Paired Transcriptional Regulatory System for Differentially Expressed Genes

Aurpan Majumder
Dept. of E.C.E, National Institute of Technology, Durgapur, India
Email: aurpan.nitd@gmail.com

Mrityunjay Sarkar
Dept. of E.C.E, Durgapur Institute of Advanced Technology and Management, Durgapur, India
Email: mrityu1488@gmail.com

***Abstract*—A fully functional gene regulatory network can be formed using gene-gene and/or gene-protein interactive patterns. To maintain a healthy cell cycle, it is necessary to have a proper control of the regulatory proteins in the network. Excess protein concentration may lead to beyond control division of the healthy cells causing cancer. In this context, transcriptional regulators (proteins) are responsible for changes in gene expression levels across different developmental stages. In our work we have extended a recently developed procedure to find out the pair(s) of TFs, which can control a target gene from a linear prospective. Here, we have explored the pairwise regulatory action through mutual information and spline regression. In the result segment we have shown that the controlling action between these two methods is dependent upon the dimension (number of samples) of the data. For large dimension spline regression based controlling shows better result that MI, and vice versa for smaller dimensions.**

***Index Terms*—transcription factor, mutual information, spline regression, differentially expressed (DE) genes, colorectal (colon) cancer**

I. Introduction and Theory

A gene regulatory network is a system where genes and proteins bind to each other and act together controlling various cellular functions [1].An eukaryotic organism can exist only when all its cells function according to the rules governing cell growth and reproduction; indicating the existence of a regulatory network under control. Though there are some external factors responsible for cell division (UV light, X-rays, chemicals, tobacco products, viruses are responsible for cancerous growth) it is ultimately the signalling proteins which cause the nucleus to stimulate the cell division. These proteins causes a signal transduction cascade which includes a membrane receptor for the signal molecule, intermediary proteins that carry the signal through the cytoplasm, and *transcription factors* (T.F) in the nucleus that activate the genes for cell division (cell cycle genes) [2].In each step of the pathway one T.F or protein activates the next; however, some TFs can activate more than one protein in the cell.

Microarray time series data gives us a possible way to identify transcriptional regulatory relationship among the genes [3]. In our previous work [4] we have utilised the concept of Pearson's correlation coefficient to identify potentiality of regulatory action of two genes on a target gene. Although this traditional method has been successfully applied to find functionally correlated genes, it has a limitation of highlighting the linear regulatory relations. However, there still remains a fair chance to explore the nonlinear association of TFs and target (regulated) genes (especially with large time series microarray data). Thus to discover the total potential of nonlinear regulatory operation we have indulged ourselves towards mutual information [5], [6] and spline regression [7] based approaches to construct the gene regulatory networks.

In this work to check the nonlinear regulation of the T.Fs over the cell cycle process we have used varied data sets. Firstly, we used the budding yeast, *Saccharomyces cerevisiae* cell cycle data [8]. In the second case, colorectal *(colon) cancer* data which accounts 10% to 15% of all cancers and is the second leading cause of cancer-related death in industrialized countries [9].

The target genes mentioned above are those having changed expression levels across different conditions, in other words they are the differentially expressed (DE) genes. Corresponding to each DE gene we predict the possible TF pairs which in combination conduct non linear regulation of the target gene towards altered expression levels across different stages of cell cycle.

The rest of the paper is as follows. In next section we have discussed about the *Methodology*. A detailed view of the problem and its implementation on the datasets has been given in *Results and Discussion* section. The paper concludes with *Conclusion and Future work.*

II. Methodology

Step 1 of the algorithm is used to find out the differentially expressed (D.E) genes between the two conditions. Here we have implemented the same using DEGseq [10] (It is an R package to find the differentially

Manuscript received July 24, 2014; revised September 24, 2014.

doi: 10.12720/lnit.2.3.266-272

expressed genes from the gene expression value itself. Here depending upon expression values of genes at different time instants and by a particular threshold ‘p-value’/’z-score’/’q-value’ D.E genes are computed).

