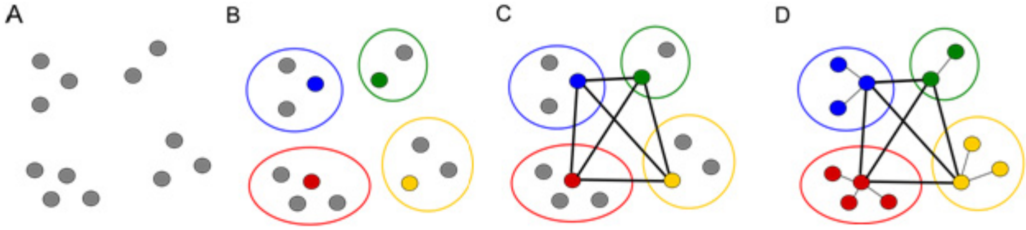

Figure 1. The workflow of our method. (A) In the schematic representation, every node indicates a gene. In (B) after clustering, the original input dataset is replaced by a compressed dataset represented by the medoids of the corresponding clusters and in (C) the weight of interactions among medoids is calculated. In (D) for interactions between two genes in different clusters, the weight of interaction between their corresponding medoids is assigned.

Under this framework, genes belonging in the same cluster appear to have maximum similarity (MIC value equal to 1). If a more detailed "picture" of the network is needed, the GRN method can be applied locally for each cluster. An extension of this framework is for each gene inside a cluster, to evaluate the interactions among the genes that belong only to a number of nearest clusters. To determine the distance between two clusters, similarity between their medoids can be used, which is already computed at the first step. However, initial experiments showed that this setup for large clustering resolution increased search space without changing substantially accuracy of the inferred network, so we have not performed these steps in this work. This is explained by the fact that the large number of clusters increased their compactness, so genes belonging in same cluster resulted in high similarity values and setting similarity to maximum was a satisfactory approximation.

The time complexity of inferring a network of N genes is at least $O(N^2)$ for any method as the ones described above. To build the medoid network of C clusters, will demand $O(C^2)$ time, with $C < N$. Precisely, the total process will require additional time for clustering, but we can neglect it, since in practice clustering is usually a lot faster process than a applying a GRN method, regardless the clustering algorithm complexity, which is usually at least $O(N^2)$. So, a significant improvement in execution time can be achieved due to quadratic complexity.

Additionally, to further reduce the size of the dataset, we can perform clustering across the samples. Most datasets contain few samples, but in cases such as the DREAM 5 datasets, where many microarrays have been aggregated, this could be very helpful. Computationally, it is known that as dimensionality increases, the points in space become very sparse and metrics underperform. From biological aspect, replicates or samples under the same condition hold similar information, so clustering can reduce this redundancy.

The proposed method is general and can utilize any clustering algorithm in combination with any metric or GRN algorithm. However, in this paper for clarity, we use Affinity Propagation Clustering (Frey & Dueck, 2007) and MIC to estimate similarity between genes. A key property for selecting this algorithm is that returns medoids instead of centroids. Affinity Propagation Clustering algorithm iteratively transmits real valued messages among data points, updated in each round with purpose to minimize squared error, until a good set of medoids (exemplars) and corresponding clusters emerges. The method is called "affinity propagation", because at any point in time, the magnitude of each message reflects the current affinity that one data point has for choosing another data point as its medoid. This algorithm accepts as input a similarity matrix (negative of Euclidean distance was used) and a parameter called preference, which indirectly affects the number of clusters. The value of preference can be adjusted so to get a specific number of clusters, if it is desired. Other algorithms can equally well be used such as k-centers (MacQueen, 1967), a medoid-based variation of k-means.

2.4 Data

To test the efficiency of our method, we used the datasets and the gold standards provided by DREAM 5 network inference challenge (Marbach, et al., 2012). The DREAM project provides benchmark datasets along with the real network topology derived from biological validated data and organized annual challenges for Systems Biology problems like network inference. We used the 3 datasets with averaged experimental conditions, containing 1643 genes - 487 samples (in silico), 4511 genes - 487 samples (E. Coli) and 5950 genes - 321 samples (S. cerevisiae), which from now on will be referred as D51, D53 and D54 respectively.

