Integrating quantitative knowledge into qualitative gene regulatory network: Supplementary data

Jérémie Bourdon, Damien Eveillard & Anne Siegel

June 1, 2011

1 Methods

1.1 ETG modeling overview

ETG modeling approach needs two inputs: (i) a qualitative abstraction of the dynamical system, and (ii) quantitative knowledge. As results, one obtains three kinds of outputs: quantitative simulations of the dynamical system that allows (i) a validation of our results based on experimental knowledge at disposal and (ii) prediction of biological compound behaviors over times. Furthermore, sensitivity analysis emphasizes (iii) a ranking of the most important ETG transitions that must occur to satisfied the overall quantitative behavior of the system. Figure 1 pictures the modeling process overview. A MATLAB script allows to perform the ETG modeling. It can be found here[1].

1.2 Qualitative inputs of the ETG modeling

1. *Defining the ETG graph (core model):* The graph describes the qualitative behaviors of the biological regulatory network by focusing on the products of the genes. As described in the manuscript, a given gene stochastically regulated by the system either product a increase of (gene_+), or a decrease (gene_-) of its protein quantity, which characterizes what we call here two events that are related to a given gene. This graph can be manually built using biological knowledge at disposal (i.e., knowing what gene activation actives or represses what gene activations). When available, one should consider to make this graph automatically from qualitative models. For illustration, the following regulatory model of two genes formalized into a Piecewise

[1] http://pogg.genouest.org

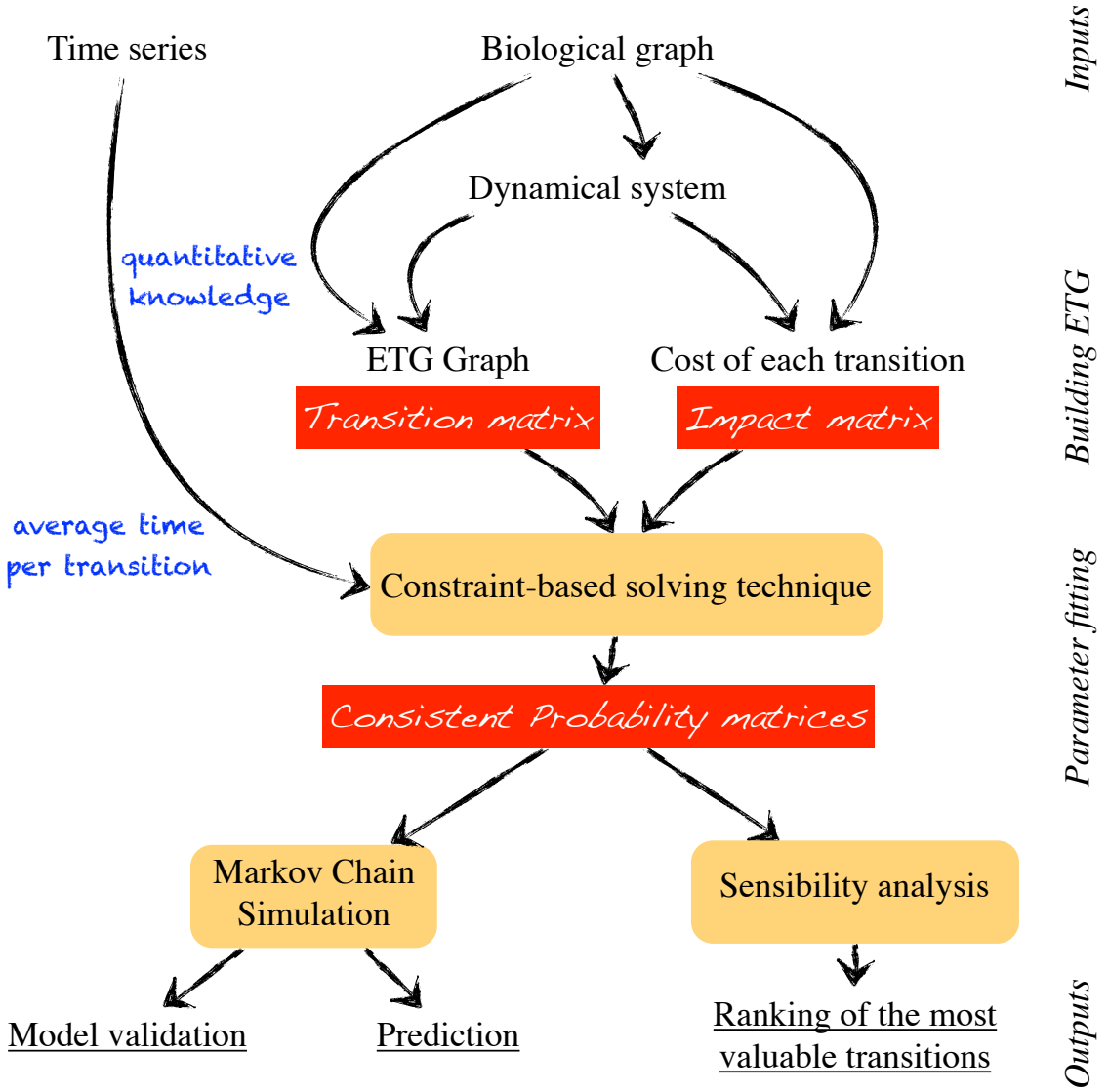
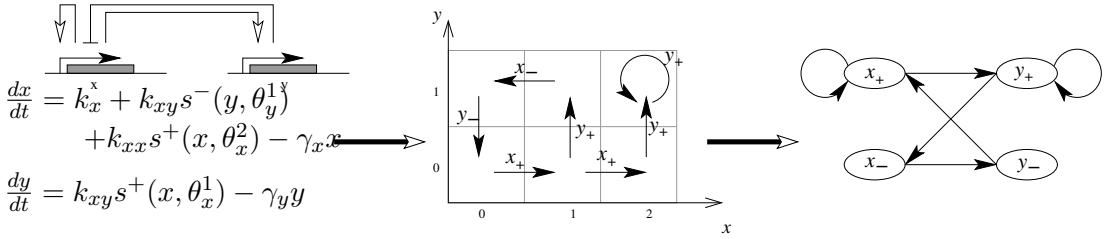