ALGORITHM. BEST POSSIBLE COMBINATION OF TF PAIRS

```
Input:  gEmtx1, gEmtx2, TF1, TF2, TFg1, TFg2, $\beta$ ←1
Output: NLM1, NLM2, Mv1, Mv2, Mv, bTFp1, bTFp2, TFp1,
        TFp2

Step1.  DE←DEGexp(gEmtx1,set expression columns for
        conditition1,gEmtx2,set expression columns for
        condition2)
Step2.  c ← choose mode of operation
        for i in 1 : X do
          for j in 1:Y do
            r1← rbind(DE1[i,],TF1[j,])
            nr1← transpose(r1)
            r2← rbind(DE2[i,],TF2[j,])
            nr2← transpose(r2)
           if c ==1 then
            NLM1 [i,j]←NonLinMI(nr1, $\beta$ )
            NLM2 [i,j]←NonLinMI(nr2, $\beta$ )
          else
            NLM1 [i,j]← NonLinSP (DE1[i,],TF1[j,])
            NLM2 [i,j]← NonLinSP (DE2[i,],TF2[j,])
          end if
         end for
        end for

Step3.  for i in 1:X do
          m←0
          for j in 1:Y-1 do
            for k in j+1:Y do
              m←m+1
              Mv1[i,m]←NLM1[i,k]*NLM1[i,j]
              Mv2[i,m]←NLM2[i,k]*NLM2[i,j]
            end for
          end for
        end for

Step4.  Mv← Mv1-Mv2
        for i in 1:X do
          M[i]← min(Mv[i,])
          n← 0
          for j in 1:Y-1 do
            for k in j+1:Y do
              n←n+1
              if (M[i,1]==Mv[i,n])
               Indx1← j
               Indx2←k
               bTFp1[i]← TFg1[Indx1]
               bTFp2[i]← TFg2[Indx2]
              end if
            end for
          end for
        end for

Step5.  M ← $Y_{c_2}$
        for i in 1 : X do
          for l in 1: M do
             TFp1[i,l]← 0
             TFp2[i,l]← 0
          end for
            V[i]=M[i]+ $\Delta v$
            n←0, s←0
          for j in 1:Y-1 do
            for k in j+1:Y do
              n←n+1
              if MV[i,n] < V[i,1] then
                s←s+1
                TFp1[i,s]←TFg1[k,1]
                TFp2[i,s]←TFg2[j,1]
              end if
            end for
          end for
        end for

%%%% End of main routine %%%%%%

NonLinMI← function (nr, $\beta$ ) {
V← mutualInfoAdjacency (nr,discretize columns, set entropy
estimation method, set the number of discretization beans) ^ $\beta$
}

NonLinSP ← function (DE,TF) {
sm.spline (DE,TF,set the order of spline function)
}
```

Step 2 of the algorithm is dedicated to compute the nonlinear association between D.E genes and Transcription Factors (T.F) between the two conditions through mutual information and spline regression. Let there be X number of D.E genes (from Step 1) and Y number of TFs. DE1 stands for gene expression matrix of DE genes under condition1 and DE2 correspondingly for condition 2. Similarly, TF1 happens to be the gene expression matrix of TFs under condition1 and TF2 for condition 2. Here X number of D.E genes will act as the target genes. We have used a function *NonLinMI* to find out the symmetric uncertainty based mutual information (MI) adjacency measure between each TF and the target gene in each condition; in a similar fashion we have used another function *NonLinSP* to find the nonlinear association between D.E and T.F by spline regression based measure. These two user defined functions compute the corresponding measures by invoking two R package functions *mutualInfoAdjacency* and *sm.spline* (in this work we have taken the order =3, cubic spline). Each of these approaches will create two matrices NLM1 and NLM2.

In Step 3 we multiply all possible pairs of row elements of each matrix separately indicating multiplication between every pair of nonlinear association values obtained between TFs and the target gene. As shown in our ALGORITHM the multiplication results are stored in the set of matrices Mv1 and Mv2 respectively. As we are multiplying each possible pair so for Y number of columns (T.Fs) there will be Y_{c_2} number of combinations and accordingly we can say that Mv1 and Mv2 will have X number of rows and Y_{c_2} number of columns.

Here high multiplication result suggests that the dependency of the target gene with respect to two TFs is high, and a low value suggests that the dependency is low. In other words a high value means strong regulatory action whereas a low value point towards weak regulation.

In Step 4 we perform subtraction between the multiplied results between the two conditions (Mv).