Additionally, we used a human ovarian cancer microarray dataset accessible through NCBI's Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) with series accession number GSE14407 (Bowen, et al., 2009), which is based on 12 normal surface epithelial cell samples and 12 unmatched cancerous epithelial cell samples from serous papillary ovarian adenocarcinoma (for the network construction, we used only the normal samples). This dataset is closer to most applications, as it has few samples and it is based on human data. Since a reference GRN does not exist, the evaluation was performed with the network of experimentally validated regulatory interactions from (Madhamshettiwar, et al., 2012), containing 6,330 interactions among 280 TFs and 2,170 targets, originally derived from TRANSFAC database.

Finally, we applied our method in a human aging dataset (NCBI GEO series accession number GSE11882), which profiles gene expression in 4 brain regions of cognitively intact humans, across the adult lifespan (173 samples of ages 20-99) (Berchtold, et al., 2008).We kept only the same 2450 genes with ovarian cancer dataset, in order to be able to use the known TF's-Targets network for evaluation.

3 Results

To assess the accuracy of the inferred networks, Area Under Curve (AUC) was used, which is computed as the area under Receiver Operator Curve (ROC) which in turn is the plot of the true positive rate versus the false positive rate at various values of threshold. This way, we avoided selecting a specific threshold, which varies for each method and dataset, since AUC essentially takes into account all possible thresholds. All algorithms were implemented in Matlab and executed in a PC with i7 3.1 GHz CPU and 8GB RAM. MIC implementation by (Albanese, et al., 2013) was used. All data were normalized per gene profile to zero mean and unit variance.

3.1 Benchmark datasets

We used the DREAM 5 benchmark datasets, to explore the efficacy of our framework and fine tune the parameters. In all cases MIC parameter c was set to 15 (default) and alpha to 0.25, which yielded optimal performance in DREAM 5 datasets in terms of AUC (AUC was increased about 4% in D51and 2% in rest dataset), despite that this value led to examination of less and larger grids. In Figure 2 we show the performance of MIC in comparison with PCC, ARACNE and CLR. It is clear that MIC is competitive with these established methods and specifically better than MI based methods in D53 and D54 datasets. Remarkably, it achieves the best performance in Dataset 4, which is the largest, not only among the methods shown in Figure 2, but also among the 35 methods presented in (Marbach, et al., 2012).

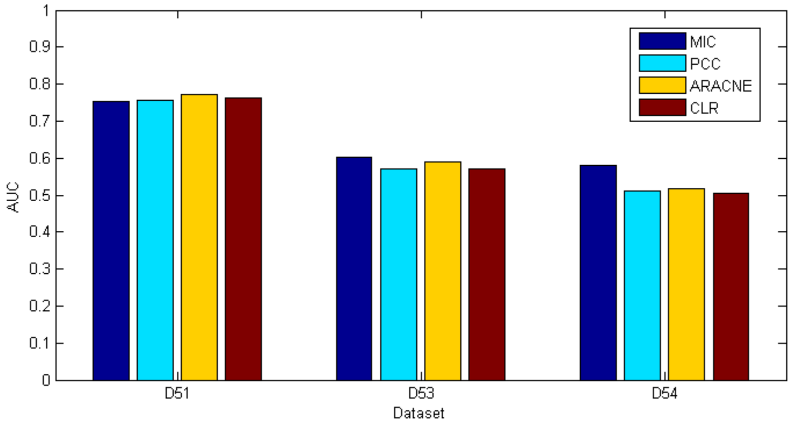

Figure 2. Performance of various metrics and methods on the 3 datasets of DREAM 5. MIC shows similar or better performance with other state of the art methods.

