uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Vida Abedi

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Cellular and Molecular Medicine)

GRADE / DEGREE

Department of Cellular and Molecular Medicine

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

System-Identification of Gene Regulatory Networks by Systematic Gene Perturbation Analysis

TITRE DE LA THÈSE / TITLE OF THESIS

Dr. Mads Kaern

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Dr. D. Bickel

Dr. C. Scott Findlay

Dr. T. Perkins

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

System-Identification of Gene Regulatory Networks by Systematic Gene Perturbation Analysis

By

Vida Abedi

This thesis is submitted as a partial fulfillment of the Masters in Science program in Cellular and Molecular Medicine

Faculty of Medicine, University of Ottawa

Your file Votre référence
ISBN: 978-0-494-59888-7
Our file Notre référence
ISBN: 978-0-494-59888-7

ABSTRACT

Systems biology is an interdisciplinary field that combines engineering and molecular biology to better understand the 'design principles' of biological systems. One of the main goals in systems biology is to understand and map complex biological networks. In order to achieve this goal, tools able to process the non-linearity and high dimensionality of biological systems are urgently needed. To develop, test and benchmark tools for investigation of biological processes, it is important to utilize a model system that is able to provide a glimpse of complexity found in higher organisms, and be simple enough such that detailed studies can be performed rapidly, accurately, and reproducibly. Here, we have developed a system-identification framework for the extraction of quantitative and mechanistic information about causal relationship among genes using the canonical galactose utilization pathway in *Saccharomyces cerevisiae*. This framework, referred to as systematic gene perturbation analysis (SGPA), is based on the effects of systematic pair-wise gene deletions. In essence, the method establishes dynamical models of the regulatory network from single-cell measurements of steady-state input-output relationships, in systematically perturbed networks. SGPA framework is successful in identifying the network structure for the GAL system. This strategy leads the way to a better application of available resources and provides a scalable framework for system-identification and reverse-engineering of biological networks based on in vivo systematic data generation.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
GA	Genetic Algorithm
GAL	Galactose
GAL2	Galactose permease
GAL3	Transcriptional regulator
GAL4	DNA-binding transcription
GAL5	Phosphoglucomutase
GAL6	Leucine AminoPeptidases
GAL7	Galactose-1-phosphate uridyl transferase
GAL80	Transcriptional regulator
GFP	Green Fluorescent Protein
GRN	Gene Regulatory Network
I/O	Input/Output
MIG1	Repressor of GAL genes when glucose is present
OFF state	State in which cells do not express GFP. Since yeast cells auto-fluoresce naturally, therefore cells with low expression level are assumed to be in the OFF state. OFF state translates to a situation where the galactose utilization pathway is not active.
OD	Optical Density
ODE	Ordinary Differential Equation
ON state	State in which cells express GFP. ON state translates to a situation in

	which the galactose utilization pathway is active.
P_{GAL1}-GFP	Promoter of GAL1 fused to GFP. GAL1 is a Galactokinase which phosphorylates alpha-D-galactose in the first step of GAL network
RE	Reverse-Engineering
RNAi	RNA interference
SGPA	Systematic Gene Perturbation Analysis
SI	System-Identification
SM	Synthetic Media
YPG	yeast/peptone/galactose (rich media containing galactose as the carbon source)
YPR	yeast/peptone/raffinose (rich media containing raffinose as the carbon source)

Acknowledgements

A special thank you to my supervisor, Dr. Mads Kaern, for his encouragements, motivation and mentorship throughout the course of my studies.

Many thanks to Dr. Beatz, Dr. Andrade, Dr. Hincke as well as Dr. Perkins for their time, support and sound advices.

Many thanks are extended to Ms. Donna Hooper, Blanche Dinelle and Sylvie Deblois for their support and administrative help. Many thanks also to Dr. Renauld for his support and commitment to graduate students.

I would also like to thank my colleagues, especially Simon St-Pierre, Mila Tepliakova, Hilary Phenix and Dawn Fraser for their continuous support and help without whom this project would not have been a success.

This work was supported by a grant from CIHR and by MITACS (Mathematics of Information Technology and Complex Systems).

Most of all, I would like to thank my family members, especially my husband, Ramin, my parents Farah and Majeed and, my sister and brother, Dona and Keyvan. Your continuous support and unconditional love have given me the strength to continue and achieve my goals that I have set for myself.

Chapter 1: Introduction

A key goal of systems biology is to develop improved methods to map complex biological regulatory systems providing insight into how biomolecular interaction networks propagate and process information. Systems Biology studies typically aims to generate new knowledge by integrating experimental inquiry, computational data analysis and model-based hypothesis generation. While methods developed in engineering could in principle be utilized to analyze biological data, the high dimensionality and often qualitative nature of experimental datasets, as well as the predominance of non-linear co-dependencies among biomolecules, make the task of adapting these methods to a biological context very challenging (Sontag E.D., 2004). The development of novel approaches for systems-level characterization thus requires not only new computational methodologies to analyze complex datasets, but also the design of new experimental assays that can provide the appropriate types of data. This thesis describes the development of a novel integrative approach, referred to as systematic gene perturbation analysis (SGPA), and demonstrates its usefulness as a method for reverse engineering transcriptional regulatory networks.

The SGPA approach is a data-driven method designed to extract quantitative and mechanistic information about causal relationships among regulatory genes in signal transduction and processing systems without *a priori* knowledge. It is inspired by a

classical engineering approach to gain insight into the properties of a "controller" by measurements of input/output relationships under different operational conditions, or perturbations. In this first implementation of the SGPA, the input signal is the extra-cellular concentration of an inducer, the output signal is the steady-state expression of a Green Fluorescent Protein (GFP) fused to an induced promoter, and the perturbations consists of systematic knockout of individual regulatory genes and combinatorial knockdown of regulatory gene pairs.

As with all technology platform developments it is essential to choose an appropriate model system that can be used for benchmarking and validation. We have chosen the galactose utilization pathway in budding yeast *Saccharomyces cerevisiae* as our model system as it satisfies a number of important criteria. As one of the most primitive eukaryotic organisms, budding yeast maintains some of the complexity found also in higher organisms (including conservation of numerous pathways) – yet detailed studies can be performed rapidly, accurately, and reproducibly. For these reasons, budding yeast has become a "gold standard" for systematic investigation of biological processes (see review by Barnett J.A., 2007) and for tool development (Ito T. *et al.*, 2001; Uetz P. *et al.*, 2000; Gavin A.C. *et al.*, 2002; Ho Y. *et al.*, 2002; Aloy and Russel, 2004; Harbison C.T. *et al.*, 2004; Lee T.I. *et al.*, 2002). Among the numerous canonical pathways found in budding yeast, the galactose utilization pathway is one of the best-studied transcriptional regulatory networks in eukaryotes (see review by Gancedo, J. M.,

1998), and has for over 60 years served as a favourite model system in molecular biology. Since the interactions among regulatory genes in pathway are well established, it is ideally suited for the development of new methodologies to map complex regulatory pathways. As this thesis will demonstrate, our SGPA methodology succeeds in identifying, with no prior knowledge, the main regulatory interactions among genes in the galactose utilization pathway setting the stage for the application of the method to other networks in yeast as well as higher organisms.

Goal and Objectives:

The goal of this thesis is the development of a novel system-identification method for the extraction of quantitative and mechanistic information about causal relationship among regulatory genes from steady-state network input-out relationships, in systematically perturbed networks. Our approach to achieving this goal can be described as a sequence of four objectives summarized briefly below.

Objective I:

Generating high quality quantitative data for benchmarking system identification method. This objective encompasses three separate steps: 1) Generating a library of deletion strains carrying systematic gene perturbations; 2) Identifying key regulatory genes; and 3) Generating high quality single-cell expression data

Objective II:

Developing mathematical model based on input/output relationship

Objective III:

Developing an optimization algorithm for fitting the model to experimental data

Objective IV:

Validating the methodology by comparing the results to the known network structure

Background

Uncovering Biological Networks

Systems biology, an emerging field in the post-genomic era, is the science of discovering, modeling and engineering biological processes. System biology is by definition multidisciplinary: it combines engineering and molecular biology to better understand the 'design principles' of biological systems. It is the idea of using a systems approach to address biological questions which can only be done if biologists work in parallel with theoreticians. Systems biology is about the interplay between experimental science and simulation studies. Complex systems are translated into mathematical models for generating new hypothesis toward investigation of key biological questions. This approach was used, for example, to understand the flow of information in signalling pathways (Mettetal J.T. *et al.,* 2008), to study the noise in gene expression (Ramsey S. *et al.,* 2006) or even to study cellular memory (Acar, Becskei and van Oudenaarden 2005).