Keeping in mind that small subtraction result suggests that the regulation of the target by both TFs in each condition is near about same and a relatively large subtraction value highlights the effective regulation of the target by both TFs between conditions are not equal. We introduce a filtering through the selection of TF pairs having small subtracted values between the two conditions and the best TF pair will be the one having the minimum of these values. Best TF pairs for a particular target gene are stored in *bTFp1* and *bTFp2*.

Step 5 is an extension of Step 4. Here by taking a range of values within a limit and the subtracted result of the best TF pair as the reference we search for that TF pairs which have subtracted values within that limit.

Let for target gene 1, T.F pair *x* and *y* have given the minimum subtraction value and let the value be *M [1]*. Now setting the limit mentioned above as Δv we select those pairs of TF genes who have subtracted outcomes within the range *M [1]* + Δv

We repeat the above mentioned procedure for remaining X-1 number of target genes.

III. Results and Discussion

As discussed previously we have tested our ALGORITHM on two publicly available data sets. First one is the budding yeast, *Saccharomyces cerevisiae* cell cycle data. Details about the data set, how the DE genes and TFs are found are discussed in [4]. As mentioned in [4] there are in total 285 DE genes and 17 TFs. At first we do proceed via *NonLinMI* function to search for significant TF pairs against target genes by mutual information based adjacency measure. In table I we have given the significant TF pairs and their corresponding target genes.

Next, we proceed via *NonLinSP* function to find the same but this time through spline regression approach. In this context Table II enlists the significant TF pairs and their corresponding target genes.

Another data set that we have used for testing is the Affymetrix expression data of colon cancer tissues, having 22,278 genes in total, in two conditions. There are in total of 111 colon tissues from tumours and adjacent noncancerous tissues out of which 49 tissues are noncancerous and 48 are cancerous tissues. Details of the data set can be found in [11].

Again on the other side, we have searched for TFs from [12]. By matching the IPI id provided in [12] with the IPI ids in our data set we found a total of 1065 TFs.

After finding the TFs we extracted the TFs in both the cases from the entire gene expression data. Then from the remaining data set we have searched for DE genes. Total number of DE genes found across these two conditions is 56. DE genes are found using DEGseq [10]. These 56 DE genes will act as targets. Now, as mentioned in Step 4 and Step 5 of ALGORITHM we have searched for best TF pairs as well as other TF pairs showing good prediction power corresponding to target DE genes through MI and spline regression measure. In Table III and Table IV corresponding to colon cancer we have enlisted some of the DE genes and with respect to each of them the possible TF pair(s) obtained using MI and spline regression methods respectively.

TABLE I. Best Combination OF TF Pairs Corresponding to Target Genes FOR Yeast Cell Cycle Data Through Mutual Information Adjacency Measure

Target (DE)	**TF pair(s)**
PRM5, BUL2	ACE2, FKH2
YDR124W, YER010C	ACE2, MCM1
PGM2	ACE2, CST6
STE2, TIF1, RSA1, YMR111C	ACE2, ASH1
YPC1, BSC1, MCT1, YML119W	MCM1, ASH1
PRB1	STE12, SWI6
DSE12, RIM21, GPD1, BAP2	RLM1, STP1
YJL160C, MOG1, LTV1, YKL044W, MFG1, GYP7, SRB7, PCM1, IME4, LSB1, DIA4, YSC84, HXT4, AYR1, YJR026W, PSO2, MRPS18, IMA2, YOR029W, YOR053W, MKK1, YPL039W, YPL062W	TEC1, STB1
GLK1, TDP1, ERG24, ESC8	SWI4, TEC1
ATO2, DIP2, PIR3	CST6, SWI4
YBR225W, KCC4, DOT5, YJL068C, CYC2, YPK2, YIL108W	TEC1, RLM1
OM14, DIA3, YET3, PAM1, ATP17, GSY1, YGL052W, YHR097C, PRM10, SMD2, VPS38, YML131W, SIP5, CIK1, BOR1, BSC6, GDH1, NTO1	ASH1, TEC1
FMP30	ACE2, STE12
FUS1, ECM4, YBR138C	CIN5, CST6
MFA1, SSU1, FDH2	TEC1, CST6

