Transcription regulation

Model implementation. There are two aspects to modeling transcriptional regulation: (1) modeling the activation or inhibition of a transcription factor (e.g., by a ligand), and (2) given an active transcription factor, modeling its effect on RNA polymerase recruitment to a promoter site. The the enhanced coverage of the regulatory network - 438 regulatory interactions described by 22 transcription factors that regulate 355 genes - is a significant difference from the *M. genitalium* model. To incorporate this network, regulation is represented by three different classes of transcription regulators: zero-component systems, one-component systems and two-component systems.

Modeling transcription factor activation. We consider three classes of transcription factors based on their mechanism of activation:

1. **Zero-component systems**: transcription factors that are considered to be active whenever they are expressed. Examples include the Fis and Hns proteins. These two proteins, for instance, are important in maintaining higher-order DNA structure and likely have complex feedback loops modulating their activity. Because this complexity is not yet fully understood, we make the simplifying assumption that these proteins are always active unless they are knocked out. Zero-component systems are modeled in the `TfBinding` process in the model, which handles transcription factor binding to promoters.

2. **One-component systems**: transcription factors that are directly activated or inhibited by a small molecule ligand. Examples of this class include the repressor TrpR which binds tryptophan, and the inducer AraC which binds arabinose. One-component systems are modeled in the `TfBinding` and `Equilibrium` processes in

the model, which handle transcription factor binding to promoters and transcription factor binding to ligands, respectively.

3. **Two-component systems**: transcription factors that are paired with a separate sensing protein that responds to an environmental stimulus (these are simple analogs to the vast, complicated signaling networks that exist in eukaryotic cells). The sensing protein phosphorylates the cognate transcription factor in a condition-dependent fashion. Examples include ArcA which is phosphorylated by its cognate ArcB in anaerobic conditions, and NarL which responds to the presence of nitrate when phosphorylated by its cognate sensor NarX. Two-component systems are modeled in the `TfBinding`, `Equilibrium`, and `TwoComponentSystems` processes in the model, which handle transcription factor binding to promoters, transcription factor binding to ligands, and phosphotransfer reactions of signaling pathways, respectively.

Zero-component systems. We assume all transcription factors of this class will bind to available promoter sites.

One-component systems. For a transcription factor with concentration T whose activity is directly modulated by a ligand with concentration L that binds with stoichiometry n, we assume that the two species achieve equilibrium on a short time scale and that the affinity of the two molecules can be described by a dissociation constant K_d:

$$nL + T \rightleftharpoons T^* \tag{1}$$

where T^* represents the concentration of the ligand-bound transcription factor.

With the dissociation constant K_d defined as:

$$K_d = \frac{L^n \cdot T}{T^*} \tag{2}$$

we have:

$$\frac{T^*}{T_T} = \frac{L^n}{L^n + K_d} \tag{3}$$

where T_T is the total concentration of the transcription factor, both ligand-bound and unbound. As we can see, the fraction of bound transcription factor is a function of ligand concentration and the dissociation constant. Importantly, if the ligand concentration is (approximately) constant over time, the fraction of bound transcription factor is (approximately) constant over time.

To computationally simulate this model we start with total counts of free transcription factor and ligand, completely dissociated from one another. We then form one molecule of the ligand-TF complex at a time and evaluate how close the ratio of $L^n \cdot T/T^*$ is to the actual K_d. We select the values of L, T and T^* that minimize the absolute difference between K_d and $L^n \cdot T/T^*$ (see Algorithm 1).