Systems biology is also the study of inferring functional relationships among genes (referred to as genetic interactions) and building regulatory networks based on gene perturbation experiments (Mani R. *et al.,* 2008). For example, synthetic genetic array (SGA) analysis, which is based on crossing a query mutation to an array of deletion mutants and visually monitoring growth rate of double mutant haploid strains, has been successful in grouping genes by function (Tong A.H.Y. *et al.,* 2001). SGA is a high-

throughput method that allows the identification of epistatic interactions among genes based on extreme pheontypes; severe sickness or death. E-MAPs, or epistatic microarray profiling, is a technology that overcomes some of the limitations of SGA by providing a quantitative approach for measurements of genetic interactions (Collins S.R. *et al.,* 2006). This method allows the detection of epistatis based on less severe phenotypes but only allows a subset of the genome at a time. E-MAP experiments consist of measuring growth rate of yeast strains carrying each pair of mutations. The quantitative measure of growth rate of the double versus single deletion strains determines aggravating (negative) or alleviating (positive) interactions. SGA as well as E-MAP analysis are but two examples of great tools for identification of functionally related genes; yet, they do not provide quantitative data needed to develop dynamical models.

Several approaches allowing the development of dynamical models have been developed in the past, some of which require data from perturbation experiments (Bansal M. *et al.*, 2007). For example, Gardner T.S. *et al.,* (2003), used a systematic transcriptional perturbation strategy (over-expression) to construct a first-order model of regulatory interactions in a nine-gene sub-network of the SOS pathway in *Escherichia coli.* The predictive model, which used steady-state expression measurements with no prior knowledge, was based on the assumption of linear interactions among the genes in the network. To identify the architecture of the network, the method requires N distinct perturbation experiments (where N is the number of elements in the system) as well as

accurate measurements of RNA concentrations for all genes of interest. The computational methodology, referred to as network inference by multiple regression (NIR), further required that that number of possible interactions per gene is low (e.g., the network needs to be sparse), and this number needs to be determined in advance of the computational analysis. The method requires therefore a large amount of experimental data even for a small regulatory network. Additionally, since the inferred dynamical model is linear, it is more suitable for association studies, discovering compound mode of action, for example, than for generating insight into biological information propagation and processing.

System Identification and Reverse-Engineering

An alternative strategy to network inference is to use system identification methods based on an input/output relationship, where the process of interest is modelled by a black-box. This type of modeling is inspired from engineering, where the black box is referred to as the 'controller' or the 'plant' of the system of interest. The amount of required information is reduced – measurement from the reporter gene (output) for each perturbation experiment; however, the quality of data and the design of perturbation experiments are critical for a successful study.

Reverse engineering (RE) and system identification (SI) are highly active fields of Systems Biology research. The goal is to determine the internal structure/behaviour of the system based on the input/output relationships, which in engineering is called transfer functions. The idea is very natural when one considers cell as a system and cellular processes such as cell growth, cell death and metabolism as sub-systems. Input signals could be physical as well as chemical: UV radiation, drug treatment, temperature fluctuations, nutrient limitations are examples of possible input signals. The measurable output signal could be change in genes expression, activation of transcription factors, cellular proliferation rates (i.e., fitness) or change in cell morphology and so on.

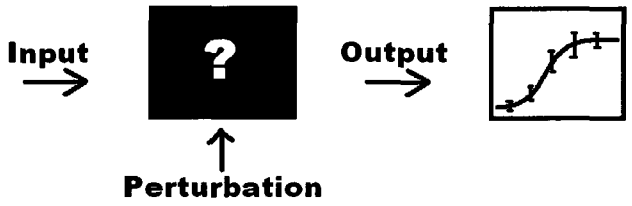


Figure 1: System Identification and Reverse-Engineering based on Input/Output (I/O) relationship.

A system (i.e. cell, cellular process or a signalling pathway) can be viewed as a black box that is subject to perturbations. The input-output relationship can then be utilized to investigate dynamical properties and potentially internal structure of the system represented by the black box. These types of problems in systems biology are very similar in nature to standard problems in control theory; however, there are challenging differences (Sontag E.D., 2004). In fact, evolution has resulted in systems that are highly non-linear, fault-tolerant and very rich in feedback loops therefore, existing techniques from SI and RE in control theory are not sufficient to unravel mechanisms underlying complex biological systems (Sontag E.D., 2004). There is strong need for powerful mathematical tools to facilitate processing of biological information.

A good example of system identification approach to gain insight into dynamics of signalling cascades is the work done by Mettetal J.T. and colleagues (2008). They used SI method based on I/O relationship to investigate the frequency dependence of osmo-adaptation in *Saccharomyces cerevisiae.* The input (extracellular osmolyte concentration) and the output (activity of the MAPK Hog1) of the network were easily identified and by comparing the response of the wild-type and different mutants, it was possible to infer a

concise predictive model. This work supports the idea that the osmo-adaptation response is dominated by a fast-acting negative feedback through kinase Hog1.

Recent computational search algorithms have been applied to shed light on, for instance, mechanisms underlying force integration during mitosis in the *Drosophila* embryo (Wollman R. *et al.,* 2008). The approach, which was based on quantitative data, eliminated thousands of possible models and identified six distinct strategies for microtubules-motor integration. Identified differences among the six predicted strategies suggest future experiments. In fact a challenging task in RE is to identify the best possible model structure if multiple models are suited to describe the data; the latter would potentially require new experimental data for model discrimination (Kremling A. 2004). Another key challenge in RE is validation of methods used, which is mainly due to the limited availability of high quality quantitative data (Gardner and Faith, 2005). In fact, lack of high quality data and the non-linearity of the available datasets make benchmarking RE methods difficult. In many cases a model allows an infinite number of parameter combinations, producing all an equally good agreement with the data (Isidori, A. 1995). Model reduction techniques such as sensitivity analysis-based techniques are usually used (Schmidt H. *et al.*, 2008) to overcome this problem. Sensitivity analysis-based relies on eliminating parts of the model that are locally unimportant for the property of interest (Brown, N.E. 1997; Danø S. *et al.,* 2006; Edelson D. 1981; Gautier O. et al., 1985; Okino and Mavrovouniotis 1998). Generally speaking, RE algorithms

seem to better perform when 'local perturbation experiments', such as single gene over-expression or knockdown, or when steady-state data are used. Time-series data and global perturbation experiments are usually less informative for the same amount of data (Bansal M. *et al.*, 2007).

In addition to that, single cell data, when available, can provide essential information about the dynamic of the network. Bistability, existence of two stable steady-states for the same external condition, is a consequence of non-linearity in the system and can only be detected if analyzing single cell datasets. Bistable systems can be found only in the presence of an appropriate genetic background (Thomas R. 1998). It has been shown in synthetic networks that positive feedback mechanism is able to convert a graded response to a binary response (Becskei A. *et al.*, 2001); however, this characteristic is not a direct consequence of positive feedback since it depends on the parameter values that characterize the gene circuitry (Becskei A. *et al.*, 2001). The positive feedback loops in biological systems have been studied extensively because of their intriguing properties (see review by Mitrophanov and Groisman, 2008; Brandman O., *et al.*, 2005; Sneppen K. *et al.*, 2008). Nonetheless, the design principles underlying the true dynamical behaviour of most biological networks are still unclear. However, the quality and resolution of single cell data-set provides valuable information and contributes significantly toward a better understanding of the structure, dynamics and functions of regulatory networks.

Biological Robustness

Robustness of biological systems is said to occur when changing system parameters yield minimal variation in the systems functional output (Kitano H. 2004; Stelling J. *et al.* 2004; Csete and Doyle 2002). It has been proposed that redundancy, modularity and increased network complexity are some of the means by which biological systems use to achieve robustness (Wagner A. 2005). However, robustness due to network structure is maintained over subsequent evolution, making structure-based robustness more stable than redundancy-based robustness (Salathe M. *et al.*, 2008). This not-so-intuitive feature of biological systems arises from the fact that for the systems to be functionally stable, the biological network has to maintain specific dynamical behaviour (i.e. oscillatory behaviour): purely redundant proteins could be lost if not needed, without compromising the dynamical aspects of the system (Salathe M. *et al.*, 2008; Deutscher D. *et al.*, 2006).

The 'robust-yet-fragile' feature of many biological networks is an integral systems property and causes robust response to certain perturbations and fragile response to other perturbations (Csete and Doyle 2002). Regulatory circuitries are important features of a system because of their role in maintaining the homeostasis or bistability (Bhalla and Iyengar 2001; Ferrell J. 2002) making a link between evolutionary robustness and system's dynamics. It has also been shown, through an exhaustive computational analysis, that the requirement for functional robustness drastically reduces the choices of viable topology, which can then be reduced even further through perturbation

experiments (Ma WZ, *et al.,* 2006); in fact the set of viable topology that can reproduce the wild-type behaviour and also the mutant behaviour in a robust manner are surprisingly limited (Ma WZ, *et al.,* 2006), opening a door for system identification and reverse engineering when experimental data are rich and accurate and incorporate a wide range of perturbation experiments.