TABLE II. Best Combination of TF Pairs Corresponding to Target Genes for Yeast Cell Cycle Data Through Spline Regression Measure

Target (DE)	**TF pair(s)**
GRX7, PRB1, CNB1	ACE2, FKH1
FUS3, SPI1, STF2, MOD5	ACE2, SWI5
UGA2, RTC3, SRP40, KTI12, YKL044W, AFR1, PRM10, COS9	MCM1, SWI4
LEU2	SWI5, ASH1
RCR1, YNL146W, YJR154W	CIN5, CST6
FIG2, NMA1, GSF2, HCH1, AFI1	TEC1, CST6
YPR142C, YBR144C, CCT4, GUD1, ECM18, GGA1, YGL117W, YGR149W, DSE2, ERG24, HXT10, PIR3, BUR2, PGM2, IMA2, FMP21, KCC4, YDR249C, YML131W, OM14, SDH4, YIL108W, YJL068C	STE12, ASH1
YBR144C, KCC4, FMP21, HXT10, GSY1, MST27, HXT4, AYR1, GUD1, YJR026W, YKL151C, CIK1, ERG24, ATO2, RPL25, YOR053W	STE12, STB1
PAM1, DIA4, SMD2, PAM16	STE12, RLM1
STP4, YET3, GYP7, YJL052W, SLT2, HXT6, HXT9, MYO3, YLR253W, PSO2, YNL043C, YMR317W, FDH2	TEC1, STB1
IME4, APE1, DIP2, YOR121C, MKK1, VPS38, FSH1, CAP2, GIP3, MF(alpha)1	ASH1, TEC1
TDP1, YBR225W, FUS1, YCR007C, ATP17, BCY1, BAP2, POR2AGA2, YSC84, TFA2, MCM5, NIT3, SIP5, VTI1, BOR1, ESC8, NTO1	ASH1, STP1

Validation of Yeast gene regulatory networks have been done by using a web based tool called YEASTRACT (www.yeastract.com) [13].

In Fig. 1 and Fig. 2 we have shown some TF pairs to target gene regulatory networks obtained using the ALGORITHM; here genes mentioned in text boxes having gray background represents TFs and genes mentioned in text boxes having white background represent target genes.

For human colon cancer data validation is performed by two web based tools namely Tfacts (www.tfacts.org) and PRISM (www.PRISM.stanford.edu) [14]. The corresponding steps to use these tools are discussed below.

TABLE III. BEST COMBINATION OF TF PAIRS CORRESPONDING TO TARGET GENES FOR HUMAN COLON CANCER DATA THROUGH MUTUAL INFORMATION ADJACENCY MEASURE

Target (DE)	TF pair(s)
CA1	CDX2, PAX2
GCG	CREB1,FOXA1,HOXC8, ST18 AHCTF1,STAT1
INHBA	ATF1,CREB1,NFYA,MEF2B,HIVEP3
CHGA	ATF1,CREB1,EGR1,ETS2,JUN,HOXC4, TFAP2A, STAT1
SPP1	TP53, FOXJ3 DEPDC6,EST1,GLI1, JUN, HOXC8
IL8	TP53, HSF2,LHX3,SOX21,IRF9, GATAD1, ZFR2,NFKB2,JUN,RELA, ZNF33B
ADH1C	TCF3,NFYA,ELF5,NFIC,TBP,DBP
CHI3L1	SP4,MAX, PARP12
ADH1B	ATF4,DBP,FOXC1,MTA1,CEBPB
MUC4	RCOR1,STAT5A,ZNF43,ZNF764,SMAD4
PDE9A	LHX6,ZNF665,ATXN7,DSP,GLI1
ANPEP	HOXD1,NFYA,NR2F1,ANKZF1,ETS2
UGT1A1,UGT1A2, UGT1A3,UGT1A4, UGT1A5,UGT1A6, UGT1A7,UGT1A8, UGT1A9	TBX21,NEUROD1,H1F0,HNF1A,RARA,SP1, RESTKLF2, NEUROD1, IRF7, ZMAT4, RBM22,SLC22A4,PPARG
UGT1A1,UGT1A2, UGT1A3,UGT1A4, UGT1A5,UGT1A6, UGT1A7,UGT1A8, UGT1A9,IL8	IRF7,ZMAT4
CLCA1	NR1H3,LHX6,RELA,GLI1
SST	ZNF287,FOXJ3,RNF113A,PAX6,CEBPE, ATF1,ATF2,ATF4,CREM