Two-component systems. For a transcription factor with concentration T; a cognate sensing protein with concentration S; a ligand with concentration L; subscripts f denoting a free (unbound) form of a molecule, b denoting a ligand-bound form of a molecule, and p denoting a phosphorylated form of a molecule; and ATP, ADP, H^+, and H_2O denoting concentrations of these molecules, we propose a system with the following:

Free (unbound) cognate sensing protein at equilibrium with ligand-bound cognate sensing protein, described by dissociation constant K_d:

$$L + S_f \rightleftharpoons S_b \tag{4}$$

The autophosphorylation of a free (unbound) cognate sensing protein at a rate k_A:

$$S_f + ATP \overset{k_A}{\to} S_{fp} + ADP + H^+ \tag{5}$$

The autophosphorylation of a ligand-bound cognate sensing protein at a rate k_B:

$$S_b + ATP \overset{k_B}{\to} S_{bp} + ADP + H^+ \tag{6}$$

The phosphorylation of a transcription factor by its free, phosphorylated cognate sensing protein at a rate k_C:

$$S_{fp} + T \overset{k_C}{\to} S_f + T_p \tag{7}$$

The phosphorylation of a transcription factor by its bound, phosphorylated cognate sensing protein at a rate k_D:

$$S_{bp} + T \overset{k_D}{\to} S_b + T_p \tag{8}$$

The auto-phosphatase activity of a transcription factor at a rate k_E:

$$T_p + H_2O \overset{k_E}{\to} T + P_i \tag{9}$$

Ligand binding is simulated in a fashion identical to the one-component systems. By assuming mass-action kinetics, we can represent the rest of this system mathematically using ordinary differential equations:

$$\frac{dS_f}{dt} = -k_A \cdot S_f \cdot ATP + k_C \cdot S_{fp} \cdot T \tag{10}$$

$$\frac{dS_b}{dt} = -k_B \cdot S_b \cdot ATP + k_D \cdot S_{bp} \cdot T \tag{11}$$

$$\frac{dT}{dt} = -k_C \cdot S_{fp} \cdot T - k_D \cdot S_{bp} \cdot T + k_E \cdot T_p \cdot H_2O \tag{12}$$

$$\frac{dS_{fp}}{dt} = -\frac{dS_f}{dt} \tag{13}$$

$$\frac{dS_{bp}}{dt} = -\frac{dS_b}{dt} \tag{14}$$

$$\frac{dT_p}{dt} = -\frac{dT}{dt} \tag{15}$$

This system of equations is simulated using a numerical ODE integrator (see Algorithm 2).

Modeling the modulation of RNA polymerase recruitment. After modeling transcription factor activation, we need to model the probability that the transcription factor is bound to DNA, p_T, and, when the transcription factor is DNA-bound, its effect on RNA polymerase recruitment to the promoter site, Δr (see Algorithm 3). Recalling the notation used in the *Transcription* section (Algorithm **??**), we want to modulate the j^{th} entry in the v_{synth} vector of RNA polymerase initiation probabilities such that:

$$v_{\text{synth},j} = \alpha_j + \sum_i p_{T,i} \Delta r_{ij} \tag{16}$$

where α_j represents basal recruitment of RNA polymerase and the second term is dependent on transcription factor activity: the probability that the i^{th} transcription factor is DNA-bound is $p_{T,i}$, and the recruitment effect of the i^{th} transcription factor on the j^{th} gene is Δr_{ij}. The α values are either computed prior to the simulation or determined from the simulation ppGpp concentration as described in the section below. The Δr values are computed prior to simulation based on gene expression values from conditions that modulate transcription factor activity. Values for p_T are calculated as described in Table 1.

Transcription factor type	Promoter-bound probability
Zero-component system	$p_T = 1$ if TF is present, 0 otherwise
One-component system	$p_T = (T^*)/(T^* + T)$
Two-component system	$p_T = (T_p)/(T_p + T)$

Table 1: Formulas used to compute the probability that a transcription factor is promoter-bound. T^* is the active form of a one-component system transcription factor, while T_p is the phosphorylated form of a two-component system transcription factor, and T is the inactive or unphosphorylated form of a transcription factor.