Yeast: Gold Standard Model System

Yeast is a Model System in Systems Biology. To unravel the network of life, comprehend the true underlying principles of biological robustness and develop concepts for quantitative description of living cells, it is vital to use a model system which can provide a glimpse of complexity found in higher organisms and, be at the same time simple enough such that detailed studies can be performed rapidly, accurately and reproducibly. A gold standard model system in the field of systems biology is the budding yeast, *Saccharomyces cerevisiae*. Yeast provides a good balance of complexity. It has about 6000 genes, 50% of which have homologue in humans and 30% are involved in diseases (Foury F. 1997). In fact most fundamental cellular processes are conserved from yeast to humans (Marton M.J. *et al.,* 1998). Furthermore, yeast was one the first eukaryotes to have its full genome sequenced. Yeast is a great candidate for its interesting features: it has one of the smallest genome sizes among well-studied eukaryotes (over 14,000,000 bp), and a high gene density with >70% of the genome encoding for proteins (Oliver, S. G. *et al.,* 1992). In 1997, yeast's full genome was sequenced, opening the door of many possibilities (Bussey, H. *et al.,* 1997; Churcher, C. *et al.,* 1997; Dietrich, F. S. *et al.,* 1997; Dujon, B. *et al.,* 1997; Jacq, C. *et al.,* 1997; Johnston, M. *et al.,* 1997; Philippsen, P. *et al.,* 1997; Tettelin, H. *et al.*, 1997). Since that significant step, many high-throughput experiments have been performed and, a 'bar-coded' complete set of deletion mutants (Giaever G. *et al.,* 2002) and sets of open reading frames tagged for

proteomics analysis (Gavin A.C. *et al.,* 2002, Ghaemmaghami S., *et al.,* 2003; Huh W.K. *et al.,* 2003) were made available in the yeast community. Thousands of microarray transcription experiments in multiple conditions (Lum P.K. *et al.,* 2004; Hillenmeyer M.E. *et al.,* 2008), fitness analysis of knockout strains (see a review by Boone, Bussey and Andrews 2007) and combinatorial gene deletion study (St. Onge R.P. *et al.,* 2007) are just a few of genome wide screens performed during the last five years. All this has motivated computational scientists and theoreticians to model regulatory interactions and deduce functional relationship among genes in order to better understand its dynamics and architectural structure. This exponential growth in availability of information and feasibility of performing genome wide screens rapidly to validate new hypothesis generated from models has created an increased interest in yeast as a model organism in systems biology. In fact systems biology is a marriage of experimental biology and mathematics and uses a systems approach to solve the complex biological questions.

Examples of a systems biology approach using modeling and computational tools to study pathways, genetic networks and transcription regulation in yeast includes, among others, the galactose utilization pathway (Ideker T. *et al.,* 2001), yeast protein interaction networks (Ito T. *et al.,* 2001; Uetz P. *et al.,* 2000; Gavin A.C., *et al.,* 2002; Ho Y. *et al.,* 2002; Aloy and Russel 2004), or yeast transcription networks (Harbison C.T. *et al.,* 2004; Lee T.I. *et al.,* 2002). The galactose utilization pathway, which will be described later in this chapter, is, however, favoured for methodology development since it is one of the

classical pathways and been studied for more than 60 years (Reviewed by Barnett J.A. 2007; Winge & Roberts, 1948).

Finally, yeast has been used extensively to pioneer the concepts and tools to tackle complex biological questions. The outlines of its genetic networks are apparent but a more complete exploration would be a substantial resource for better understanding of more complex systems. The true value is also in the development of methods and implementation of tools in biotechnology and mathematics: the introduction of microarray technology, single cell flow Cytometry and high resolution microscopy along with mathematical tools such as sensitivity analysis, continuation analysis, and standard tools in computer science (i.e. graph theory, mutual information and general modeling approaches). The goal is more than understanding how yeast functions; the goal is to move beyond generating large complex networks to a deeper understanding of human inheritance, including its susceptibility to environmental stress.

Galactose Utilization Pathway in Yeast

In a changing environment, nutrient availability and different sources of energy are the major factors controlling growth and development of living micro-organisms, such as yeast. The preferred source of carbon and energy for yeast is glucose. Upon depletion of glucose, yeast cells will switch to other available sugars (i.e. galactose, maltose or raffinose). In fact, glucose downregulates genes involved in the uptake and metabolization of alternative carbon sources such as GAL genes, which are required for galactose utilization pathway. GAL genes are subject to dual control of expression: they are downregulated by glucose and upregulated by galactose (see review by Gancedo J.M. 1998). In this situation, the GAL 'system' has two input signals: galactose and glucose. When glucose is unavailable, there will only be one input signal to the system. Raffinose on the other hand is a less attractive carbon source; when depleted from glucose and galactose, cells will start to metabolize raffinose (Goffrini, Ferrero and Donnini 2002). Galactose Utilization Pathway in *S. cerevisiae* is the best documented eukaryotic regulatory pathway (see review by Barnett J.A. 2007). Through several steps, this pathway, which contains positive and negative feedbacks, metabolizes galactose as the available carbon source to an utilizable carbon source and energy. The regulatory elements of the network are Gal4p, Gal2p, Gal3p, Gal6p, Gal7p, Gal80p and Mig1p. The galactose pathway has four major stages: 1) galactose is imported into the cell by Gal2p; 2) the positive regulator, Gal3p, gets activated by intracellular galactose (Timson, Ross,

and Reece 2002; Suzuki-Fujimoto T. *et al.,* 1996); 3) Gal80p is sequestered in the cytoplasm by the active form of Gal3p, causing a shuttling of Gal80p from nucleus to the cytoplasm (Peng and Hopper 2002); and, 4) Gal4p, the transcriptional activator bound to the promoter of the GAL genes (Mizutani and Tanaka 2003), is released from the inhibitory action of nuclear Gal80p. Gal4p will therefore activate the GAL genes including GAL2, GAL3 and GAL80 forming two positive and one negative feedback loops.

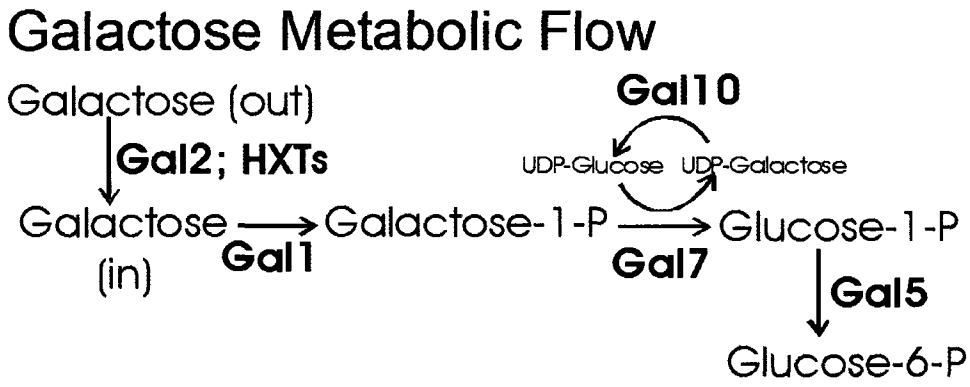


Figure 2: Galactose metabolic flow. External galactose is transported into the cell by mainly Gal2p; other transporters of galactose have also been identified (HXTs). Intracellular galactose is converted to galactose-1-phosphate by Gal1p. Galactose-1-P was shown to be toxic for cells; the latter is converted to glucose-1-P by Gal7p. Gal5p then catalyzes the conversion from glucose-1-P to glucose-6-P, which is a key step in hexose metabolism.