In TFactS we first need to give the target gene(s) names as input. The *P value, Evalue, Q value and FDR (Benjamini-Hochberg)* thresholds are set as 0.01. They are given to control the rate of false positives for multiple testing conditions [15]. Remaining parameters are left as it is (default value). After the computation is complete we check the result in next step by clicking the link “Submitted Lists and the corresponding TFs”. Here it gives us the corresponding TFs for the particular target genes. As mentioned in these steps we have given the DE gene (from Table III and Table IV) names as input (target) and noticed that most of the TFs discovered by our algorithm to be significant against that DE gene using TFactS.

TABLE IV. BEST COMBINATION OF TF PAIRS CORRESPONDING TO TARGET GENES FOR HUMAN COLON CANCER DATA THROUGH SPLINE REGRESSION MEASURE

Target (DE)	TF pair(s)
CA1	CRX, PAX2,PAX5, SP140 SEMA4A
GCG	PAX2,POU6F2,ZNF236,NKX6-1, ZNF638,ZC3H10,FOXA1,CREB1
INHBA	ATF1,CREB1,DHX57
CHGA	NR2E3,MYF6,ZNF236,NEUROD6, ZSCAN12,ZNF155,LASS6,ZNF638, CBX2,EGR1,TFAP2A,ATF1
SPP1	HOXA9,MYB,SMAD1,KLF10,ETV4, ZNF750,ZSCAN16,ID4,POU5F1,TP5, DLX5,CTNNB1,ETS1,GLI1,HOXC8,
IL8	TP53, NFE2L3,SMAD5,MSX2, ZNF444,STAT2,VENTX, BACH2,TCF20,MET, RORA,CDX1,TEAD4,ZNF257, TOX3, MET
ADH1C	CEBPB,NFYA,ELF5,SMARCA1,TBP,DBP
CHI3L1	ELF4,BACH2,SPI1,MLLT3,YEATS2, GTF3A,JRKL,USF1
SLC26A2	TFAP2C,CTNNB1,RBPJL,SP1,EMX1, ZNF257,NEUROG2,PLEKHA4,
ADH1B	CEBPA,CEBPB,BACH2,HHEX, HOXB2,ZNF155,WNT8B
MUC4	ZNF236,ST18,HOXC10,ZNF750, SMAD7, TFAP2B
PDE9A	ELF4,GLI1,TFAP2A,FOXJ3,BMP2
CEACM7	EZH2,PLAGL2,SRY,RERE,HMGB3, MBNL2,GLI2,NFAT5,HOXB6, BCL11B,PBX1,ZNF177
ANPEP	ELF4,PHOX2B,IRF8,ESRRA,HLF, TSC22D2,CUL3,EST1,EST2
UGT1A1,UGT1A2, UGT1A3,UGT1A4, UGT1A5,UGT1A6, UGT1A7,UGT1A8, UGT1A9	RARA,CDX1,GATA6,HOXD12, CHD7,HNF1A,PPARG,HHEX, NPAS2, MNX1,ZBTB3,TOX3,RAPGEF
CLCA1	HOXA9,ATF2,HOXC13,GLI1,BMP2, ZC3H7B,
SST	GATA1,DLX2,NR4A3,SRY, NEUROD4,C11orf9,PLEKHA4, CREB1,CEBPA,CEBPG
HSD17B2	RBPJL,PGR,HOXC6,EN1,FBN1, CTNNB1,

On the other hand the corresponding steps to use the PRISM tool are given below:

At first we need to choose the species on which the computation is being conducted. Accordingly we have selected “Human NCBI build 36.1”.

In the subsequent step we have to give either the target gene name or the Transcriptional regulator name. If we choose to give the target gene name as input then like TFactS it will give us TFs corresponding to that target gene, but if the inputs are TFs then it will give us target gene names (controlled by those TFs) as output. In addition to these results, PRISM also gives us the corresponding *ontology*, *biological context*, *E value*, *P* *value*, *fold enrichment*, *genes hit* and *binding sites* as output.

