5 Gene regulation functions

Regulation of gene expression is one of the main control mechanisms in cells. Biochemically, mRNA transcription is controlled by regulatory proteins (e.g. σ factors and transcription factors), which bind to regulatory sites on the DNA and modulate the promoter activities of genes or operons.

To obtain dynamic models of gene networks, the simple qualitative arrows need to be replaced by quantitative ***gene regulation functions***, the rate laws of transcription. Gene regulation functions have been determined accurately for individual promoters(e.g. for the Lac operon in *E. coli*) by fitting predicted mathematical functions to measured transcription data. Based on high-throughput expression data, simple gene regulation functions and regulator activities can be estimated even for larger transcription networks.

5.1 Equilibrium binding of transcription factors

5.1.1 Transcription factor binding to promotor

In boolean models, gene regulation is described as an all-or-none decision. In kinetic models, it is described by continuuous functions that *arise* from a model of transcription factor (TF) binding (all-or-none on the microscopic level.) In the following, we assume (i) a binding equilibrium for transcription factors and (ii) a certain average transcription rate in each binding state.

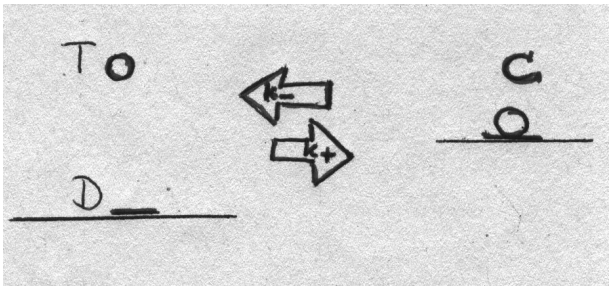

Figure 1: Example of transcription factor binding to a promotor

We treat the following example (see Fig.1): T is an active, free transcription factor, D is the DNA binding site and C is the complex, which accures when T and D bound together. $T + D \Longleftrightarrow C$, with $D_t = D + C$ (1 DNA molecule per cell $\rightsquigarrow$* $1\mu m^3$ in e.coli) $\rightarrow C = D_t - D$

The rate equation for the concentrations are:

$$\frac{dC}{dt} = k_+ * T * D - k_- C \tag{1}$$

Chemical eqilibrium:

$$C = \frac{T * D}{k_-/k_+}, \tag{2}$$

with $\frac{k_-}{k_+} = K_D$ =dissociation constant.

The greater the value of k_D, the more easily the complex will dissociate because the binding energy is lower. Thus, the binding energy determines k_D. We equal the equations for C and get:

$$D = \frac{D_t}{1 + T/k_D} \tag{3}$$

and further:

$$C = \frac{D_t T}{k_D + T}. \tag{4}$$

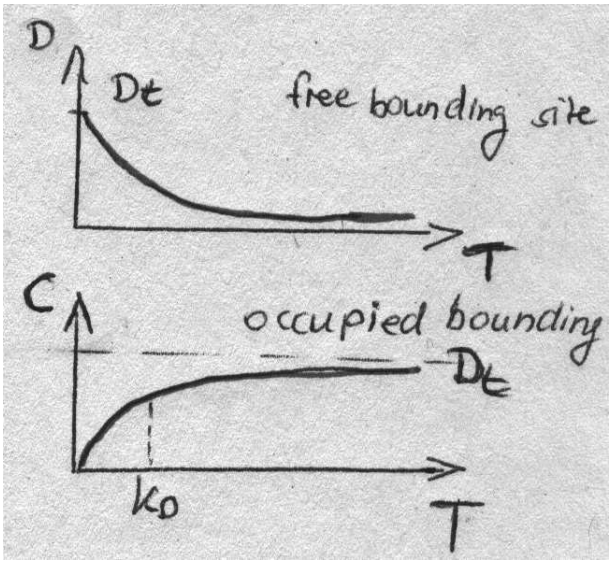

Figure 2: Concentration of DNA-binding sites (top) and the concentration of complexes (bottom) plotted against the concentration of active transcription factor T molecules.

At the concentration $x = k_D$, exactly half of the binding sites are empty (see Fig.2). For $x \to \infty$, $C \to D_{tot}$. We get the probability of a single free binding site with the equation: $p = \frac{1}{1+T/k_D}$

5.1.2 The transcription rate for a repressor gene

If no repressor is bound, Pol II binds and the transcription starts. The maximum transcription rate β depends on the promotor quality and can be changed with single-point mutations.

