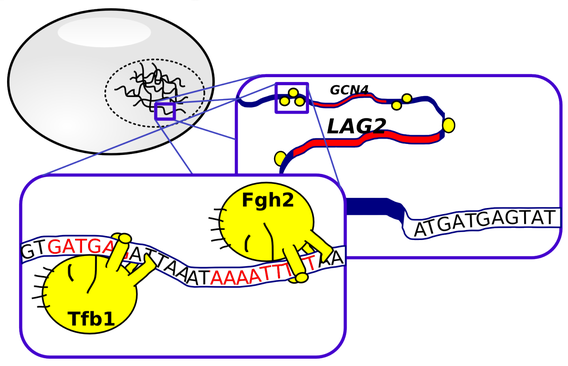
G=MAT:
Linking Transcription Factor Expression and DNA Binding

Konstantin Tretyakov *(kt@ut.ee),* ***Jaak Vilo*** *(vilo@ut.ee)*

Institute of Computer Science, University of Tartu

In Brief:

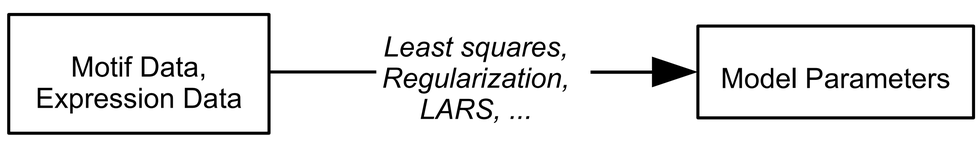
- Transcription factors *bind* to motifs on the DNA and thus influence the expression of genes.



- We attempt to *model* the effect that transcription factor binding has on *gene expression* using a certain *linear model*.
- Consider an example:

$$g \approx 2\, m_{CGG}\, t_{GAL1} + 0.5\, m_{GAT}\, t_{ROX3} - 3\, m_{TCA}\, t_{BYE1}$$

- Here we imagine that the expression of a gene depends *additively* on the expressions of transcription factors GAL1, ROX3 and BYE1, the latter here being a strong suppressor. However, the effect of GAL1 is only present if the gene has the CGG motif in the promoter, ROX3 will only act if the gene has the GAT motif, and BYE1 needs the motif TCA to act.
- The coefficients 2, 0.5 and -3 in the example model tell us about the effect of each transcription factor, as well as about the potential relationship between the factor and the motif. Pairs of transcription factors and motifs that have no effect on gene expression are implicitly present in the model with coefficients 0.
- We don't actually *know* the coefficients. But if we take the expression and motif data, we can *derive* the best-fitting coefficients. These can tell us something about motif and transcription factor significance and relations.

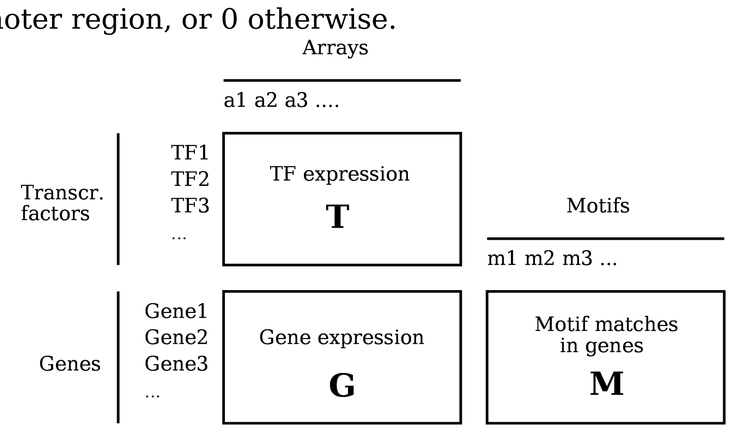


- Experiments show that the method can indeed detect interesting protein-motif connections.
- The name “G=MAT” stems from the matrix form of the model. Read on.

In Detail:

The G=MAT Model

Let $\mathbf{G} = (g_{ij})$ denote an *expression matrix*, i.e. g_{ij} is the expression of gene i on microarray j. Let $\mathbf{T} = (t_{kj})$ denote the transcription factor expression matrix, i.e. t_{kj} is the expression of transcription factor k on microarray j. Finally, let $\mathbf{M} = (m_{il})$ denote the *motif matrix*: m_{il} is 1 if gene i has motif l in the promoter region, or 0 otherwise.