Optional feature: ppGpp regulation. ppGpp regulation of gene expression can be enabled with a simulation option (`--ppgpp-regulation`). When enabled, it is assumed that ppGpp binds to RNA polymerases and the free and bound forms of RNA polymerase have different amounts of expression for each gene. The binding can be represented with a reversible reaction:

$$RNAP_{free} + ppGpp \rightleftharpoons RNAP_{ppGpp} \tag{17}$$

To calculate the impact of ppGpp on RNA expression the following equation is used:

$$exp_j = (1 - f) \cdot exp_{free,j} + f \cdot exp_{ppGpp,j} \tag{18}$$

where $exp_{free,j}$ and $exp_{ppGpp,j}$ represent the amounts of expression expected from the free and ppGpp bound RNA polymerases, respectively, and are determined prior to simulations. f is the fraction of RNA polymerases that are bound to ppGpp as defined below:

$$f = \frac{RNAP_{ppGpp}}{RNAP_{total}} = \frac{C_{ppGpp}^2}{K_M^2 + C_{ppGpp}^2} \tag{19}$$

where C_{ppGpp} is the concentration of ppGpp, K_M is determined prior to simulations and represents the Michaelis constant representing the concentration of ppGpp where half the RNA polymerases are bound. There is also a Hill coefficient of 2 representing the fact that ppGpp has multiple sites of interaction with RNA polymerase [**?**].

To calculate K_M of ppGpp binding to RNA polymerase, several assumptions are made. The first is that the amount of RNA in a cell at different growth rates changes due to changes in stable RNA expression. The second is that this change in RNA is controlled by ppGpp concentration changes which cause differential expression based on the amount of RNA polymerase that is bound. Finally, the amount of RNA polymerase bound to ppGpp follows the relationship in Eq. 19. Using population level data for RNA mass fraction, RNA polymerase concentration and ppGpp concentration at different cell doubling times from Bremer and Dennis [?] allows K_M (as well as overall expression rates from free and ppGpp bound RNA polymerase) to be determined with a least squares fit based on the following relationship:

$$RNA_{fraction,i} = C_{RNAP,i} \cdot ((1 - f_i) \cdot exp_{free} + f_i \cdot exp_{ppGpp}) \tag{20}$$

$$f_i = \frac{C_{ppGpp,i}^2}{K_M^2 + C_{ppGpp,i}^2} \tag{21}$$

where $RNA_{fraction,i}$ is the mass fraction of the cell that is RNA, $C_{RNAP,i}$ is the cellular concentration of RNA polymerase and $C_{ppGpp,i}$ is the cellular concentration of ppGpp for each growth rate i.

ppGpp dependent fold change data and basal expression data can be used to solve for $exp_{free,j}$ and $exp_{ppGpp,j}$ for each gene, j, using data from Sanchez-Vazquez *et al.* [?]. To ensure consistency with annotated ppGpp regulation, only genes known to be regulated by ppGpp as curated from EcoCyc [?] and with consistent regulatory direction with the fold change data are regulated by ppGpp in the model. The fold change data from Sanchez-Vazuez *et al.* represents the change from an uninduced (low ppGpp) condition in rich media to a RelA induced (high ppGpp) condition. Assuming the overexpression of RelA leads to ppGpp concentrations that are much greater than K_M so that all RNA polymerases are bound to ppGpp, the fold change can be represented as:

$$FC_j = \log_2 \frac{exp_{induced,j}}{exp_{uninduced,j}} = \log_2 \frac{exp_{ppGpp,j}}{(1 - f_{rich}) \cdot exp_{free,j} + f_{rich} \cdot exp_{ppGpp,j}} \tag{22}$$

where FC_j is the measured fold change for gene j and f_{rich} is the fraction of RNA polymerases bound to ppGpp in rich media as determined by Eq. 21 with the concentration for ppGpp in rich media.