If we assume constant transcription for the empty promoter and no transcription at all for the represser-bound promoter, the promoter activity (=transcription rate) with repressor is given by $Y = \frac{\beta}{1+R^*/k_D}$, where R^* is the activated repressor.

5.1.3 Cooperative binding

A transcription factor (TF) can be a dimer, tetramer, et cetera containing various identical subunits with n binding sites. Now we are only interested in occupied transcription factors because the cooperative binding has to be examined. Consider the reaction of an unbound to a bound TF: $n * S + x \Longleftrightarrow \bar{x}$

- Bound TF: $\bar{x} = \frac{x_t S^n}{{k_x}^n + S^n}$
- Free TF: $x = \frac{x_t}{1+S^n/{K_x}^n}$

We obtain a step-funtion with a step at k_x for $n \to \infty$ (see Fig. 3).

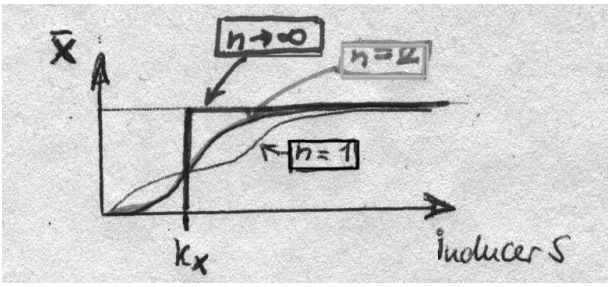

Figure 3: Relation between the number of inducer molecules S and bound TF x

5.2 Gene regulation functions derived from equilibrium binding: the general case

A kinetic law is used for the transcription rates to describe gene expression. The rate y is then given by a ***gene regulation function***

$$y(t) = f(\boldsymbol{x}(t), \boldsymbol{p}). \tag{5}$$

y is the transcription rate of a gene. It depends on regulator activities x_l. The parameter vector $\boldsymbol{p}$ and the mathematical form of f are specific for each gene (however, different genes may be modeled with the same functional form). The vector $\boldsymbol{x}$ contains the activities of all regulators for the gene. In eukaryotes, promoters can process a large number of inputs. They have complicated input functions. A **gene input function** f_i describes microscopic processes like binding of regulators. It is determined by the nucleotide sequence of the promoter region. Fig. 4 shows different binding states of the Lac promoter in a simplified scheme with five states. Figure (5) shows the relationship between promoter sequence and gene input function: a gene

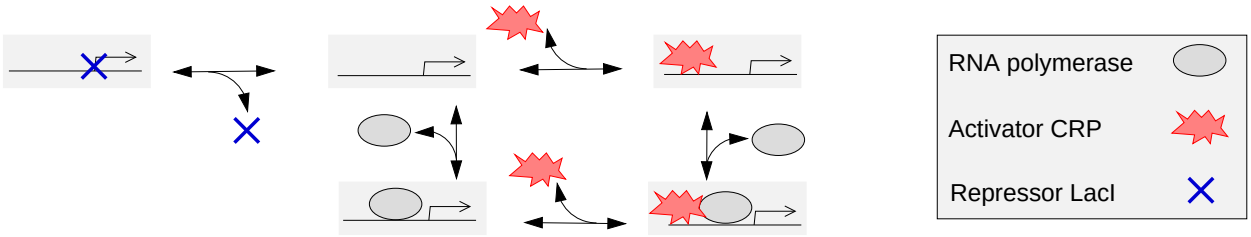

Figure 4: Microscopic states of the Lac promoter (schematic model). Bound activator increases the probability of polymerase binding (right). Transcription can occur in states with bound polymerase (bottom). The promoter can be bound by RNA polymerase and the activator CRP. The bound repressor LacI inhibits the binding of other molecules (left).

promoter can assume various microscopic states, characterized by different regulators bound to its binding sites and by different conformations of the DNA. Two basic assumptions For a quantitative model:

- There is a thermodynamic equilibrium between the different states. The probability for each state depends on its binding energy and the regulator molecules availability.
- The transcription initiation occurs randomly at a certain rate in each state.

