Kalli Kappel
December, 2014

Gene Copy Number Project

Gene expression data describes expression for an exponentially growing population at steady state. We want to translate this into per gene copy expression. For each mRNA, i, we can solve for the rate of synthesis, $k_{synth,\ pop\ S.S.,\ i}$ using the following:

$$\frac{dm_i}{dt} = k_{synth,pop\ S.S.,i} - \left(\frac{\ln{(2)}}{h_i} + \frac{\ln{(2)}}{\tau_d}\right) m_i = 0$$

$k_{synth,\ pop\ S.S.,\ i}$ can be decomposed as follows:

$$k_{synth,pop\ S.S.,i} = (p_{pop,SS,i})(RNAP\ available\ to\ bind)\frac{k_{elong,i}}{L_i}$$

where $p_{pop,SS,i}$ is the population level steady state promoter strength, $k_{elong,i}$ is the transcript elongation rate, and L_i is the length of the transcript. The RNAP available to bind is assumed to be proportional to the concentration of RNAP and therefore constant in time. Now we want to decompose $p_{pop,SS,i}$ into an intrinsic promoter strength for a single copy of the promoter, p_i. To do this we need to know what fraction of cells in an exponentially growing population have one versus two copies of the gene (for doubling time of 60 minutes). In 1956, Powell derived an expression for this:

$$\phi(a) = 2\left(\frac{\ln{(2)}}{\tau_d}\right)e^{-\left(\ln{(2)}/\tau_d\right)a}$$

where $\phi(a)$ is the fraction of cells in the population at age a, where a is between 0 and τ_d, the doubling time. Then,

$$p_{pop,S.S.,i} = p_i \times 1 \int_0^{A_i} \phi(a)da + p_i \times 2 \int_{A_i}^{\tau_d} \phi(a)da$$

Where A_i is the age at which the gene is replicated.
Solving for $p_{i,}$

$$p_i = \frac{p_{pop,S.S.,i}}{2e^{-\left(\ln{(2)}/\tau_d\right)A_i}}$$

Assuming DNA replication happens at approximately a constant rate,

$$A_i = \frac{End\ position_i - OriC}{k_{DNA\ elongation}}$$

Then, during the simulation, at each timestep, we multiply p_i by the copy number of gene i and normalize.

Modifying the fitting procedure:
In the knowledge base, I set initial p_i values, $p_i(0)$, as described above. Then I also find the time average p_i value, $\overline{p_\iota}$, for each gene i. This is nontrivial because the time dependence comes from renormalization after the replication of each gene, so there is no analytical expression for the time evolution of p_i. To simplify the calculation, I fit an expression for $\overline{p_\iota}/p_i(0)$ as a function of A_i with the values from an initial simulation. Let's call this fit expression $f(A_i)$. Then, $\overline{p_\iota} = f(A_i) \times p_i(0)$. The current fitting procedure in fitkb1 fits population level steady state expression values. Instead, we want to fit single cell time average expression:

$$expression\ to\ fit_i = \overline{p_i}/\left(\frac{\ln{(2)}}{h_i}+\frac{\ln{(2)}}{\tau_d}\right)$$

Once this expression is fit, initial synthesis probabilities can be calculated as follows:

$$p_i(0) = \frac{fit\ expression_i \times \left(\frac{\ln{(2)}}{h_i}+\frac{\ln{(2)}}{\tau_d}\right)}{f(A_i)}$$

These are the p_i values that are used in the simulation. At each timestep, they are multiplied by the copy number of gene i, and then renormalized.

Variation in synthesis probability with gene copy number really only makes sense for genes that are constitutively expressed. For a temporary implementation that doesn't mess things up too much, I created a list of highly expressed genes, which include all the genes for rRNAs and rProteins. The genes in this list are not subject to any of the effects of gene copy number described above. Their synthesis probabilities are constant throughout the cell cycle and are based on the steady state population level expression data.