Measured RNA expression in M9 + glucose conditions (basal) provides another relationship between measured data and $exp_{free,j}$ and $exp_{ppGpp,j}$ for each gene:

$$exp_{basal,j} = (1 - f_{basal}) \cdot exp_{free,j} + f_{basal} \cdot exp_{ppGpp,j} \tag{23}$$

where $exp_{basal,j}$ is the measured RNA expression data for each gene, j, and f_{basal} is the fraction of RNA polymerases bound to ppGpp in basal media as determined by Eq. 21 with the concentration for ppGpp in basal media.

Taken together, Eq. 22 and 23 can be used to solve for the unknown values $exp_{free,j}$ and $exp_{ppGpp,j}$. This provides expression for nearly all genes but some adjustments are needed. First, some genes do not have fold change data (transcript is too small, difference was not significant, etc) but are annotated as being regulated by ppGpp. For those with positive regulation, the average fold change of all positively regulated genes, FC_+, is used. For those with negative regulation, the fold change calculated from fitting exp_{free} and exp_{ppGpp} in Eq. 20, FC_-, is used since this mostly represents the change in rRNA and tRNA expression which is not measured in the fold change data set. Another adjustment is needed for certain genes with high positive fold changes. In these cases, solving the two equations results in a negative value for $exp_{free,j}$. Since negative RNA expression is not possible, these values are truncated at 0.

Expression adjustments to ribosome and RNA polymerase related genes are currently done outside the framework of ppGpp regulation in order to get the appropriate doubling time in given conditions. To match ppGpp regulation expression levels to these expression adjustments, a least squares fit is performed to determine new expression values for the adjusted genes. Based on Eq. 18, a system of equations can be set up for the expression in each of three conditions (rich, basal, anaerobic):

$$F \cdot r = e \tag{24}$$

$$\begin{bmatrix} 1 - f_{rich} & f_{rich} \\ 1 - f_{basal} & f_{basal} \\ 1 - f_{anaerobic} & f_{anaerobic} \end{bmatrix} \cdot \begin{bmatrix} exp_{free,j} \\ exp_{ppGpp,j} \end{bmatrix} = \begin{bmatrix} exp_{rich,j} - tf_{rich,j} \\ exp_{basal,j} - tf_{basal,j} \\ exp_{anaerobic,j} - tf_{anaerobic,j} \end{bmatrix} \tag{25}$$

where f_{rich}, f_{basal} and $f_{anaerobic}$ are the fractions of RNA polymerase bound to ppGpp in different conditions, $exp_{free,j}$ and $exp_{ppGpp,j}$ are the gene specific expression values for free and ppGpp bound RNA polymerases, $exp_{rich,j}$, $exp_{basal,j}$ and $exp_{anaerobic,j}$ are adjusted expression values for genes of interest in each of the conditions and $tf_{condition,j}$ is the contribution to expression that is expected to be controlled by transcription factors as defined below:

$$tf_{condition,j} = \frac{exp_{condition,j} \cdot \Delta r_{condition,j}}{p_{condition,j}} \tag{26}$$

where $\Delta r_{condition,j}$ is the change in probability expected from the average transcription factor binding in the condition as described above and $p_{condition,j}$ is the expected synthesis probability in the condition (not considering ppGpp regulation). This calculation assumes that the ratio between expression and synthesis probability will be constant for each gene, which can be used to convert the expected change in synthesis probability from transcription factors to an expected change in expression. Finally, solving with least squares provides the following solution:

$$\hat{r} = (F^T F)^{-1} F^T e \tag{27}$$

During simulations, α_j in Eq. 16 becomes dependent on ppGpp concentrations as shown with the equations below:

$$\alpha_j = \frac{exp_j \cdot loss}{n_{genes}} \tag{28}$$

where exp_j is defined in Eq. 18. $loss$ is the expected loss rate of the given transcript approximated by:

$$loss = \frac{\ln(2)}{\tau} + \frac{\ln(2)}{t_{1/2}} \tag{29}$$

where τ is the expected doubling time from the current concentration of ppGpp (interpolation performed from data from Bremer and Dennis [**?**]) and $t_{1/2}$ is the measured RNA half life. n_{genes} is the expected gene copy number which is a function of τ as determined above and the gene's position in the genome.

```
Algorithm 1: Equilibrium binding
Input : $c_m$ counts of molecules where $m = 1$ to $n_{molecules}$
Input : $S$ matrix describing reaction stoichiometries where $S[i,j]$ describes the
        coefficient for the $i^{th}$ molecule in the $j^{th}$ reaction
Input : $reactants_j$ set of indices for $c_m$ of reactant molecules that participate in
        the $j^{th}$ reaction
Input : $product_j$ index for $c_m$ of the product molecule formed by the $j^{th}$ reaction
Input : $f$ conversion factor to convert molecule counts to concentrations
Input : $K_{d,j}$ dissociation constant where $j = 1$ to $n_{reactions}$
1. Dissociate all complexes in $c$ into constituent molecules to get total reactants
  $(d)$ since some reactants participate in multiple reactions:
    $d = c$
    for each ligand-binding reaction, $j$ do
        for each molecule, $i$ do
            $d_i = d_i + c_{product_j} \cdot S[i,j]$
        $d_{product_j} = 0$
2. Find the number of reactions to perform $(n_j)$ to minimize the distance from
  $K_{d,j}$, where $r$ is a positive integer and not greater than the total products that
  can be formed by the reactants:
    for each ligand-binding reaction, $j$ do
        $n_j = \underset{r}{\operatorname{argmin}} \left| \frac{\prod_{i \in \text{reactants}_j} (f \cdot (d_i - r))^{S[i,j]}}{f \cdot r} - K_{d,j} \right|$
3. Update counts $(c)$ based on number of reactions that will occur. Starting from
  the dissociated counts, reactants will decrease by the number of reactions and
  their stoichiometry and one product will be formed for each reaction.
    $c = d$
    for each ligand-binding reaction, $j$ do
        for each molecule in $reactants_j$, $i$ do
            $c_i = c_i - S[i,j] \cdot n_j$
        $c_{product_j} = n_j$
Result: Ligands are bound to or unbound from their binding partners in a
        fashion that maintains equilibrium.
```

```
Algorithm 2: Two-component systems
  Input : $\Delta t$ length of current time step
  Input : $c_m$ counts of molecules where $m = 1$ to $n_{molecules}$
  Input : $k_A$ rate of phosphorylation of free histidine kinase
  Input : $k_B$ rate of phosphorylation of ligand-bound histidine kinase
  Input : $k_C$ rate of phosphotransfer from phosphorylated free histidine kinase to
          response regulator
  Input : $k_D$ rate of phosphotransfer from phosphorylated ligand-bound histidine
          kinase to response regulator
  Input : $k_E$ rate of dephosphorylation of phosphorylated response regulator
  Input : solveToNextTimeStep() function that solves two-component system
          ordinary differential equations to the next time step and returns the
          change in molecule counts ($\Delta c_m$)
  1. Solve the ordinary differential equations describing phosphotransfer reactions
   to perform reactions to the next time step ($\Delta t$) using $c_m$, $k_A$, $k_B$, $k_C$, $k_D$ and $k_E$.
        $\Delta c_m$ = solveToNextTimeStep($c_m$, $k_A$, $k_B$, $k_C$, $k_D$, $k_E$, $\Delta t$)
  2.  Update molecule counts.
        $c_m = c_m + \Delta c_m$
  Result: Phosphate groups are transferred from histidine kinases to response
          regulators and back in response to counts of ligand stimulants.
```