Figure 1: Flowchart of the ETG modeling process. For two given inputs, quantitative knowledge of time series and qualitative biological knowledge, the ETG modeling technique provides a simulation of the model and an automatic sensitivity analysis. Those results can be further analyzed like for validation purposes.

Affine Differential Equations (PADE) system can be transformed into ETG graph in two steps:



a) Two genes interaction network and its PADE definition

b) Discrete dynamics or state transition graph

c) Concentration graph

The transformation from PADE systems to ETG structures is straightforward. First, one produces a state transition graph from a PADE. It is classically obtained by fixing some inequality constraints on the parameter of the system's equations (see GNA webpage for details). Second, the ETG is made from the state transition graph by considering all possible successions of events, which is here mRNA changes. Precisely, there exists an edge from state A to state B in the ETG graph if the two successive transition paths $\xrightarrow{A} \cdot \xrightarrow{B}$ is present in the state transition graph.

This example is available in MATLAB. The ETG graph is described as a MATLAB matrix in `ExampleSqueleton` that defines the corresponding $\{0,1\}$-transition matrix.

This simple example has two latent variables $v_1 = p_{x_+ \to x_+}$ and $v_2 = p_{y_+ \to y_+}$ allowing to express the unknown probability transition matrix of the model

$$\begin{pmatrix} v_1 & 1-v_1 & 0 & 0 \\ 0 & v_2 & 1-v_2 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

2. *Estimation of the impact of each transition:* we consider a cost for taking each transition of the ETG graph. The cost is a direct impact on the protein production. As a biological assumption, we assume the passive degradation rate (free parameter) to be equal to 5% for both protein X and protein Y. The value of the active production and degradation rates d_+ and d_- for protein X satisfies an equilibrium principle saying that for a uniform choice of transition probabilities, protein X expected concentration is constant. Assuming that $d_- = p$ and $d^+ = 1/p$ leads to the resolution of an equation of order 2 in p where the coefficients depend on the stationary distribution π of the transition matrix. Here, the equation is

$$1/p\ \pi_{x^+} + p\ \pi_{x^-} + 0.95 \times [1 - \pi_{x^+} - \pi_{x^-}] = 1.$$

This equation have only one solution smaller than 1, $p = 0.8818$. Thus $d_+ = 1.1341$

and $d_{-} = 0.8818$. The impact matrix for protein X is then

$$\begin{pmatrix} 1.1341 & 0.95 & 0 & 0 \\ 0 & 0.95 & 0.8818 & 0 \\ 0 & 0 & 0 & 0.95 \\ 1.1341 & 0 & 0 & 0 \end{pmatrix}$$

This solution is described in MATLAB format in the matrix `CostProtein_X` that describes the cost of ETG graph transitions on the protein X quantities (mainly, the transitions involved and the cost for these transitions).

1.3 Probability inference

To demonstrate the interest of our modeling approach, we consider fictive experimental data. The dataset is hence defined by the protein X concentration multiplied by 100 in 100 time unit or iterations. Assuming a multiplicative effect of the regulatory network on the protein concentrations, it corresponds to an asymptotic growth rate (observable variable of the system) of:

$$\exp(\log(100/1)/100) = 1.0471$$

One must find probabilities that allow to obtain such a growth rate. The inference of these probabilities is performed by our MATLAB script using:

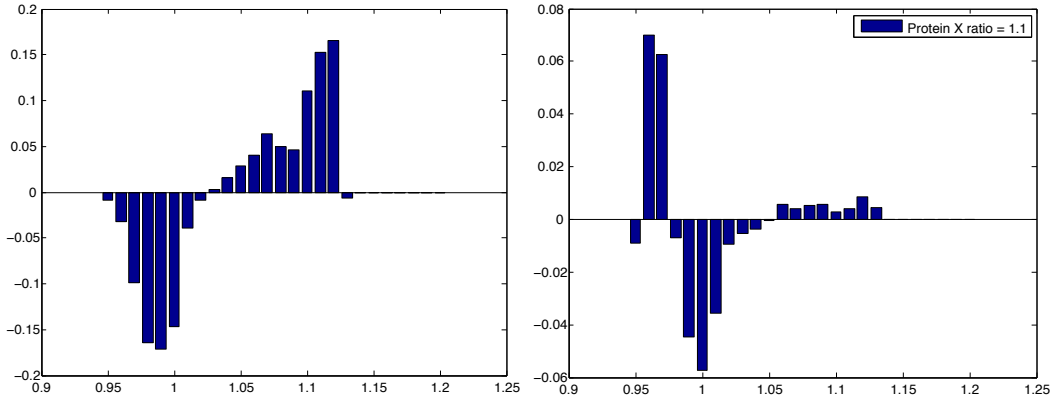
```
[BS,BM]=ETG_solve(ExempleSqueleton, {CostProtein_X}, {1.0471}, confidence);
```

where `ExempleSqueleton` is the $\{0,1\}$-transition matrix and `CostProtein_X` is a structure describing the cost of protein X (mainly, the transitions involved and the cost for these transitions), `confidence` if the maximal allowed error in the numerical computations. `BM` is the matrix of probabilities that satisfy the protein growth rate, whereas `BS` is the euclidean distance between the estimated growth rate, as computed using `BM`, and the optimal growth rate as given by the experiments.