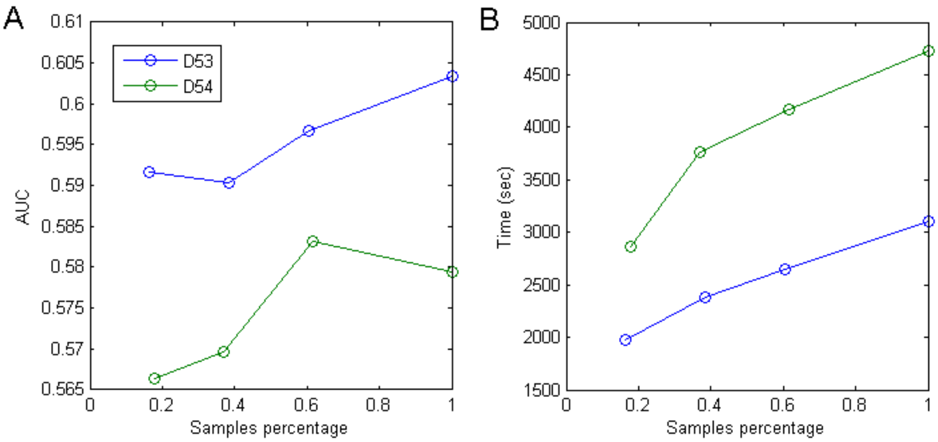

Figure 3. Plots of (A) AUC and (B) time when clustering the 2 large DREAM 5 across samples varying the number of medoids (x axis represents the percentage of samples kept) and inferring the network with MIC. Keeping only about 60% of medoids results in small loss in performance and in one case in increment, while time is reduced significantly.

Next, we performed clustering across the samples for the two largest DREAM 5 datasets D53 and D54, keeping only the medoids, as shown in Figure 3, and then inferred the network using MIC. We observed that despite the large dimensionality reduction, performance has decreased in a relatively very small percentage and in one case there is improvement. The last observation indicates that a metric might perform better with less variables and that clustering managed to successfully compress the data. Therefore, it is a worthy process to get a quick estimation of the results, since the time needed for any method is directly proportional to number of samples.

Moving forward, we applied our clustering framework in combination with MIC to DREAM 5 datasets and the ovarian cancer dataset. In DREAM 5 datasets, GENIE3 was the best or among the top algorithms in performance, while for the ovarian cancer dataset 8 unsupervised GRN methods provided predictions slightly better than random guess (0.50), with Relevance Networks having maximum performance of 0.55 (Madhamshettiwar, et al., 2012). In Table 1, some indicative results of our method are shown, obtained by setting the number of clusters to 30% of the original genes in DREAM datasets and 10% in the case of ovarian cancer. As we observe, on one hand performance better than or close to maximum can be achieved with our method and on the other, in comparison to MIC, results remained unchanged or improved, while lowering the execution time 10 to 100 times. Similar results were observed for a wide range of cluster numbers (10%-30%).

One important property of our method is that it can work under different metrics. Thus, replacing MIC with PCC, despite that it is an extremely fast method and time performance is not an issue, we observed similar behavior and improved accuracy. Specifically, in D54, PCC over the whole dataset

Method	**Dataset**							
	D51		D53		D54		Ovarian Cancer	
	AUC	Time (sec)	AUC	Time (sec)	AUC	Time(sec)	AUC	Time (sec)
GENIE3	0.81	1612	0.61	5832	0.52	4788	0.50	150
MIC	0.76	411	0.60	3128	0.58	4727	0.51	204
Our method	0.72	43	0.60	284	0.58	433	0.54	1.8

Table 1. Indicative examples of performance and execution time of various methods in comparison to our method using MIC.

yielded an AUC of 0.52, while with our method and 2000 medoids (i.e. 1/3 of the genes), 0.58 was achieved. Similarly, in ovarian cancer dataset, PCC resulted in AUC=0.51 and with our method and 1000 medoids (40% of genes), it was increased to 0.52. The conclusion from all experiments is that in order to retain a high performance, it is critical to have a very large number of clusters, with 30% representing a good balance between speed and accuracy.