Algorithm 3: Transcription factor binding

```
Input : $c_a^i$ counts of active transcription factors where $i = 1$ to $n_{\text{transcription factors}}$
Input : $c_i^i$ counts of inactive transcription factors where $i = 1$ to $n_{\text{transcription factors}}$
Input : $P_i$ list of promoter sites for each transcription factor where $i = 1$ to
        $n_{\text{transcription factors}}$
Input : $t_i$ type of transcription factor (either one of two-component,
        one-component, or zero-component) where $i = 1$ to $n_{\text{transcription factors}}$
Input : randomChoice() function that randomly samples elements from an
        array without replacement
for each transcription factor, i do
    if active transcription factors are present then
        1. Compute probability $p$ of binding the target promoter.
        if $t_i$ is zero-component transcription factor then
                transcription factor present → $p_T = 1$
                transcription factor not present → $p_T = 0$
        else
                $p_T = \frac{c_a^i}{c_a^i + c_i^i}$
        2. Distribute transcription factors to gene targets.
            $P_i^{bound}$ = randomChoice(from $P_i$ sample $p_T \cdot len(P_i)$ elements)
        3. Decrement counts of free transcription factors.
Result: Activated transcription factors are bound to their gene targets.
```

Associated data

Parameter	Symbol	Units	Value	Reference
Ligand::TF dissociation constant	$k_d = k_r/k_f$	μM	[2e-15, 5e3]	See GitHub
Free HK phosphorylation rate	k_A	μM/s	[1e-4, 5e2]	See GitHub
Ligand::HK phosphorylation rate	k_B	μM/s	1.7e5	See GitHub
Phosphotransfer rate from free HK-P to TF	k_C	μM/s	1e8	See GitHub
Phosphotransfer rate from ligand::HK-P to TF	k_D	μM/s	1e8	See GitHub
Dephosphorylation rate of TF-P	k_E	μM/s	1e-2	See GitHub
DNA::TF dissociation constant	K_d	pM	[2e-4, 1.1e5]	See GitHub
Promoter sites	n	targets per chromosome	[1, 108]	See GitHub
Fold-change gene expression	FC	$log_2(a.u.)$	[-10.48, 9.73]	See GitHub
ppGpp::RNAP binding Michaelis constant	K_M	μM	24.0	[?], See GitHub
ppGpp::RNAP binding Hill coefficient	h	-	2	Assumed
Fold-change from ppGpp regulation	FC	$log_2(a.u.)$	[-3.11, 3.40]	[?]
Default negative fold-change from ppGpp regulation	FC_-	$log_2(a.u.)$	-0.51	[?], See GitHub
Default positive fold-change from ppGpp regulation	FC_+	$log_2(a.u.)$	1.16	[?]

Table 2: Table of parameters for equilibrium binding, two-component systems, and transcription factor binding Processes. HK: histidine kinase, TF: transcription factor, HK-P: phosphorylated histidine kinase, TF-P: phosphorylated transcription factor. Note in this and future tables we reference the source code for our model, which will be freely available at GitHub as noted in the main text.

Associated files

wcEcoli Path	File	Type
wcEcoli/models/ecoli/processes	equilibrium.py	process
wcEcoli/models/ecoli/processes	tf_binding.py	process
wcEcoli/models/ecoli/processes	two_component_system.py	process
wcEcoli/reconstruction/ecoli/dataclasses/process	equilibrium.py	data
wcEcoli/reconstruction/ecoli/dataclasses/process	transcription_regulation.py	data
wcEcoli/reconstruction/ecoli/dataclasses/process	two_component_system.py	data
wcEcoli/reconstruction/ecoli/flat	equilibriumReactions.tsv	raw data
wcEcoli/reconstruction/ecoli/flat	foldChanges.tsv	raw data
wcEcoli/reconstruction/ecoli/flat	tfIds.tsv	raw data
wcEcoli/reconstruction/ecoli/flat	tfOneComponentBound.tsv	raw data
wcEcoli/reconstruction/ecoli/flat	twoComponentSystems.tsv	raw data
wcEcoli/reconstruction/ecoli/flat	twoComponentSystemTemplates.tsv	raw data
wcEcoli/reconstruction/ecoli/flat	ppgpp_fc.tsv	raw data
wcEcoli/reconstruction/ecoli/flat	ppgpp_regulation.tsv	raw data

Table 3: Table of files for transcription regulation.