Each conformation state of the gene input function f_i is characterized by a free energy $F = E - TS$ where E and S denote the energy and the entropy of the state and T is the temperature. On the one hand, the free energy F captures energies related to regulator binding or bending in DNA loops and these energies depend on presence and sequences of regulator binding sites (of the promotor sequence). On the other hand, the entropy term depends on the number of free regulator molecules. The free energy F of a promoter

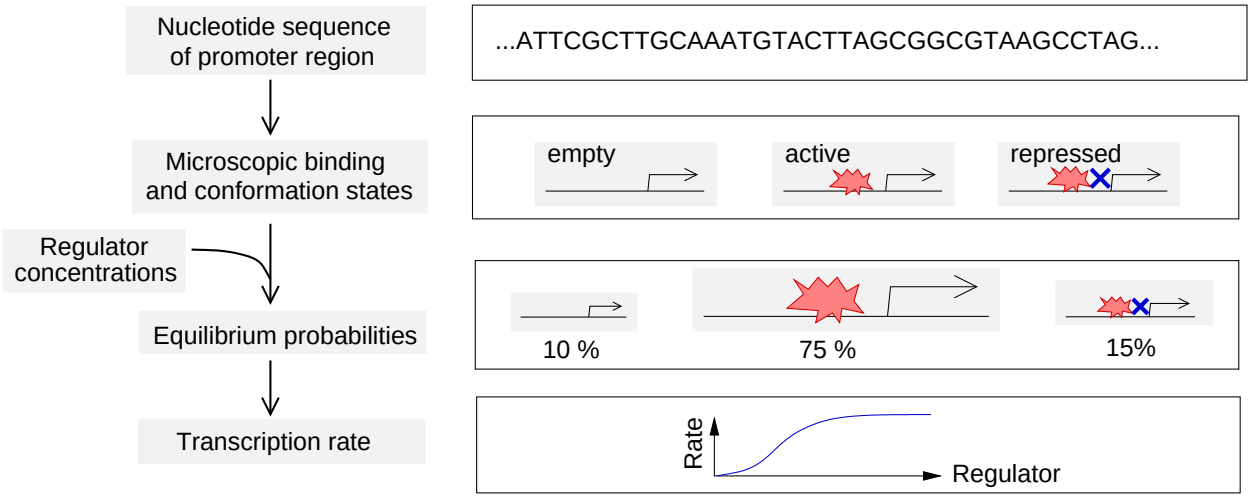

Figure 5: Schematic relation between nucleotide sequence and transcription rates. There is shown the transcription function of a single transcription factor (= activator), (bottom, right). The function of an inhibitor is reversed.

state determines its statistical weight $w_i = \exp(-F_i/(k_B T))$ in the Boltzmann distribution, and the total transcription rate:

$$y = \frac{\sum_i w_i v_i}{\sum_i w_i} \tag{6}$$

is computed as the weighted average over the synthesis rates in the different states.

We obtain an expression for the gene regulation function, if we write the transcription rate as a function of regulator concentrations .

5.3 The Lac operon in *Escherichia coli*

Metabolites can control the enzymes that catalyze their own production or consumption. With the resulting feedback loop the protein levels can be constantly adapted to the current needs of the cell. *Escherichia coli* bacteria prefer glucose as their energy source. For this reason, they sustain enzymes for glucose metabolism under all conditions. Bacteria can utilize other sugars such as lactose except of glucose. The enzymes β-galactosidase, permease, and thiogalactoside transacetylase (they are important for the consumption of lactose) are coded and regulated together in transcription unit *Lac operon*.

When cells are shifted from a glucose-rich medium to a glucose-free, but lactose-rich medium, they need an adaption time before they can assimilate lactose (at a high rate). The expression level of the Lac operon is increased when glucose is missing and lactose is present. A strong Lac expression follows the logical rule 'low glucose AND high lactose' (in approximation).

The transcription rate is controlled by combining two (biochemical) signals (Figure 6 (a)). On the one hand, a high glucose level decreases [cAMP], an intracellular messenger that activates the transcriptional activator CRP. (That is why at high glucose levels, CRP remains inactive, and Lac transcription is low.) Lactose, on the other hand, is sensed via allolactose, an isomer formed by converting the 1-4 bond of lactose into a 1-6 bond. Allolactose activates the transcriptional repressor LacI, which shuts down Lac expression by blocking the binding of polymerase and by promoting a DNA loop. If no lactose is present, Lac expression will also be low.