1.4 Plasticity of Event Transition Markov chain face to extreme parameters

In complementary to the results shown in the manuscript, one compares the distributions of protein Y growth rates under various conditions. All these conditions are compared to the distribution of a random case where both variables v_1 and v_2 are drawn uniformly. Then, one considers two extreme conditions for the X growth rate. For instance, this parameters may be equals to 0.97 for simulating a fast degradation of X protein, or 1.1 for simulating a major burst of protein X production. In both cases, distributions of the Y growth rate

have been estimated by considering 10000 inferred probability matrices. One are then able to compute their difference of estimated Y growth rate distributions between the "random" condition and X growth rates in both extreme conditions.



Random condition V.S. 0.97 X growth rate Random condition V.S. 1.1 X growth rate

The difference in the distributions highlights the inherent plasticity of the Event Transition Markov chain to estimate protein growth rates in various conditions, despite the major constraints given by the Event Transition Graph that restricts the quantitative behaviors.

2 Application to the gene regulatory network of *Escherichia coli* under carbon starvation

2.1 Experimental data and model

As an application of our modeling approach, we consider the time series data extracted from:

Ball CA, Osuna R, Ferguson KC, Johnson RC (1992) Dramatic changes in Fis levels upon nutrient upshift in *Escherichia coli* *J Bacteriol* **174**: 8043–56.

and represented in Fig. 2. For sake of simplification, concentrations of both proteins are not represented by their absolute concentrations but rather converted into their respective percentages of maximal observed concentration.

We used the biological model and corresponding biological assumptions as published in:

times (min)	[FIS] %	[CYA] %	growth phase
0	75	75	stationary
2	10	1	stationary
30	20	1	stationary
55	50	1	stationary
70	80	1	stationary
80	100	1	exponential
100	50	100	exponential
110	30	-	exponential
130	10	130	exponential

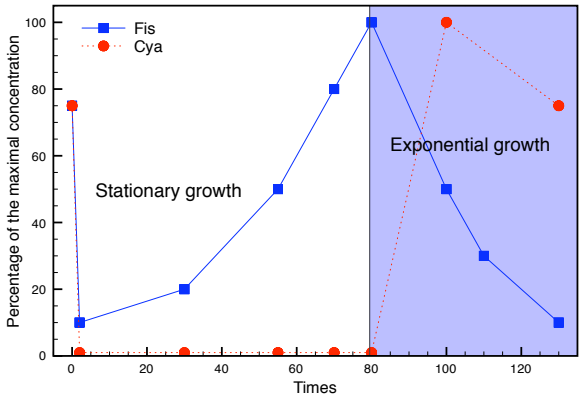


Figure 2: Experimental data at disposal

Ropers, D., de Jong, H. D., Page, M., Schneider, D., & Geiselmann, J. (2006). Qualitative simulation of the carbon starvation response in *Escherichia coli*. *BioSystems*, **84**(2), 124-152.

Since the model is formalized in a Piecewise Affine Differential Equation system (PADE), both its biological graph (ex: gene x activates the gene y transcription) or its formalization of the dynamical system can be used to build an ETG. As an application, we used herein the biological graph. The corresponding ETG is pictured in Figure 6 of the manuscript. Notice that the switch between the two phases impacts the event transition graph by suppressing two transitions ($fis_+ \rightarrow crp_-$ and $complex \rightarrow fis_-$ in the stationary growth phase.

2.2 Training dataset and costs

For the sake of clarity, we expose here the data used for the training the model (i.e. estimation of the probability matrices). Notice here that 3 data points are needed for finding the information.

Protein costs. Following the equilibrium principle, one deduces these values for the relative protein rates.

Protein	d_+	d_-
FIS	1.4653	0.6825
CYA	2.0636	0.4846
CRP	1.4362	0.6963
TOPA	1.6181	0.6180
GYRAB	1.8662	0.5358

times (min)	[FIS] %	[CYA] %	growth phase
2	10	-	stationary
80	100	-	exponential
130	10	-	exponential