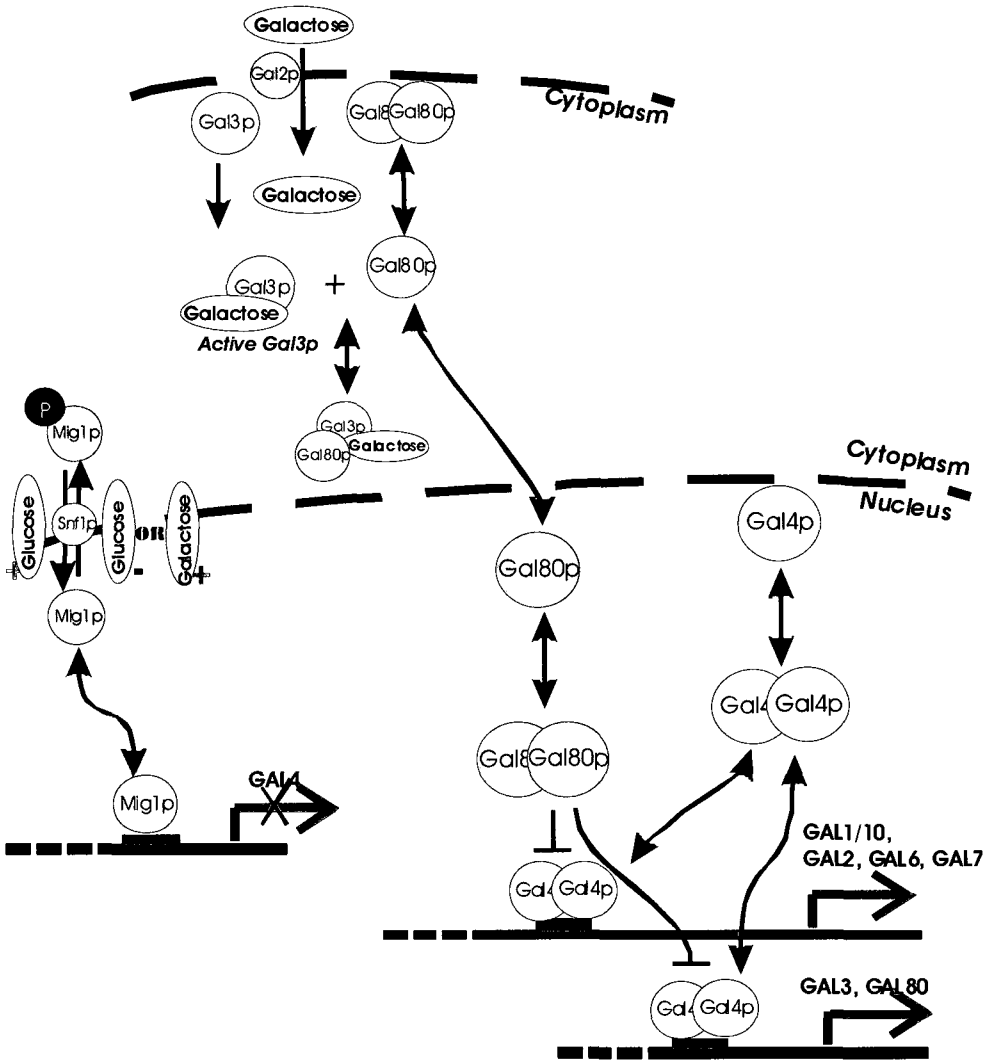


Figure 3: The Galactose Utilization Pathway in yeast has four major stages: **1)** galactose is imported into the cell by Gal2p; **2)** the positive regulator, Gal3p, gets activated by intracellular galactose **3)** Gal80p is sequestered in the cytoplasm by the active form of Gal3p causing a shuttling of Gal80p from nucleus to the cytoplasm; and, **4)** Gal4p, the transcriptional activator bound to the promoter of the GAL genes, is released from the inhibitory action of nuclear Gal80p. Gal4p will therefore activate the GAL genes including GAL2, GAL3 and GAL80 forming two positive and one negative feedback loops. Mig1p is rapidly moved to the nucleus under glucose condition and represses GAL4 expression.

GAL4 is the main regulatory element of the system. It is responsible to turn the transcriptional machinery ON and OFF when needed. Number of upstream activating sequence sites and their affinity for GAL4 are different among the GAL genes, causing differential activation (Lohr D. *et al.,* 1995). In addition to that, there exists > 300 potential GAL4 binding sites in yeast genome; however, it is not clear what are the various factors influencing proper binding (Lohr D. *et al.,* 1995). GAL genes regulated by GAL4 are sensitive to the concentration of the latter; in fact is has been shown that upon repression by glucose, GAL4's transcription decreases about five folds (Griggs and Johnston 1991), causing more than 100 folds decrease in expression of GAL1 (Griggs and Johnston 1991). In the absence of glucose, Gal80p is the main repressor of GAL4. Gal80p is the main negative regulatory element in the network. Its deletion causes the system to be in an ON state independent of the inducer level (Lamphier and Ptashne 1992). It has also been demonstrated that Gal80p tend to dimerize with high affinity and that appears to be the main factor that stabilizes the $Gal4p_2$-$Gal80p_2$ interaction (Melcher and Xu 2001). This stabilization is not due to a stronger interaction between $Gal4p_2$-$Gal80p_2$; it has been shown by cooperative binding experiments that the dimer-dimer interaction probably best covers Gal4p activation domain (Melcher and Xu 2001). Amino acids determinant for the Gal80p self association are also important for Gal80p-Gal3p association. Gal80p binds to Gal3 and this binding is competitive to the dimerization of Gal80. Gal80p-Gal80p is required for binding to Gal4p-Gal4p and the latter is in competition with Gal3p-Gal80p association (Pilauri, V. *et al.,* 2005).

Therefore, upon activation of Gal3p by galactose, Gal80p will be sequestered by the active form of Gal3p; the inhibitory effect of dimer form of Gal80p though its binding to Gal4p-Gal4p is therefore relieved (Peng and Hopper 2002), and the strength of the inhibition will depend on the concentration of inducer (galactose). The model was further refined to include NADP (adenine dinucleotide phosphate) as a player in the initial induction event (Kumar P.R. *et al.,* 2008). The original model which relies on an indirect positive feedback through Gal3p, was not able to explain the rapid induction of the GAL system. It is suggested, through biochemical and in vivo experiments, that NADP might be acting as a second messenger in initiating the GAL system (Kumar P.R. *et al.,* 2008).

GAL6, a Cytoplasmic Amino-Peptidase (Trumbly and Bradley 1983), has also some regulatory role and its deletion has been shown to have an inhibitory action on Gal4p (Zheng, Xu, and Johnston 1997); its deletion leads to an increase of expression of GAL1, GAL2 and GAL7 genes. GAL6, as well as all the known galactose-regulated genes, have the well characterized 17bp Gal4p binding site in their promoter; the latter is used by Gal4p to activate transcription upon induction by galactose. GAL3, GAL80 and GAL6 have all one binding site for Gal4p (Bajwa, Torchia and Hopper 1988; Shimada and Fukasawa 1985) compared to two to five binding site for structural genes (Melcher, K. 1997). Furthermore, structural genes are induced 50-1000 folds upon induction where

regulatory genes including GAL6 are induced only 2-10 folds (Johnston M., 1992). Nonetheless, how GAL6 conveys the negative regulation is still not clear.

Mig1p is also involved in the regulatory machinery of the GAL network and is controlled primarily by Snf1, a protein kinase. Mig1p is rapidly moved to the nucleus under glucose condition (Ahuatzi D. *et al.* 2007; Devit, Waddle, and Johnston 1997) and is back again in the cytoplasm upon glucose depletion (Papamichos-Chronakis, Gligoris, and Tzamarias 2004). Mig1p binds to two sites (site 1 and 2) in GAL4 promoter (Frolova, Johnston, and Majors 1999) and represses GAL4 expression causing repression of galactose (Ronne H. 1995).

Gal7p catalyzes the conversion of galactose-1-phosphate to UDP-galactose in the galactose utilization pathway (De Robichon-Szulmajster H. 1958; Douglas and Hawthorne 1964). GAL7 is conserved from *E. coli* to man (Seiboth B. *et al.,* 2002) and, mutations in human *GALT*, the ortholog of yeast Gal7p, have been associated with a potentially lethal disease (Riehman, Crews and Friodovich-Keil 2001). GAL7 deletion shows altered fitness when grown on galactose (Stagoj, Comino and Komel 2005). It has been observed that in the presence of galactose, GAL7 deletion causes a reduced expression of the GAL genes. Since an accumulated level of galactose-1-phosphate is toxic, cells lacking GAL7 seem to have a mechanism to downregulate the GAL genes in order to reduce the level of galactose-1-phosphate to its threshold level (Ideker T. *et al.,* 2001). The mechanism of this regulation is still unknown.

Modeling Gene Regulatory Networks

Systems biology is by definition multidisciplinary: it combines engineering and molecular biology to better understand biological principles. The interplay between hypothesis generation through simulation studies and experimental corroboration has been very successful during the recent years. Building a model of biological process can be used to predict behaviour of the system under conditions not tested experimentally and therefore guide experimentalists to design and implement experiments that are likely to provide more information. Simulations that are paired with experimental validations have gained significant interest in systems biology community because of their success in providing useful insight into cellular processes.

In order to build a biologically relevant model, one needs to consider the biological question of interest. However, a successful model should allow prediction of system's behaviour under various conditions. The non-observed conditions would have to be experimentally supported; this iterative process between experimental corroboration and model prediction would eventually refine our understanding of the underlying design principles of the system of interest (see review by Di Ventura B. *et al., 2006*).

For systems of limited size and complexity, it is possible to obtain high quality quantitative data in order to better understand for instance the effect of external or internal perturbations. Quantitative models, such as ODE (Ordinary Differential

Equations) based models, are often used when quantitative data is available. An ODE based model could be very general or more specific with detailed information about for instance complex formation or cooperative binding of transcription factor to a promoter region of a gene.

Being a relatively well-defined system, the Galactose Utilization Pathway, was subject to many simulation studies (Smidtas, Schachter, and Kepes 2006; Verma, Bhat, and Venkatesh 2003; Ideker T. *et al.* 2001; Ramsey S. *et al.* 2006; Bhat and Venkatesh 2005; De Atauri, P. *et al.,* 2004), where ODE based models were used to generate new hypothesis.

Modeling Galactose Utilization Pathway

Galactose Utilization Pathway has been of great interest because of its intriguing structural dynamics. In a complex-yet-elementary fashion, this network which is under dual control of glucose and galactose, is able to display persistent memory of its previous consumption state (Acar, Becskei and van Oudenaarden 2005). A model of galactose utilization pathway was constructed using ODEs to investigate the design principles underlying this functionality (Acar, Becskei and van Oudenaarden 2005). In that model, the idea was to separate the fast reactions from the relatively slow reactions in order to study the stochastic transition between stable states. The fast reactions such as protein-protein and protein-galactose interactions were assumed to be in equilibrium with respect to the slow reactions (i.e. mRNA synthesis). The rate equations were then used to evaluate the dynamics of these reactions. Some of the parameters such as relative concentration of Gal80p were determined experimentally, using fluorescence microscopy (where Gal80p was fused to CFP, cyan fluorescent protein). Some other parameters were estimated by fitting the model to some experimental data. Using the constructed model, it was possible to explore the space in a systematic manner to better understand the design principles that produce destabilization of the memory (Acar, Becskei and van Oudenaarden 2005). The model predicted that GAL80p reduces the strength of the positive feedback loop and, "tunes the system from having destabilized memory to having a persistent memory" (Acar, Becskei and van Oudenaarden 2005). The regions

where the model predicted the presence of 'persistent memory' were characterized for validation. This hypothesis was then verified experimentally and the results were in accordance with model prediction (Acar, Becskei and van Oudenaarden 2005). This beautiful marriage between model prediction and experimental validation is a great example of a systems biology approach to shed lights on some fundamental biological questions.