We shall be interested in *predicting* the gene expression values g_{ij} using both motif and transcription factor expression information. We shall exploit the following model for that:

$$g_{ij} = \sum_{k,l} \alpha_{kl} m_{il} t_{kj} + \varepsilon\,,$$

where each parameter α_{kl} tells us about the potential regulatory role of transcription factor k in tandem with motif l, and ε is the random noise not accounted for by the model.

It turns out, that the above equation can be conveniently expressed in matrix form as

$$\mathbf{G} \approx \mathbf{MAT},$$

where $\mathbf{A}$ is the matrix of coefficients (α_{kl}).

The least squares fit to $\mathbf{A}$ can be computed quite efficiently as:

$$\mathbf{A}_{\mathrm{LS}} = (\mathbf{M}^{\mathrm{T}}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{G}\mathbf{T}^{\mathrm{T}}(\mathbf{T}\mathbf{T}^{\mathrm{T}})^{-1}.$$

Due to the relative scarcity of data, it mostly makes sense to fit the model in presence of *regularization*. Although classical l_2-regularization does not result in a nice expression, a result close to it can be obtained by computing

$$\mathbf{A}_{\mathrm{RR}} = (\mathbf{M}^{\mathrm{T}}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{G}\mathbf{T}^{\mathrm{T}}(\mathbf{T}\mathbf{T}^{\mathrm{T}} + \lambda\mathbf{I})^{-1}.$$

We have also implemented true l_2-regularization using gradient descent, as well as a specialized version of the *least angle regression (LARS)* algorithm for l_1-regularization. All versions were efficient enough to be usable with yeast data (~6000 genes, ~300 transcription factors, ~100 motifs, ~100 arrays).

Experimental Evidence:

We performed a number of experiments to confirm the viability of the approach. Consider, for example, the following set up:

Expression data:	Yeast cell cycle data by Spellman et al. (1998)
Motif data:	38 S. cerevisiae motifs from Transfac.
Sequence data:	800bp upstream sequences from SGD.
Transcription factors:	317 genes annotated in GO as transcription regulators.

We obtain the model fit $\mathbf{A}_{\mathrm{RR}}$ using the simple regularized G-MAT as presented above. Next we need to determine which of the coefficients α_{kl} are large enough to be considered significant. For that we perform a randomization experiment: repeatedly randomly permute the rows and columns of $\mathbf{G}$, fit the model and obtain the *distribution* of each coefficient in these random permutations. We then compare the true value of each coefficient with the distribution of values of this coefficient under random permutations and use its *p-value* and *z-score* to detect significance.

The *p-value* is the percentage of random instantiations of the coefficient greater than the true value. *Z-score* is the distance of the true value from the mean of the random distribution, divided by standard deviation.

Here are the 10 coefficients with the largest z-score resulting from that experiment:

Z-score	Motif	Transcription Factor
9,94	F$GAL4_01 (Bram et al, 1986)	GAL1, galactokinase, regulated by Gal4p.
7,95	F$MCM1_02 (Spellman et al, 1998)	SFG1, nuclear protein, putative TF.
7,45	F$GAL4_01 (Bram et al, 1986)	GAL3, involved in activation of GAL genes.
6,28	F$GAL4_01 (Bram et al, 1986)	GAL80, involved in repression of GAL genes.
6,16	F$MCM1_01 (Wynne et al, 1992)	ASH1, zinc-finger inhibitor of HO transcription.
5,27	F$MCM1_02 (Spellman et al, 1998)	ACE2, activates early G1-specific genes.
4,62	F$RAP1_C	IME1, master regulator of meiosis.
4,45	F$MCM1_02 (Spellman et al, 1998)	WTM2, transcriptional repressor, involved in meiosis.
4,28	F$STRE_B	WTM1, transcriptional repressor, involved in meiosis.
3,82	F$MCM1_02 (Spellman et al, 1998)	SWI5, activates transcription at M/G1 phase boundary.

It can be clearly seen how the transcription factors relevant for the cell cycle ended up with high scores. Also the association between the GAL motif and the corresponding group of transcription factors was nicely detected.

References:

H. J. Bussemaker, H. Li, E. D. Siggia, *Regulatory element detection using correlation with expression.* Nat. Gen., 2001.

L. A. Soinov. *Supervised classication for gene network reconstruction.* Biochem. Soc. Trans., 2003.

M. Middendorf, et al. *Predicting genetic regulatory response using classication.* Bioinformatics, 2004.

J. Ruan, W. Zhang. *A Bi-dimensional regression tree approach to the modeling of gene expression regulations.* Bioinformatics, 2005.