3.2 Aging

Finally, we applied our method in a dataset studying human aging. Initially, we split the data into 3 subsets, containing each young (20-30 years, 30 samples), middle (31-65, 54 samples) and old (66-99, 89 samples) ages. Next, we built a network for each age group, by applying our method with MIC, clustering data into 500 groups. We observed a major difference in time and an improvement in performance. Specifically, for young ages we obtained by MIC AUC of 0.48 in 240 seconds versus 0.54 in 10 seconds for our method. In both middle and old ages, the values for MIC were 0.49 in 580 seconds and 690 seconds respectively, while our method achieved 0.51 in 21 and 22 seconds respectively. Consequently, we examined the genes which their interactions changed significantly during aging. To determine the differentiated regulatory links, we isolated the interactions that had more than 0.8 difference in MIC value between two networks. We detected 135 differentiated genes between young and middle networks, 34 between middle and old and 178 between young and old. For these genes we performed Gene Ontology (GO) and Pathway enrichment analysis using DAVID Functional Annotation tool (http://david.abcc.ncifcrf.gov) and some indicative results are shown in Table 2. These pathways and GO categories have been known to be associated with aging, for example Cell cycle and p53 signaling pathway (Rufini, et al., 2013) and MAPK signaling pathway specifically in brain tissue (Zhen, et al., 1999). Also, some cancer related pathways appeared, which is no surprise, since literature supports that aging and cancer involve some common mechanisms (Finkel, et al., 2007).

An important remark is that low AUC was achieved here and in previous datasets, but since gold standard is derived from experimentally validated interactions between TFs and targets, on one hand many true interactions are not known yet and on the other some interactions might not occur in the specific conditions of the experiment. Finally, the usefulness of our method lies in the fact that results can reveal valuable information when examined under different perspectives, such as genes cooperating in pathways.

Condition	**Pathway / GO Term**	**P-Value**
Young-Middle	p53 signaling pathway	4,9E-5
	Insulin signaling pathway	3,3E-3
	TGF-beta signaling pathway	8,3E-3
	regulation of cell death (GO)	1,7E-9
	organ development (GO)	9,8E-13
	system development (GO)	5,7E-12
Young-Old	Insulin signaling pathway	5.7E-3
	MAPK signaling pathway	1,8E-2
	Cell cycle	1,3E-2
	Pathways in cancer	8,3E-5
	Thyroid cancer	2,5E-4
	Endometrial cancer	3,8E-3

Table 2. Gene Ontology categories and KEGG Pathways detected to differentiate significantly among different age groups.

4 Conclusions and Future Work

In this work we presented a general clustering strategy, which can be combined with any GRN method and achieve similar level of accuracy, by applying the GRN method in a fraction of the dataset, accelerating so the network inference. We introduced usage of MIC metric for GRN inference and achieved high performance and in some occasions better than any method reported in recent literature. Finally, in the application on aging dataset, groups of genes showing differentiation among different ages, successfully identified pathways relevant to the problem in study.

In future work we plan to investigate the inclusion of different kind of a priori information, such as Gene Ontology (GO), Protein - Protein Interactions, Pathway Maps and TF's - targets (the latter is currently used for evaluation, so it is not possible to utilize it in network reconstruction). Embedding experimentally validated information into the network has been proved to increase the biological consistency of results (Maraziotis, et al., 2012). This is the only way to increase accuracy in datasets, in which all unsupervised methods provided results marginally better than random guess. A natural adaptation of our method is, complementary to clustering, to use GO terms or Pathway Maps, which provide a grouping of genes known to cooperate to perform a function. This way, researchers who are studying a specific biological problem can easily focus only on GO categories or pathways related to problem in study, rather than reconstructing the whole network. Moreover, Protein - Protein Interaction is a different kind of a priori knowledge that is available, but it does not provide a grouping of genes unless more preprocessing steps are followed (for example creation of a PPI graph and extraction of functional modules).

Acknowledgement

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thalis. Investing in knowledge society through the European Social Fund.

References

Albanese, D., Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G. and Furlanello, C, (2013). Minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. *Bioinformatics,* 29(3): 407-408.