Galactose utilization pathway has also been used to investigate cellular heterogeneity, or transcriptional noise (Ramsey S. *et al.,* 2006). Using kinetic modeling based on ODEs, it was possible to demonstrate and further confirm experimentally that the positive feedback by Gal3p and the negative feedback by Gal80p are evolved together to control cell-to-cell variability and ensure the rapid transition to an induced state upon galactose uptake (Ramsey S. *et al.,* 2006). The ODEs used for this model are very clear and include all the known reactions: import of galactose to the cell by Gal2p, activation of Gal3p by galactose, sequestration of Gal80p by activated Gal3p, dimer formation of Gal80, shuttling of Gal80p-Gal80p to and from the nucleus, deactivation of Gal4p-Gal4p by nuclear form of Gal80p-Gal80p, mRNA production and decay, protein production and decay, and finally cooperative binding of $Gal4p_2$-$Gal80p_2$ for induction of GAL genes depending of the number of binding sites in the promoter regions (GAL3 and GAL80 are assumed to have only one binding site for Gal4p and GAL2 and GAL1 five and four binding sites respectively (Melcher and Xu 2001; Davidson E.H.. 2001; Fell D.A.1997).

Based on a previous model by De Atauri P. et al., (2004), this model is derived from the observation that Gal3p is localized only in the cytoplasm (Peng and Hopper 2000 and 2002) and the gal80p-gal4p-gal3p complex which was shown to exist in vitro will not be taken into consideration. The model includes 19 species (19 ODEs) and 44 parameters (Ramsey S. et al., 2006). The first and last metabolic steps in the network are the import of galactose into the cell and the conversion of galactose to galactose-1-phosphate by galactokinase Gal1p respectively. The model assumes a constant extracellular concentration of galactose which is also a free parameter in the model and corresponds to a continuous culture system. Protein-DNA interactions are assumed to occur sufficiently rapidly such that the system could be assumed at quasi-steady-state (QSS) with respect to free transcription factor concentration in the nucleus. The model also assumes no feedback from Gal1p which is 40 times less effective than its homolog Gal3p (Platt and Reece 1998) and, no glucose repression through MIG1. The model also assumes galactose uptake only through the galactose transporter Gal2p. The rate and equilibrium constants have been taken from various sources and are all referenced in the original publication (Ramsey S. *et al.* 2006). Finally, this refined model of the GAL network generated new hypothesis and directed experimentalists to design and implement experiments that would potentially give useful insights into the biological properties of the system. By understanding how, for instance, simple feedback loops operate in a biological networks, thought simulations and experimental studies, we will be able to

understand the rationale behind existence of certain molecular blueprints that are often present in biological networks.

Generic Modeling of Gene Regulatory Networks

Other modeling approaches are based on a more generic formulation in order to tackle the system for uncovering undetermined interactions. In those instances, through high-throughput or other systematic experiments, it is possible to identify key players involved in the system of interest; however, the physical or genetic interaction among these players will have to be identified by more accurate means. For instance, perturbation experiments or time-series data could provide some insights into regulatory interactions of the network. In a recent study, a modeling framework to study dynamical system was applied to expression data of early *Drosophila melanogaster* development. The goal of the study was to determine the structure and dynamics of Drosophila's genetic network (Perkins, T.J. *et al.,* 2006); the approach was more generic, but required a careful parameter optimization technique to handle the significant computational time due to the large parameter space. The model was based on ordinary differential equations and incorporated protein production, decay and diffusion (see figure 4).

$$\frac{\partial v^a(x,t)}{\partial t} = \zeta(t)P^a(v(x,t)) - \lambda^a v^a(x,t) + D^a \frac{\partial^2 v^a(x,t)}{\partial x^2},$$

$$P^a(v(x,t)) = R^a g\left(\sum_b T^{ab} v^b(x,t) + h^a\right)$$

Figure 4: A generic modeling approach for reverse-engineering of dynamical systems. First equation represents the Gap gene dynamics using reaction-diffusion partial differential equation. First and second terms represent respectively the production and decay rates, the third term accounts for the protein diffusion rate. Second equation represents the production rate, where R is the maximum synthesis rate, T encode the effect of one protein on the protein of interest, h is the basal level of expression (fixed to -3.5 for all genes) and $g(u)=1/2((u/\sqrt{(u^2+2)})+1)$ (adapted from Perkins, T.J. *et al.*, 2006)

The production component of the model contained terms for maximum production rate, interaction among proteins (activating or repressing signals) and basal level of expression. The model fitting exercise had three stages: I) finding an initial estimate of the decay and diffusion parameters for each gene, ii) determining an initial estimate of the regulatory parameters and finally iii) using a local optimization procedure to fit the model to experimental data (Perkins, T.J. *et al.,* 2006). The main challenge of that system was its spatio-temporal aspect which was translated into the diffusion term of the model (Perkins, T.J. *et al.,* 2006).

Once a model of a biological system is designed, then the challenging next step is to fit the available data to that model in order to be able to predict the behaviour of the system for other non-tested conditions. The latter seems simple, yet it requires a significant computational power. Elegant solutions have been proposed over the years (see review by Jaqaman and Danuser 2006; and by Gardner and Faith 2005) and depending on the complexity of the model, different techniques are being used and, in many cases a combination of different approaches is necessary. This problem in computer science is known as parameter optimization and model discrimination. The former is an initial attempt to find a set of models to reproduce the data and the latter is to design a set of experiments or narrow the constraints to discriminate among these models (review by Kremblig A. 2004; Lorenz S. *et al.,* 2007). Sensitivity analysis-based methods are also means to reduce the model by eliminating parts of it that are locally unimportant for the

property of interest (Brown, N.E. *et al.,* 1997; Danø S. 2006; Edelson D. 1981; Gautier and Carr 1985; Okino and Mavrovouniotis 1998). In fact, it is important to note that biological networks generally follow a power-law distribution where most of nodes in the network have limited connectivity among them and a few of the nodes are highly connected (Barabasi and Albert 1999; Grandison and Morris 2008). It is therefore important to reduce the number of links among the genes if that does not contribute significantly to the fitness score. It has also been shown, though a systematic experiment in *E. coli,* that in 95% of the cases (~600 strains), adding a new link to a biologically evolved network increases the fitness of that network only slightly. In few cases, the added interaction was detrimental (Isalan M. *et al.,* 2008).

Parameter Optimization

Evolutionary Algorithms: Principles and Concepts. Optimization technique and heuristic search are active fields of research in artificial intelligence (AI) and computer science (see textbook by Luger G.F., 2005). The simplest method is a random search-and-test method, where an arbitrary candidate solution is selected from a pool of possibilities and tested; the test stops when it is successful or the optimized criteria is found to be at its minimum. The random search is a brute force technique and is usually used to scan the search space prior to a rigorous optimization.

The optimization criterion is well understood by the notion of a fitness function (also known as cost function, penalty function or objective function). Fitness function is measure of distance from the candidate solution to the exact solution. This measure depends on the specific case under investigation. In biological setting, it could be the difference between observed expression level and the simulated value. Examples include and are not limited to the difference between observed doubling time of a given strain and the doubling time predicted by the model, or the up-regulation of genes predicted by a model and observed by microarray experiment.

Other well investigated search-and-optimization techniques are hill-climbing, simulated annealing and genetic algorithms. Hill-climbing is a very intuitive and simple technique. The idea is to select and test a candidate (a parameter set) at random; this

individual is then modified slightly yielding to a new candidate that will be also tested. If the new candidate produces an improved fitness then the modification is kept and the latter will be used to generate a third candidate and so forth until there exist no modification that will decrease the fitness function of the last optimum candidate. The last candidate will be called a local optimum (or suboptimal) individual, which could be or not the best solution. To find out if the local optimum is in fact the best solution, the algorithm has to be able to find this candidate independent of the initial candidate. Thus, to have fair confidence, it is required to run this algorithm many times, starting each time from a different parameter set.

Simulated annealing is based on the Metropolis algorithm at well-chosen, ever decreasing temperatures. Metropolis algorithm differs from the basic hill-climbing algorithm in the following: the algorithm will accept an inferior individual with a given probability which is a function of the difference in fitness and the temperature at which the algorithm is operating. Simulated annealing algorithm guarantees to find the optimum solution.