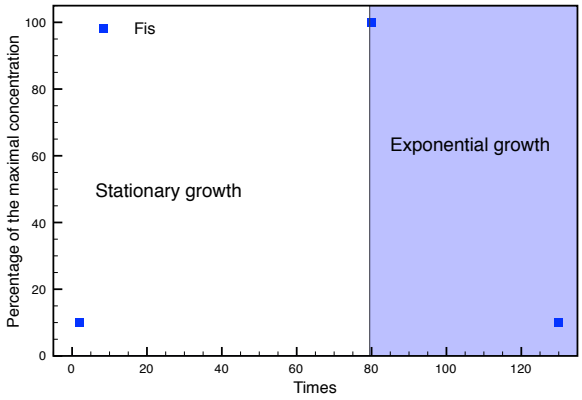


Figure 3: Experimental data used for the parameter identification

Datasets. The concentration evolution rates can be determined for both phases, according to Figure 2.2. For instance, the growing rate for FIS in the stationary growth phase, computed by using is relative values at times 2 and 80 minutes, equals

$$\frac{\log(100) - \log(10)}{80 - 2} = 1.03.$$

This value says that FIS protein concentration increases by 3% each minute. In order to use it in our model, it is necessary to obtain the corresponding rate per transition of the model, and thus to know the number of iterations performed by the model in a one minute duration. We argue that FIS, CYA and other proteins are degraded as soon as a sufficient number of its amino acids are degraded. In accordance to the N-end rule [Alexander, Varshavsky (1997). "The N-end rule pathway of protein degradation". Genes to Cells 2 (1): 13-28], we take a duration of 2 minutes as the minimal half-life for these animo-acids. Thus, when taking a natural degradation rates of 5% per transition, the model runs n iterations to degrade half of the present proteins, where n satisfies $0.95^n = 0.5$. Here, $n \approx 14$ implying that 7 iterations are reached per minute. Known concentration evolution rates in both phases, expressed in a per iteration scale, are synthesized in the following table:

Protein	Stationary growth	exponential growth
FIS	1.0042[1]	0.9934[1]
CYA	0.7346[2]	1.0334[2]
CRP	?	?
TOPA	?	?
GYRAB	?	?

[1] used for inference
[2] used for validation

Mean values of the predicted concentration evolution rates in both phases (mean computed using 100 points of the solution set):

Protein	Stationary growth	exponential growth
FIS	1.0042	0.9934
CYA	0.9506	1.0745
CRP	0.9506	1.0073
TOPA	0.9819	0.9987
GYRAB	0.9700	0.9675

Figure 4 depicts an example of probabilities assignment that satisfies the expected growth ratios for protein FIS.

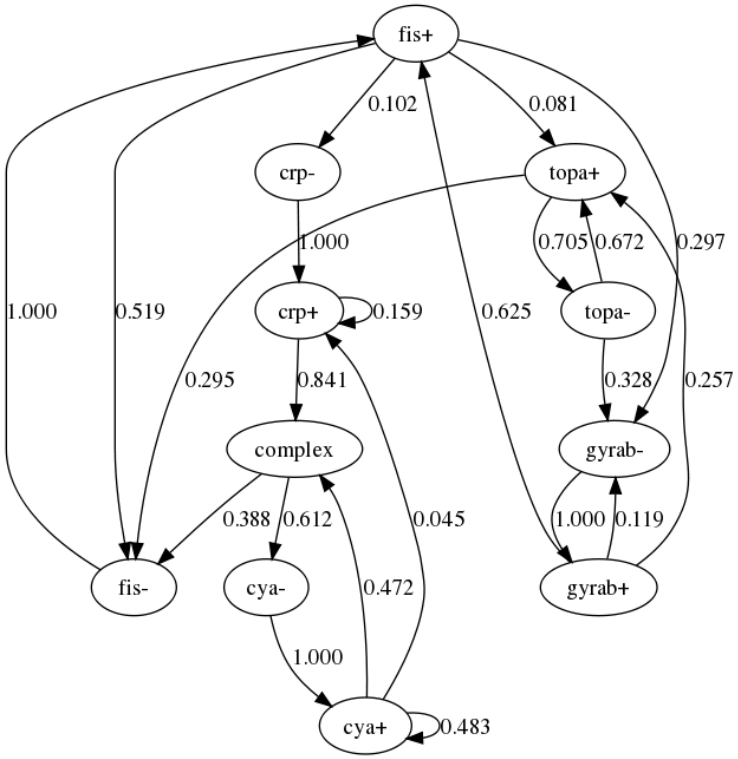
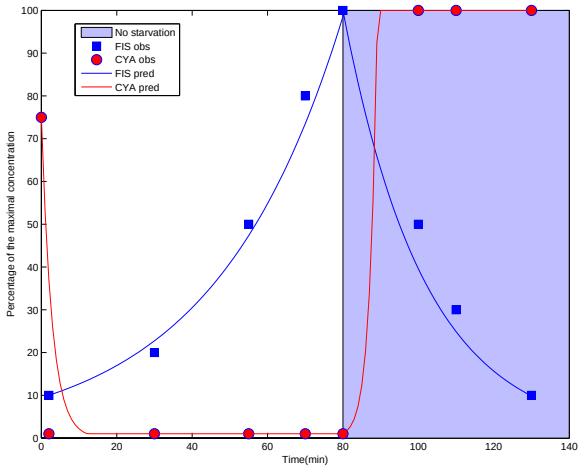