The Lac operon becomes a strong expression, if the repression is released and the activator CRP is bound, (happening when Glucose is absent but Lactose is available). Experiments have shown, CRP and LacI (the

regulators) can be controlled by the extracellular levels of cAMP and IPTG, (a substitute for allolactose) shown in Figure 6 (b).

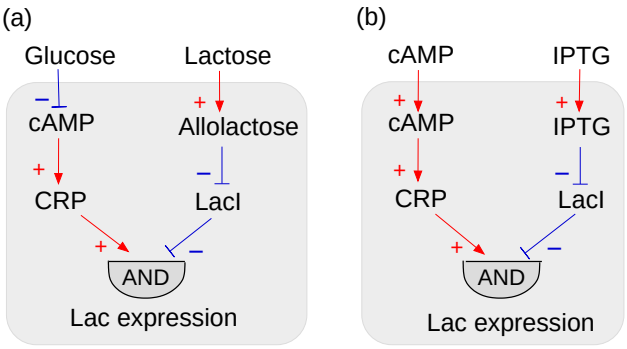

Figure 6: Regulation of the Lac operon. (a) The Lac operon is controlled by the transcriptional regulators CRP and LacI, which respond to extracellular levels of lactose and glucose. High expression of the Lac operon requires that lactose is present and glucose is absent. (b) In an experiment, the activities of CRP and LacI are regulated by extracellular levels of the ligands cAMP and IPTG. Effectively, both substances activate Lac expression.

5.4 Gene regulation function of the Lac operon

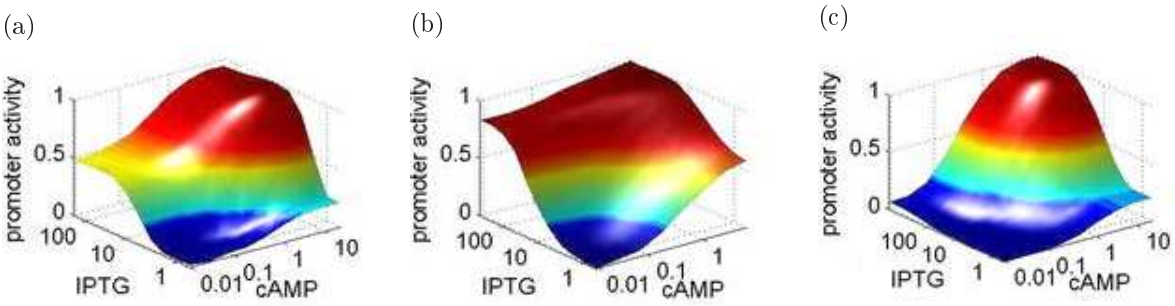

Figure 7: Gene regulation functions of the wild type Lac operon and two variants: (a) Gene input function in an *E. coli* wild-type strain. (b) An OR-like input function. (c) An AND-like input function.

Does the gene regulation function in living cells follow the above prediction? Setty et al. have experimentally determined the input function of the Lac operon in living *E. coli* cells. The transcription rates were measured by *GFP* under the control of the Lac promoter. The regulator activities were regulated via extracellular levels of cAMP and IPTG. The transcription rate is plotted against logarithmic concentrations of extracellular cAMP and IPTG. It shows four plateaus corresponding to the possible combinations of low and high concentrations (Fig. 7 (a)). At high cAMP and IPTG levels the transcription rate is high, too.

In contrast to the boolean input function, the expression rates for low cAMP and low IPTG are not exactly zero and this baseline activity has an important biological function:
in order to switch the Lac system to a high expression level, some lactose has to be imported into the cell to produce the messenger allolactose.(This requires that the lactose transporter LacY is already present, at least at a low level.)

If the levels of cAMP and IPTG are used as proxies for transcriptional activiators (inhibition of the repressor LacI effectively counts as activation), the simplified microscopic model in Figure (4) leads to a gene regulation function shown in Figure 7 (a)).

The binding energy is determined by the sequence of polymerase transcription factors an binding sites. Mutations in regulator binding sites will change the binding energies and thereby all the other parameters

$\Rightarrow$ Change the shape of the input function (= the adaption of the input function to new conditions)

Changes of the sequences of polymerase and transcription factor have a global impact, but changes of the promotor area only have lokal impact. For this reason, they are easier changed.

Point mutations (in promoter sequence of the Lac operon) lead to variants of the gene regulation functions (compare with Figure 7). The plasticity of gene input functions allows for evolutionary fine-tuning of the gene regulatory system. In the case of the Lac operon, a pure AND-like input function (Fig. 7) could have evolved rather easily.