Genetic algorithms have the benefit of maintaining a population of candidates from which new set of candidates are generated through crossover. To introduce diversity in the population, the new set of candidates will go through mutation where a random parameter will be varied by a random percentage. The algorithm will then remove the least fit individuals to keep the population size fixed. This iterative process will continue

until convergence is reached or a maximum number of iteration is performed. In the next section, I will limit myself to further explain in detail Genetic Algorithm; GA, as its name indicates, is the father of biologically inspired heuristics.

Genetic Algorithm

Genetic Algorithm is an evolutionary algorithm as described briefly above. A basic version of the algorithm is adapted from Reeves and Rowe (2003) and presented in the following section:

```
Choose an initial population of chromosomes;
While termination condition not satisfied do
        Repeat
                If crossover condition satisfied then
                        Select parent chromosomes;
                        Choose crossover parameters;
                        Perform crossover.
                If mutation condition satisfied then
                        Select chromosome(s) for mutation;
                        Choose mutation points;
                        Perform mutation;
                        Evaluate fitness of offsprings
        Until sufficient offsprings created;
        Select new population;
End while
```

Figure 5: A genetic algorithm template adapted from Reeves and Rowe (2003). The general formula could be adapted to many specific optimization problems.

Initial population: The first question is the size of the population; as mentioned, GA keeps the population size fixed and therefore it is important to investigate the proper size of the population to allow for effective search space. There are some suggestions (Grefenstette J.J. 1986; Rowe, Vose and Wright 2004) that population sizes as small as 30 are quite adequate in many cases; however there has been indication that population size should grow with string length (Goldberg and Deb 1992). The second question is how to choose the initial population. Usually the initialization is random; however seeding the initial population with a known good solution (Ahuja and Orlin 1997; Reeves C.R. 1994), obtained from another heuristic technique, could help the genetic algorithm find better solutions faster. It should also be noted here that seeding technique could induce a premature convergence and therefore to a poor solution (Kapsalis, Rayward-Smith and Smith 1993; Levine D. 1997). Careful analysis is needed to ensure proper convergence if the initial condition is not randomly selected.

Termination: Since genetic algorithms do not terminate when a local optimum is reached, they could in practice run forever or for a very long time; therefore a limit must be set to avoid such situations. Termination criteria could be simply based on the number of iterations or computer clock time. The GA will terminate if 1) the predefined condition is reached or 2) fitness values have reached convergence based on some given threshold value.

Crossover: In evolutionary terminology, chromosomal crossover is the process of

exchanging genetic material between two chromosomes. The idea is used to define crossover in GA. The following figures will show different types of crossover used in GA. The n-point crossover has great value mainly because it does not treat the problem at hand in a sequential manner; in that case, n could be the number of parameters in an ODE, where set of equations do not have any linear dependency among them.

The generalized n-point crossover is general framework for recombination and could easily be further generalized to allow more than two parents for the creation of children.

Selection: Selection is related to fitness; the intuitive idea is to select the population based on their relative fitness. The individuals with less favourable fitness values will be discarded to keep the population size constant. However, as the population converges, there exist duplicate parents and a crossover between them will recreate the parent again. This could be easily detected and force the termination.

Mutation: In biology, changes to the base pair sequences of DNA are called mutations. Again the same principle applies when defining mutation in genetic algorithms. Mutations are simple processes to maintain genetic diversity among the population from one generation to the next. As John Holland, the pioneer of GA said "the essence of a good GA design is the retention of diversity" (Holland J.H., 2000). In a binary setting, mutation is the random selection of a bit from a given sequence of bits and

the inversion of its value (i.e. from 0 to 1 or from 1 to 0). For instance, a bit-string 1001101 can go through mutation and become 1001100, where the last bit has been altered. In a more generalized case, there are two issues to consider: 1) how many mutation points will affect a given chromosome, 2) where these mutations will take place. Next figure shows this concept.

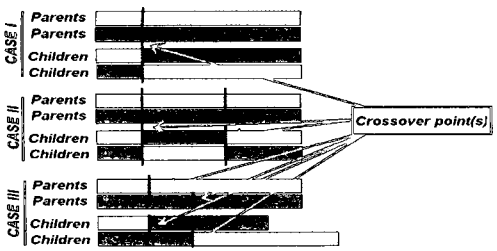


Figure 6: Three basic cases of crossover techniques in GA. Case I is the simple one-point crossover, case II represents two-point crossover, the latter could be expanded to a generalized form of *n*-point crossover, where n is the number of genes in the chromosome. Case III represents a variant approach where the length of the children is different from the length of the parents.

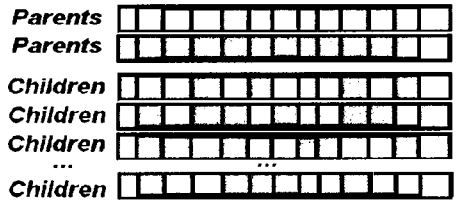


Figure 7: *N*-point crossover. The children are created based on a random combination of the parents. A binary mask could be applied for creation of each child. For instance 1001010001010 could represent the first child, where 1s are genes taken from parent 1 and 0s are genes inherited from the second parent.

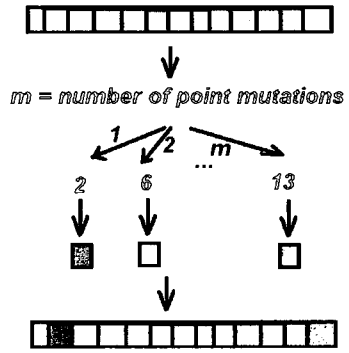


Figure 8: Mutation in GA. A random number is drawn (usually from a Poisson distribution) representing the number of mutation per chromosomes (m). M random numbers are then drawn from a uniform distribution representing the loci where the mutation would take palace. Last step is to apply the mutation and create offsprings. A changing mutation rate is considered by many as a diversity maintenance policy (Fogarty T.C. 1989; Reeves C.R. 1994).

The exact algorithm to introduce mutation will depend on the question of interest. The only idea is to introduce reasonable level of population diversity. Nonetheless a balance between mutation and crossover rate if often problem-specific.

New population: New population is generated by the crossover and mutation. There are different techniques used to select the new best population. The original GA (Reeves and Rowe, 2003) assumed replacement strategy where the new population created from crossover and mutation would replace the old population; however, this is not very efficient. A second approach developed by De Jong (De Jong K.A. 1975) introduced the idea of population overlaps. This would ensure the survival of the best individuals by preserving them and replacing only the remaining (N-1) members of the population with new strings. The idea was further developed by Davis (Davis L. 1991); his idea was to use an incremental strategy in which only one new individual is created at each stage. There are other variations for creation of the new population in the evolutions-strategy community (Goldberg and Deb 1991; Reeves C.R.1994).

Post-Optimization and Model Discrimination

Biological networks are sparse: some nodes have many interactions and others have a limited number of connections. This characteristic provides a mean of robustness to the system (Stelling *et al.,* 2004; Wagner A. 2005). In addition, using systematic means to optimize a set of parameters would generally result in networks with full connectivity. Different formulations have been implemented to avoid this problem: In some cases, the technique would limit the number of connections. (Gardner T.S. *et al.,* 2003) in other cases, parameter optimization is performed with no limitation and a refinement (or post-optimization procedure) is then implemented to complement the optimization (Morohashi, M. *et al.,* 2002). Network refinement is crucial in system-identification and reverse-engineering, and is followed by model discrimination technique. Optimization process will identify a set of possible candidate solutions which will be equally good to reproduce the data. Model discrimination, on the other hand, consists of ranking the optimized parameters based on more stringent criteria, or by analyzing new experimental data (Feng X.J., 2004; Kremling, A. *et al.,* 2004).

Biological networks are robust to perturbations, leading to an intuitive strategy where the selection criteria is based on the fact that an acceptable parameter set should be robust to withstand changes to its values. This selection method, known as parameter sensitivity analysis (Morohashi, M. *et al.,* 2002), is based on evaluating the range of parameter space capable of reproducing the experimental data. This range provides

insight into robustness of parameter set.

Chapter 2: Material and Methods

The Four Stages of SGPA Framework

SGPA framework is based on a data-driven approach for investigation of relationships among regulatory elements in a network. There are five major steps, all equally important for a successful experiment. The general idea is to generate a library of strains carrying one or two copies of the genes of interest. The pair-wise gene deletion assures that the network topology keeps its natural structure. Decreasing gene copy number from two copies to one copy will disrupt the balance of different protein levels. This information will generate sets of hypotheses to address the question of reverse-engineering regulatory networks. However, since gene dosage does not translate to an accurate reduction in protein level across the genome, full knock-out strains will be used to validate and test potential candidate network structures, and select the one(s) that are conform to the experimental observations. The following figure (figure 9) outlines a coarse-grain view of these five steps.