Berchtold, N. C., Cribbs, D. H., Coleman, P. D., Rogers, J., Head, E., Kim, R., Beach, T., Miller, C., Troncoso‖, J, Trojanowski, J. Q., Zielke, H. R. and Carl W. Cotman, (2008). Gene expression changes in the course of normal brain aging are sexually dimorphic. *PNAS*, 105 (40): 15605-15610.

Bowen, N.J., Walker, L.D., Matyunina, L.V., Logani, S., Totten, K.A., Benigno, B.B. and McDonald, J.F., (2009). Gene expression profiling supports the hypothesis that human ovarian surface epithelia are multipotent and capable of serving as ovarian cancer initiating cells. *BMC Med Genomics*, 2:71.

Butte, A. J. and Kohane, I. S., (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, pp. 418-429.

de Siqueira Santos, S., Takahashi, D. Y., Nakata, A., and Fujita, A., (2013). A comparative study of statistical methods used to identify dependencies between gene expression signals, *Brief Bioinform.* bbt051.

Emmert-Streib, F., Glazko, G. V., Altay, G. and de Matos Simoes, R., (2012). Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Front. Gene.*, 3:8.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J. and Gardner, T. S., (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5:e8.

Fernald, G. H., Capriotti, E., Daneshjou, R. and Karczewski, K. J., (2011). Bioinformatics challenges for personalized medicine. *Bioinformatics*, 27 (13): 1741–1748.

Finkel, T., Serrano, M. and Blasco, A. M., (2007). The common biology of cancer and ageing. *Nature*, 448, 767-774.

Frey, B. J. and Dueck, D., (2007). Clustering by Passing Messages Between Data Points. *Science*, 315 (5814), 972-976.

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E. and Guthke, R., (2009). Gene regulatory network inference: data integration in dynamic models-a review. *Biosystems*, 96(1):86-103..

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. and Geurts, P., (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, 5:e12776.

Langfelder, P. and Horvath, S., (2008). WGCNA: an R package for weighted correlation network analysis, *BMC Bioinformatics*, 9:559

MacQueen, J. B., (1967). Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press,* 1:281-297

Madhamshettiwar, P., Maetschke, S., Davis, M., Reverter, A. and Ragan, M., (2012). Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Med*, 4:41.

Maraziotis, I. A., Dragomir, A. and Thanos, D., (2010). Gene Regulatory networks modeling using a dynamic evolutionary hybrid. *BMC Bioinformatics*, 11:40.

Maraziotis, I. A., Dimitrakopoulos. G. N. and Bezerianos, A., (2012). Gene Ontology Semi-supervised Possibilistic Clustering of Gene Expression Data. *SETN 2012: 7th Hellenic Conference on AI, Lamia, Greece, Proceedings, LNCS (LNAI), Springer-Verlag Berlin Heidelberg*, 7297: 262–269.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M. and Allison, K. R., (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796-804.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R. and Califano, A., (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1):S7.

Meyer, P. E., Kontos, K., Lafitte, F. and Bontempi, G., (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J Bioinform Syst Biol* , 79879.

Mordelet, F. and Vert, J. P., (2008). SIRENE: supervised inference of regulatory networks. *Bioinformatics*, 24:i76-82.

Reshef, N. D., Reshef, A. Y., Finucane, K. H., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, S. E., Mitzenmacher, M. C. and Sabeti, P. (2011). Detecting Novel Associations in Large Data Sets. *Science* , 16 12, 334 (6062).

Rufini, A., Tucci, P., Celardo, I. and Melino, G., (2013). Senescence and aging: the critical roles of p53. *Oncogene*, 32, 5129–5143.

Song, L., Langfelder, P. and Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices, *BMC Bioinformatics*, 13:328

Zhen, X., Uryu, K., Cai, G., Johnson, G. P.and Friedman, E. (1999). Age-Associated Impairment in Brain MAPK Signal Pathways and the Effect of Caloric Restriction in Fischer 344 Rats. *J Gerontol A Biol Sci Med Sci*, 54 (12): B539-B548.