Similar to TFactS, in PRISM also we have given D.E gene names as input in search of TFs. Here also we have noticed that most of the TFs discovered by our algorithm are present as output for that DE gene entry. However in few cases TFs discovered (against some DE genes) by our algorithm are not present directly as output, but are present as *similar proteins* corresponding to the Transcriptional Regulators found for that target (D.E) gene by PRISM.

In Table V we have given some examples of TFs corresponding to DE genes found by PRISM which tally our result produced by MI method. Apart from the matching we have also given the *E value*, *P value* and *fold enrichment.*

In Table VI we have enlisted similar results but this time tallying with the spline regression method.

TABLE V. VALIDATING MUTUAL INFORMATION BASED TFS CORRESPONDING TO TARGET GENES USING PRISM WITH CERTAIN SIGNIFICANT SCORES

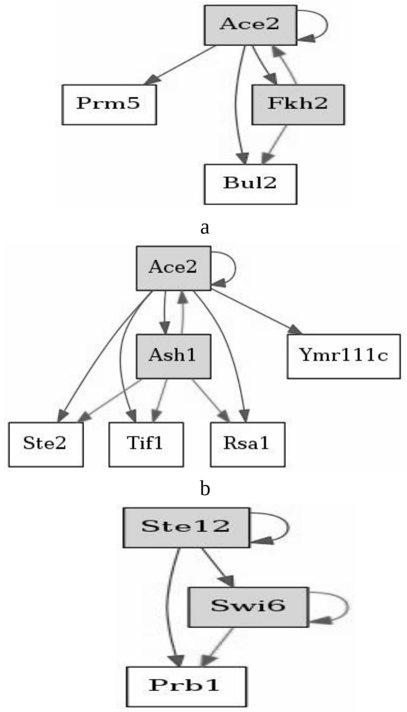
Target	**TF**	**E-value**	**P-value**	**Fold enrichment**
MUC4	STAT5A (Similar Protein to STAT1)	0.000	1.07E-16	2.16
IL8	JUN	0.116	2.71E-11	2.12
SPP1	JUN (Similar Protein to JPD2)	0.000	1.25E-42	2.02
GCG	STAT1	0.116	1.43E-22	2.57
SST	PAX6	0.349	6.21E-09	2.12

TABLE VI. VALIDATING SPLINE REGRESSION BASED TFS CORRESPONDING TO TARGET GENES USING PRISM WITH CERTAIN SIGNIFICANT SCORES

Target	**TF**	**E-value**	**P-value**	**Fold enrichment**
MUC4	TFAP2B	0.697	7.16E-06	2.62
CA1	CRX, PAX2, and PAX5 (both similar protein to CRX)	0.697	2.78E-13	3.25
IL8	BACH2	0.116	4.29E-25	2.06
SST	DLX2 (Similar Protein to BARHL2)	0.349	5.99E-08	2.14

IV. CONCLUSION AND FUTURE WORK

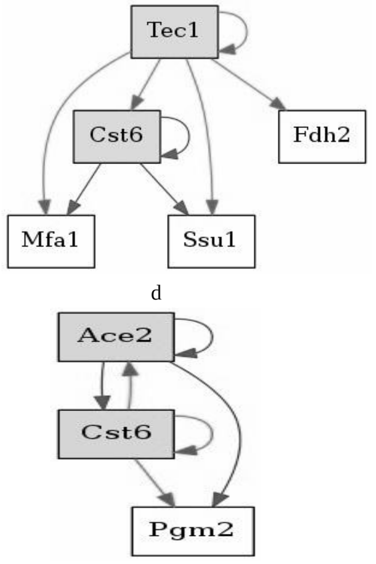
In this work we have extended a recently proposed procedure [4] to find the regulation between the Transcription Factors (TF) and differentially expressed (D.E) genes. D.E genes are those genes that have different expression values across varied conditions.