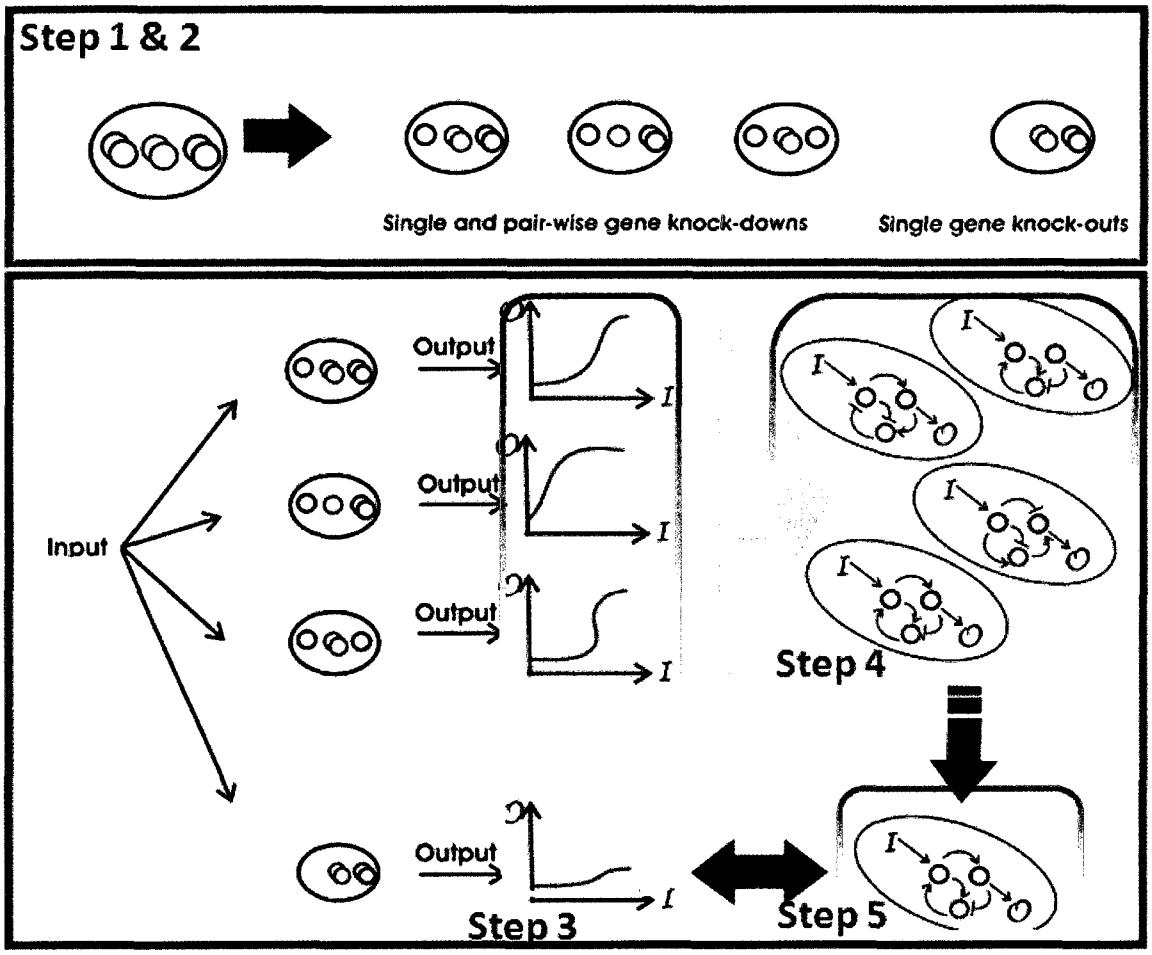


Figure 9: System Identification based on Systematic Gene Perturbation Analysis. Step 1: identification of key regulatory players in the network of interest; step 2: systematic generation of diploid strains with single and pair-wise gene knock-downs and single gene knock-outs; step 3: high quality data generation for all the strains for a range of input values; step 4: model development and generation of potential topologies conform with the experimental data for knock-down strains; step 5: network selection based on experimental data from knock-out strains.

Strains and Culture Conditions

Saccharomyces cerevisiae strains used are from BY4741 and BY4742 background. Media composition is YPG (yeast peptone galactose) with 2% galactose and YPR, 2% raffinose. SM stands for Synthetic Media. The integration strategy was performed to replace the Ade2 site in the yeast genome with a P_{Gal1}-GFP-His3 cassette in Mat (alpha) and URA3 cassette in Mat(a) strains. Primers for integration are included in the appendix.

Strains carrying Ade2 deletion are red when plated on media without adenine. These strains require adenine to gain their natural color (white). This characteristic is used for colony selection after transformation and prior to validation by PCR. Deletion strains were obtained from yeast deletion library and were PCR verified using the two ORF-specific confirmation primers and two KanMX4 specific primers (http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html). The diploids were constructed by mating the Mat(a) and Mat(alpha) strains; further PCR validation was performed and the primers for PCR validation are included in the appendix section.

An extension to the library of heterozygote strains is constructed where each strain has one copy of two genes (e.g. Gal3+/-; Gal80+/-) with two selection markers (URA3 and HIS3). The list of haploid and diploid strains and their genotypes is presented in table 1 and 2.

Strain name	ID	Genotype
BY4741;Ura3;MIG1Δ	A1-URA3	BY4741 (MATa); ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; **MIG1**::kanMX4; ADE2::**Ura3**
BY4741; Ura3;GAL80Δ	B1-URA3	BY4741 (MATa); ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; **GAL80**::kanMX4; ADE2::**Ura3**
BY4741; Ura3;GAL4Δ	C1-URA3	BY4741 (MATa); ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; **GAL4**::kanMX4; ADE2::**Ura3**
BY4741; Ura3;GAL2Δ	D1-URA3	BY4741 (MATa); ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; **GAL2**::kanMX4; ADE2::**Ura3**
BY4741; Ura3;GAL3Δ	E1-URA3	BY4741 (MATa); ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; **GAL3**::kanMX4; ADE2::**Ura3**
BY4741; Ura3;GAL6Δ	F1-URA3	BY4741 (MATa); ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; **GAL6**::kanMX4; ADE2::**Ura3**
BY4741;Ura3	G1-URA3	BY4741 (MATa); ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; ADE2::**Ura3**
BY4742; $P_{Gal1/10}$-yEGFP-His3;MIG1Δ	A2-HIS3	BY4742 (MATα); ura3Δ0; leu2Δ0; his3Δ0; lysΔ0; **MIG1**::kanMX4; ADE2::**$P_{Gal1/10}$-yEGFP-His3**
BY4742; $P_{Gal1/10}$-yEGFP-His3;GAL80Δ	B2-HIS3	BY4742 (MATα); ura3Δ0; leu2Δ0; his3Δ0; lysΔ0; **GAL80**::kanMX4; ADE2::**$P_{Gal1/10}$-yEGFP- His3**
BY4742; $P_{Gal1/10}$-yEGFP-His3;GAL4Δ	C2-HIS3	BY4742 (MATα); ura3Δ0; leu2Δ0; his3Δ0; lysΔ0; **GAL4**::kanMX4; ADE2::**$P_{Gal1/10}$-yEGFP- His3**
BY4742; $P_{Gal1/10}$-yEGFP-His3;GAL2Δ	D2-HIS3	BY4742 (MATα); ura3Δ0; leu2Δ0; his3Δ0; lysΔ0; **GAL2**::kanMX4; ADE2::**$P_{Gal1/10}$-yEGFP- His3**
BY4742; $P_{Gal1/10}$-yEGFP-His3;GAL3Δ	E2-HIS3	BY4742 (MATα); ura3Δ0; leu2Δ0; his3Δ0; lysΔ0; **GAL3**::kanMX4; ADE2::**$P_{Gal1/10}$-yEGFP- His3**
BY4742; $P_{Gal1/10}$-yEGFP-His3;GAL6Δ	F2-HIS3	BY4742 (MATα); ura3Δ0; leu2Δ0; his3Δ0; lysΔ0; **GAL6**::kanMX4; ADE2::**$P_{Gal1/10}$-yEGFP- His3**
BY4742; $P_{Gal1/10}$-yEGFP-His3	G2-HIS3	BY4742 (MATα); ura3Δ0; leu2Δ0; his3Δ0; lysΔ0; ADE2::**$P_{Gal1/10}$-yEGFP-HIS3**
BY4741; $P_{Gal1/10}$-yEGFP-His3;GAL7Δ	H1-HIS3	BY4741 (MATa); ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; **GAL7**::kanMX4; ADE2::**$P_{Gal1/10}$-yEGFP- His3**
BY4741; Ura3;GAL7Δ	H1-URA3	BY4741 (MATa); ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; **GAL7**::kanMX4; ADE2:: **Ura3**
BY4742;Ura3	G2-URA3	BY4742 (MATα); ura3Δ0; leu2Δ0; his3Δ0; lysΔ0; ADE2::**Ura3**
BY4742; $P_{Gal1/10}$-yEGFP-His3;GAL7Δ	H2-HIS3	BY4742 (MATα); ura3Δ0; leu2Δ0; his3Δ0; lysΔ0; **GAL7**::kanMX4; ADE2::**$P_{Gal1/10}$-yEGFP-His3**

Table 1: List of haploid strains. All strains have been PCR confirmed. The haploid strains are used to construct a library of diploid strains as shown in the next table.