Figure 4: Event Transition graph and **an** example of corresponding probabilities after an estimation based on experimental data such as given in Figure 3. Note herein that several probabilities allow to fit the experimental knowledge.

Model validation For the sake of validation of our modeling technique applied on E. coli, we compare the time series predicted with those observed experimentally during both growth phases. As previously mentioned, CYA and Fis concentration behaviors were investigated. A comparison between FIS and CYA observations and their respective predictions by the model is performed. A Pearson correlation test confirms the accuracy of the predictions. Notice that in the computed time-series, we set the value to 1 if the computed value is smaller than 1 and to 100 is the computed value is greater than 100.

Time	FIS		CYA	
	obs	pred	obs	pred
0	-	-	75	75
2	10	10	1	37
30	20	23	1	1
55	50	47	1	1
70	80	73	1	1
80	100	99	1	1
100	50	40	100	100
110	30	25	100	100
130	10	10	75	100
R^2	0.9937		0.9599	
p-value	$6.5\ 10^{-8}$		10^{-5}	



2.3 Model predictions

It is also possible to predict other concentration evolutions, assuming, for instance, an initial concentration of 50% for each unknown proteins. The resulting predictions are depicted in Figure 5.

2.4 Sensitivity of the ETG transitions

Computing the sensitivity of the model allows to rank the transitions according their partial derivative. The higher is the sensitivity of the transition, the higher it is constrained to be equal to a fixed value. It is expressed in percentage having the following meaning: if the given probability is changed by 1%, then the euclidean distance between the expected growth ratio and their predictions is modified by X% (the given sensitivity). Each returned sensitivity is computed as the mean over 100 transition matrices satisfying FIS observed protein evolution. Such an information is useful to classify the transitions according to their importance on the system. The sensitivities are depicted in Figure 6.