a

b

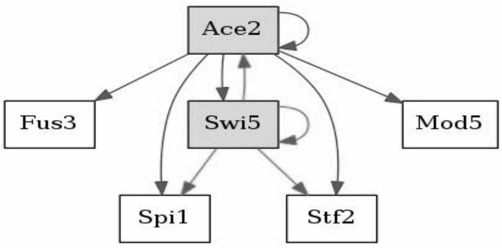
c



d

e

Figure 1. Some significant TF pair-target gene regulatory networks *using Mutual Information Adjacency Measure.*



a

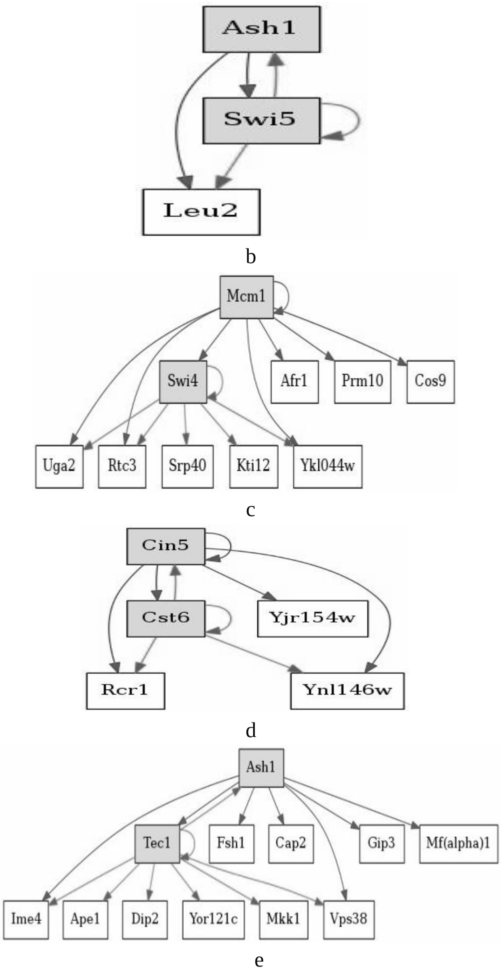


Figure 2. Some significant TF pair-target gene regulatory networks *using Spline Regression Measure*

Assuming in a Gene Regulatory Network it is the Transcription Factor protein which can affect the cell division procedure we can say that TFs must be regulating the target (DE) genes across both the conditions. The thought behind this work extends our previously introduced correlative procedure to find the best possible combination of TFs which can regulate a target gene highlighting only the linear regulation of target gene by TFs. Recent microarray studies yield datasets having large number of expression values, where the nonlinear regulatory nature may be more prominent than the linear one. To enforce this conception we have investigated the nonlinear regulation of targets by T.F genes through mutual information and spline regression based measures and tested it using yeast cell cycle as well as by human colon cancer dataset as the dimension (number of samples) of human colon cancer data is high allowing us to correctly reflect the effect of nonlinear regulations on large dataset [16]. We have taken a pair of TFs to minimize false negatives [17]. Here, as mentioned in the ALGORITHM we are considering the co-regulating effect of the TFs to the targets across both conditions to predict the TF pairs.

We have also gone through a comparison between the current nonlinear association based procedures with the previous linear association based procedure. Now while comparing the current method with the previous one we have used the result obtained for yeast cell cycle data because the previous work was carried out on this data only. Here, we have found that the numbers of possible combinations of TFs are high in linear method compared to the nonlinear counterpart. This is may be due to the small dimension of the data where as stated previously linear measure works better than the nonlinear counterparts [16]. We have also found that a good number of TF pair combinations found by the linear approach is present in either of the nonlinear based methods applied out here. Again the number of common TF pairs which have higher number of differentially regulated genes through nonlinear methods compared to our earlier linear approach are found to be more significant than the corresponding outcomes having more D.E genes in our previous linear correlative method compared to the nonlinear counterpart.

As mentioned in the *Results and Discussion* validation of these TF pairs corresponding to a target (D.E) gene have given transcriptional regulatory networks through which we can understand rolls of the TFs to control the target genes. Gene regulatory networks formed by the spline regression method are more prominent than mutual information based method.

In our work to date we have incorporated linear and nonlinear interactive cases to unveil the relationship between proteins and corresponding target genes. We can further extend our work in a protein-protein interaction network by applying these techniques to check out the interactions between *bait* and *prey* proteins to investigate which functional transcriptional factors are formed by them. This is useful especially when the hub protein is *date hub* [18] (Here interaction partners are expressed at different times. So we can assume that it will create many pair wise interactions; each time with a different partner).