5.5 Transcriptional regulation in larger networks

If the regulator activities $\boldsymbol{x}(t)$ and the transcription rate $y(t)$ for a gene regulation function (5) have been measured, the parameters $\boldsymbol{p}$ can be obtained from nonlinear regression. However, it is difficult to control or measure the active form of transcription factors. In the Lac operon study, for instance, external levels of IPTG and cAMP had to be used as controllable proxies. An alternative (if the regulator activities are completely unknown)is to compare the levels of different target genes and to estimate the regulator acitivities along with the gene input functions.

Microarrays allow to measure the mRNA levels of thousands of genes at the same time. The expression levels of a single gene, measured in different cell samples, form an expression profile. Such data contain: (i)valuable information about the regulators of a gene, (ii) their activities, and (iii) the corresponding gene regulation functions.

Data-driven methods like clustering or biclustering compute similarity measures between the expression profiles of different genes, assess their statistical significance, and hypothesize that genes with significant coexpression may be coregulated. Even if genes respond to the same regulators, their expression profiles may differ due to (i) different gene input functions ; (ii) additional transcription factors that control only some of the genes; (iii) different rates of mRNA degradation.

Dynamical models of gene expression can account for these effects and help to infer co-regulation more reliably than by using simple similarity scores. Most genes respond to several transcription factors, and transcription factors can regulate large numbers of target genes.

To determine the gene regulation functions from expression data, the effects of different transcription factors have to be disentangled. One such method is network component analysis, which uses simple linear gene regulation functions and can thereby tackle fairly large networks.

5.6 Network component analysis

Network component analysis (NCA) is a method to translate a known genetic network structure into a quantitative model of gene regulation. While the transcription rates and the transcription factors are known, the gene regulation functions need to be computed. In NCA, we assume linear gene regulation functions, so the temporal activity $y_i(t)$ of a promoter is a weighted sum of the regulator activities $x_l(t)$

$$y_i(t) = \sum_l a_{il}\ x_l(t). \tag{7}$$

The index t refers to different samples and can represent time points in an experiment. The input weights a_{il} indicate whether a regulator acts as an activator ($a_{il} > 0$), or as a repressor ($a_{il} < 0$), or has no effect ($a_{il} = 0$) on the promoter activity.

Network structures can be obtained from databases or from experiments. By these structures, many of the coefficients a_{il} are already limited to zero values. Known modes of regulation (activation/repression) may limit the signs of the remaining elements a_{il}.

The linear NCA model 7 resembles the statistical model used in principal component analysis. But in contrast, it is based on biological knowledge about the structure of the genetic network. To estimate the model parameters, we rewrite equation (7) as a matrix product

$$\boldsymbol{Y} = \boldsymbol{A}\ \boldsymbol{X}. \tag{8}$$

(see Fig. 8)

The matrix $\boldsymbol{A}$ contains the linear coefficients of input functions (rows: promoters, columns: regulators) and $\boldsymbol{X}$ contains the profiles of the regulators (rows: transcription factors, columns: time points or conditions). The structure of $\boldsymbol{A}$ (positions and possibly signs of non-zero entries) is prescribed by the network structure, and only the numerical values (the influence strengths) need to be determined from data.

The aim of NCA is to estimate the regulator activities $x_l(t)$ and the input weights a_{il} from measured expression values $y_i^{\exp}(t)$. We require that

$$\boldsymbol{Y}^{\exp} \quad \approx \quad \boldsymbol{A}\ \boldsymbol{X}. \tag{9}$$

with least square errors. Given a data matrix $\boldsymbol{Y}$ and the above-mentioned constraints on $\boldsymbol{A}$, the matrices $\boldsymbol{A}$ and $\boldsymbol{X}$ can be determined by an iterative optimization:

1. $\boldsymbol{A}$ is initialized with random values and $\boldsymbol{X}$ is chosen by least squares estimation.
2. $\boldsymbol{X}$ is kept fixed and $\boldsymbol{A}$ is updated
3. The mutual updating is iterated until convergence.

For ideal artificial data (obtained from an NCA model without noise), this biquadratic optimization converges to a global optimum for both matrices $\boldsymbol{A}$ and $\boldsymbol{X}$.