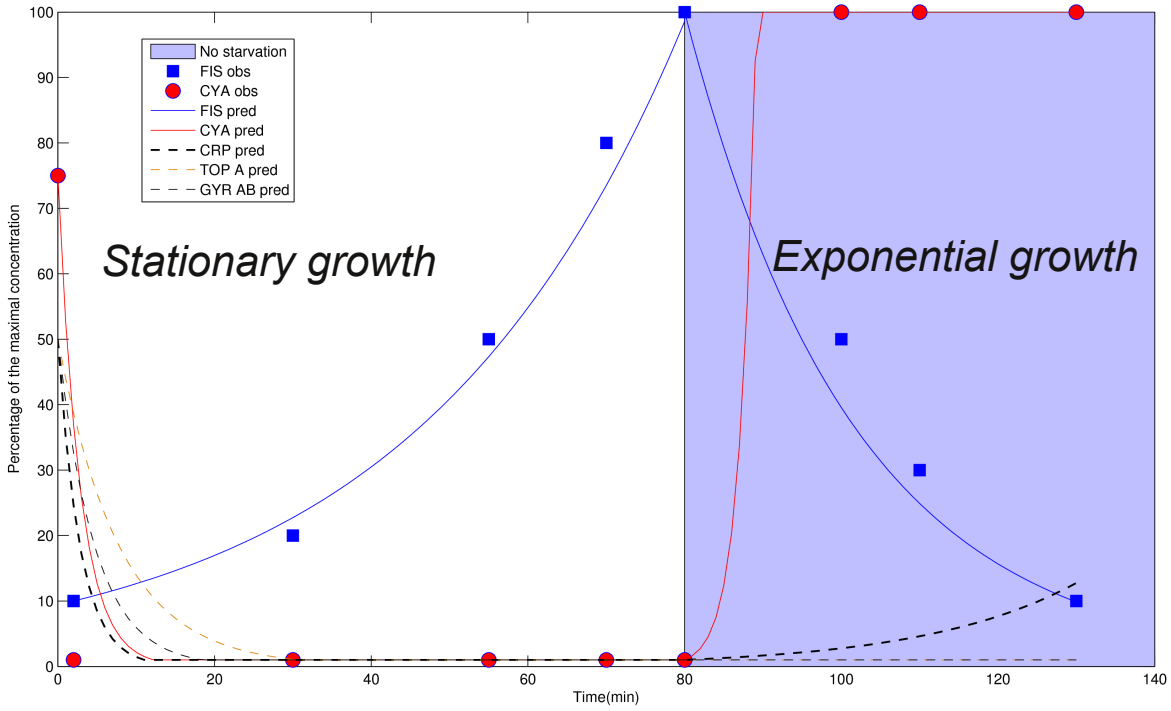


Figure 5: Predictions of diverse protein concentrations related to the studied system.

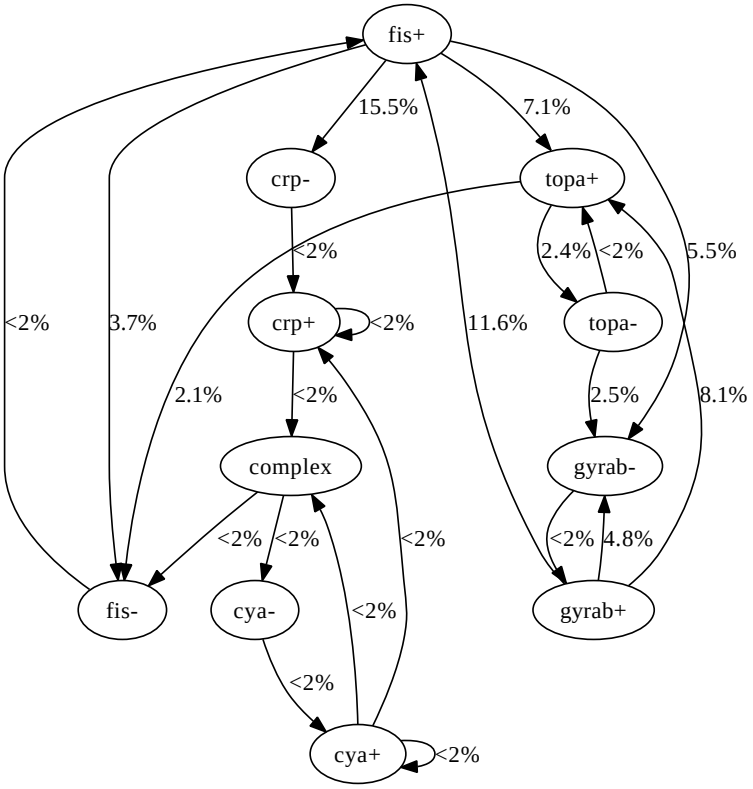


Figure 6: Event Transition graph and corresponding sensitivities after an estimation based on experimental data such as given in Figure 3