Further extension can be performed by broadening the concept given in [19]. Interaction (association) value between proteins will not only give the influence of one towards another but it can also suggest whether proteins are *localized* or not. For annotation of new proteins, particularly for indirect interactions this procedure will cite a new direction in the domain of protein-protein interaction networks.

References

[1] A. T. Kwon, H. H. Hoos, and R. Ng, “Inference of transcriptional regulation relationships from gene expression data,” *ACM*, 2003.

[2] J. M. P. Desterro, M. S. Rodriguez, and R. T. Hay, “Regulation of transcription factors by protein degradation,” *Cell. Mol. Life Sci*, 57, pp. 1207-1219, 2000.

[3] L. K. Yeung, L. K. Szeto, W. C. Liew, H. Yan, “Dominant spectral componenet analysis for transcriptional regulations using microarray time series data,” *Oxford University Press*, Jan 2004.

[4] A. Majumder and M. Sarkar, “Simple transcriptional networks for differentially expressed genes,” in *Proc. ICSPCT*, pp. 642-647, 2014.

[5] L. Paninski, “Estimation of entropy and mutual information,” *Neural Computation*, pp. 1191-1253, 2003.

[6] L. Song, P. Langfelder, and S. Horvath, “Comparison of co-expression measures: Mutual information, correlation, and model based indices,” *BMC Bioinformatics*, vol. 13, no. 328, 2012.

[7] W. K. Chen, *Feedback, Nonlinear and Distributed Circuits, CRC* Press, 3 rd ed., pp. 9-20, 2009

[8] [Website] http://genome-www.stanford.edu/ ; http://www.yeastgenome.org/ (Last accessed on January 2014)

[9] S. P. Hussain, L. J. Hofseth, and C. C. Harris, “Radical causes of cancer,” *Nature*, vol. 3, pp. 276-285, 2003.

[10] L. Wang, Z. Fenq, X. Wang, X. Wang, and X. Zhang, “DEGseq: An R package for identifying differentially expressed genes from RNA-seq date,” *Bioinformatics*, vol. 26, no. 1, pp. 136-8, January 2010.

[11] B. M. Ryan, *et. al.*, “Germline variation in NCF4, an innate immunity gene, is associated with an increased risk of colorectal cancer,” *Int J. Cancer*, vol. 134, no. 6, pp. 1399-1407, March 2014.

[12] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe, “A census of human transcription factors: Function, expression and evolution,” *Nature*, vol. 10, pp. 252-263, April 2009.

[13] P. T. Monterio, *et al*., “Yeastract-discoverer: New tools to improve the analysis of transcriptional regulatory associations in Saccharomyces cerevisiae,” *Nucl. Acids Res (Oxford University Press)*, vol. 36, pp. D132-136, 2008.

[14] A. M. Wenger, *et al*., “PRISM offers a comprehensive genomic approach to transcription factor function prediction,” *Genome Res*., vol. 23, no. 5, pp. 889-904, May 2013.

[15] A. Essaghir, F. Toffalini, L. Knoops, A. Kallin, J. V. Helden, and J. B. Demoulin, “Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data,” *Nucleic Acids Research*, vol. 38, no. 11, March 2010.

[16] S. Chatterjee and A. S. Hadi, *Regression Analysis by Example*, 4th ed., Wiley, 2006, ch.2, pp. 21-45

[17] W. S. Wu and W. H. Li, "Systematic identification of yeast cell-cycle transcription factors using multiple data source,” *BMC Bioinformatics*, December 2008.

[18] A. Jaimovich, “Understanding protein-protein interaction network,” Ph.D thesis, Hebrew University, Israel, 2010.

[19] A. Jaimovich, G. Elidan, H. Margalit, and N. Friedman, “Towards an integrated protein-protein interaction network: A relational Markov network approach,” *Journal of Computational Biology*, vol. 13, no. 2, pp. 145-165, 2006.

A. Majumder is associated with N.I.T Durgapur, India, as an assistant professor of the dept. of E.C.E. His research interests include gene differential **co**-expression analysis, biomedical signal and image processing.

M. Sarkar is associated with D.I.AT.M Durgapur, India, as an assistant professor of the dept. of E.C.E. He carries out research in the domain of differential gene expression analysis through statistical and soft computing techniques.