If this optimum is non-unique, (depending on the network topology) NCA models may be unidentifiable (because different parameter choices could lead to equally good results). Identifiability of the NCA model depends on the network structure. It can be checked by analyzing the wiring between regulators and their target genes.

The linear NCA model can also be interpreted in terms of nonlinear gene regulation functions: if the inputs x_l and outputs y_l represent logarithmic regulator activities $x_l = \ln c_l$ and logarithmic promoter activities $y_l = \ln v_l$, Eq. (7) is equivalent to a nonlinear gene regulation function of the form

$$v_i(t) = \prod_l (c_l(t))^{a_{il}} \tag{10}$$

for the original values c_l and v_i. This form accounts for multiplicative effects between regulators (but not for saturation).

Example: Assumption for the input function: $\bar{X}_{i(t)} = \bar{X}_{i(o)} * \prod \left(\frac{\bar{b}_{j(t)}}{\bar{b}_{j(0)}} \right)^{a_{ij}}$, with

$a > 0 \quad \rightarrow \quad$ activation
$a = 0 \quad \rightarrow \quad$ no effect
$a < 0 \quad \rightarrow \quad$ inhibition.

Consider the logarithms: $x = \log \frac{\bar{X}_{(t)}}{\bar{X}_{(o)}}$ and $b = \log \frac{\bar{b}_{(t)}}{\bar{b}_{(0)}}$
$\Rightarrow X_{i(t)} = \sum_j a_{ij} b_{j(t)}$.

Look at figure (8): The structure of matrix $\boldsymbol{A}$ is determined by the genetic network and the logarithmic

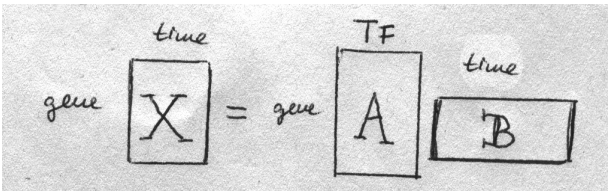

Figure 8: The matrix product usued in network component analysis.

data $\boldsymbol{X}$ comply the equation.
A Partition of this would be: $\boldsymbol{X} = \boldsymbol{A} * \boldsymbol{B} = \underbrace{\boldsymbol{A}S}_{\boldsymbol{A}'}\underbrace{S^{-1}\boldsymbol{B}}_{\boldsymbol{B}'}$ (− >S is the diagonal matrix)
Furthermore, the question arises: Is this partition unique or are there other partitions that fulfill the same structural condition?

Liao et al. developed in 2003 a criterion for the NCA, which says the partition is explicit, if:

1. $\boldsymbol{A}$ has full column rank.
2. $\boldsymbol{A}$ keeps its full column rank, although a freely chosen TF (and all of its target genes) is deleted.
3. $\boldsymbol{B}$ has full row rank.

All of these conditions have to be fulfilled by the matrices. To check, if everything is fulfilled, one can use random numbers and test.

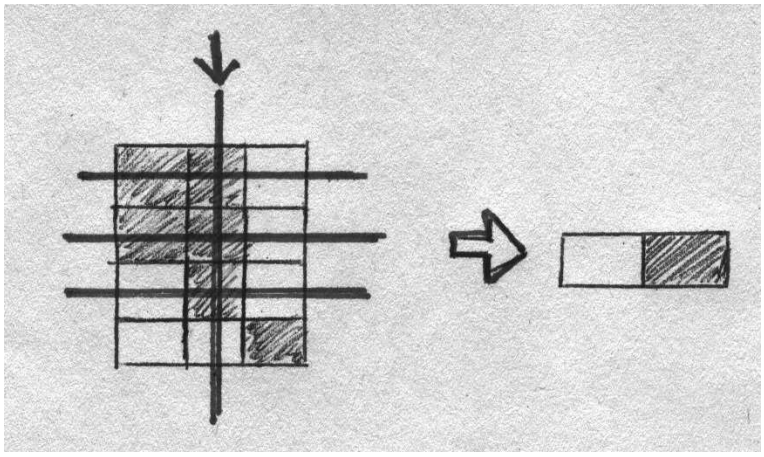

Figure 9: This is an example: the target genes of TF1 are target genes of TF2, too. In this case, the matrix has no full column rank and two columns are linearly dependent. For this reason, the second condition of the criterion of Liao et al. is not satisfied. The third condition requires that there have to be more points of time than TFs.

Figure (9) displays an example that does not satisfy the criteria.