Nb	Strain name	BY4741 (MATa)	BY4742 (MATα)
1	$MIG1^{-/-}$	A1-URA3	A2-HIS3
2	$GAL80^{-/-}$	B1-URA3	B2-HIS3
3	$GAL4^{-/-}$	C1-URA3	C2-HIS3
4	$GAL2^{-/-}$	D1-URA3	D2-HIS3
5	$GAL3^{-/-}$	E1-URA3	E2-HIS3
6	$GAL6^{-/-}$	F1-URA3	F2-HIS3
7	GAL7-/-	H1-URA	H2-HIS3
8	$MIG1^{+/-}$	G1-URA3	A2-HIS3
9	$GAL80^{+/-}$	G1-URA3	B2-HIS3
10	$GAL4^{+/-}$	G1-URA3	C2-HIS3
11	$GAL2^{+/-}$	G1-URA3	D2-HIS3
12	$GAL3^{+/-}$	G1-URA3	E2-HIS3
13	$GAL6^{+/-}$	G1-URA3	F2-HIS3
14	GAL7+/-	G2-URA3	H1-HIS3
15	$GAL80^{+/-}$;$MIG1^{+/-}$	B1-URA3	A2-HIS3
16	$GAL80^{+/-}$;$GAL2^{+/-}$	D1-URA3	B2-HIS3
17	$GAL80^{+/-}$;$GAL3^{+/-}$	E1-URA3	B2-HIS3
18	$GAL80^{+/-}$;$GAL4^{+/-}$	B1-URA3	C2-HIS3
19	$GAL80^{+/-}$;$GAL6^{+/-}$	B1-URA3	F2-HIS3
20	$GAL80^{+/-}$;$GAL7^{+/-}$	H1-HIS3	B2-HIS3
21	$GAL2^{+/-}$;$MIG1^{+/-}$	E1-URA3	A2-HIS3
22	$GAL2^{+/-}$;$GAL3^{+/-}$	D1-URA3	E2-HIS3
23	$GAL2^{+/-}$;$GAL4^{+/-}$	D1-URA3	C2-HIS3
24	$GAL2^{+/-}$;$GAL6^{+/-}$	E1-URA3	F2-HIS3
25	$GAL2^{+/-}$;$GAL7^{+/-}$	H1-HIS3	D2-HIS3
26	$GAL3^{+/-}$;$MIG1^{+/-}$	E1-URA3	A2-HIS3
27	$GAL3^{+/-}$;$GAL4^{+/-}$	E1-URA3	C2-HIS3
28	$GAL3^{+/-}$;$GAL6^{+/-}$	E1-URA3	F2-HIS3
29	$GAL3^{+/-}$;$GAL7^{+/-}$	H1-HIS3	E2-HIS3
30	$GAL4^{+/-}$;$MIG1^{+/-}$	C1-URA3	A2-HIS3
31	$GAL4^{+/-}$;$GAL6^{+/-}$	C1-URA3	F2-HIS3
32	$GAL4^{+/-}$;$GAL7^{+/-}$	H1-HIS3	C2-HIS3
33	$GAL6^{+/-}$;$MIG1^{+/-}$	F1-URA3	A2-HIS3
34	$GAL6^{+/-}$;$GAL7^{+/-}$	H1-HIS3	F2-HIS3
35	$GAL7^{+/-}$;$MIG1^{+/-}$	H1-HIS3	A2-HIS3
36	Wild-type	G1-URA3	G2-HIS3

Table 2: List of diploid strains. All strains have been PCR confirmed. The strain names will be used in the report for simplification. The genotypes for Mat(a) and Mat(alpha) that were used to generate the diploid strains are outlined in table 1.

Growth Conditions and Fitness Profiling

Fitness profiling is performed for identification of key regulatory elements of the GAL system. Fitness is defined as the ratio of wild type and perturbed strains' doubling times; doubling time is calculated based on assumption of exponential growth (see figure 10). To measure the fitness score for perturbed strains, growth of wild-type and perturbed strains is monitored for a total of 24 hours.

Initial overnight inoculation was done in YPR; saturated cells were washed and diluted to very low density (initial optical density of 0.02). Different media composition used for the growth assay contained SM without URA-HIS, adenine and the indicated amount of sugar. Plate reader was used for continuous measurement of optical density (600nm) in a 96 well plate at 30 degree Celsius.

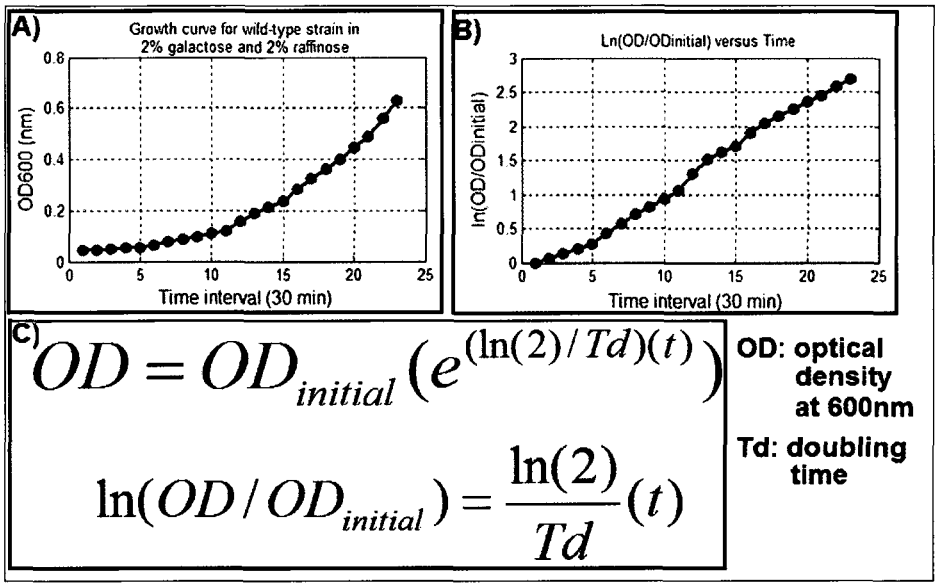


Figure 10: Fitness measure. a) Growth curve (optical density versus time) of wild-type strain. B) Growth curve plotted as $\ln(OD/OD_{initial})$ versus time; this plot has a slope of ln(2)/Td. c) Equation describing growth based on exponential growth assumption. A plot of $\ln(OD/OD_{initial})$ versus time has a slope of ln(2)/Td; therefore by calculating the slope of such a graph it is possible to estimate the doubling time.

Expression Profiling by Flow Cytometry

Fitness profiling is used for 1) identification of key regulatory elements of the GAL network and 2) generation of high quality quantitative data for system-identification of gene regulatory networks.

Overnight inoculation of yeast cells is done in YPR; saturated cells were washed and diluted to a range of different optical density based on their growth in various galactose concentrations. Cells were diluted in synthetic media (URA-, HIS- for selection) containing 2% raffinose, adenine and the indicated percentage of galactose. A total of 400ul of cell cultures were incubated in a 96 well plate (96 well plate with 1.0ml well capacity) for 7 hours at 30 degrees Celsius in a tilted position (45 degree angle) while shaking in the incubator. Seven hours incubation was shown to be sufficient to reach steady state for the wild-type and heterozygous strains in rich media – as also reported by previous publications (Acar, Becskei and van Oudenaarden 2005). The aim was to have low density cells growing logarithmically such that the final OD prior to flow Cytometry measurement would be less than 0.1. Low density cell population would ensure a relatively constant sugar concentration throughout the seven hours incubation time. In all the measurements, the OD prior to expression profiling was between ~0.08 and 0.1. The presence of raffinose ensures the sugar and energy supply even with very low galactose concentrations. The flow Cytometry experiment is performed with the same setting for all the strains and all the conditions specified in the report.

Data analysis and Data Fitting

Results from expression profiling assay, done by flow Cytometry, were analyzed using Matlab. A total maximum of 30,000 cells were captured by flow Cytometry; using gating in CXP (software for analysis of flow Cytometry data), 10% of the total population (~3000 cells) were selected for analysis. The small gate captured cells with similar size and shape, based on the fact that small cells are from newer generation and larger cells are from an older generation. This process maximizes reproducibility since cells with similar morphology, are likely to be in the same cell cycle and comparable to each other.

Experimental data was fitted to two Gaussian distributions for further comparative study; this procedure was based on the fact that all the experimental data captured cells as one or two populations. Data was further normalized to account for number of cells for each measurement. The following figures (figure 11 and 12) demonstrate the data fitting and normalization procedures.

Figure 11: Normalization of flow Cytometry data using curve fitting methods in Matlab. **A)** Representation of data points along with a fitted curve. The data fitting process is automated for all the flow Cytometry data. The fits are however individually verified; in some cases, a manual fit with a better coverage replaced the automated fit. **B)** Fitted curve is normalized to have a p.d.f. of 1. The equation and the coefficients along with R^2 value are indicated. Note that cell count is normalized in this representation. **C)** Six general examples of experimental data along with fitted curves. **D)** Normalization of x-axis: conversion of logarithmically spaced axis to a linear axis. Red plot represents two normal distributions on a log axis (channel intensity, FL1, in the flow Cytometry data acquisition setting) and the respective green plot is its equivalent distribution on a linear scale. **E)** Example of different possible measures obtained from the single cell experimental data: mean expression of two populations; and cell percentage with respect to total population.

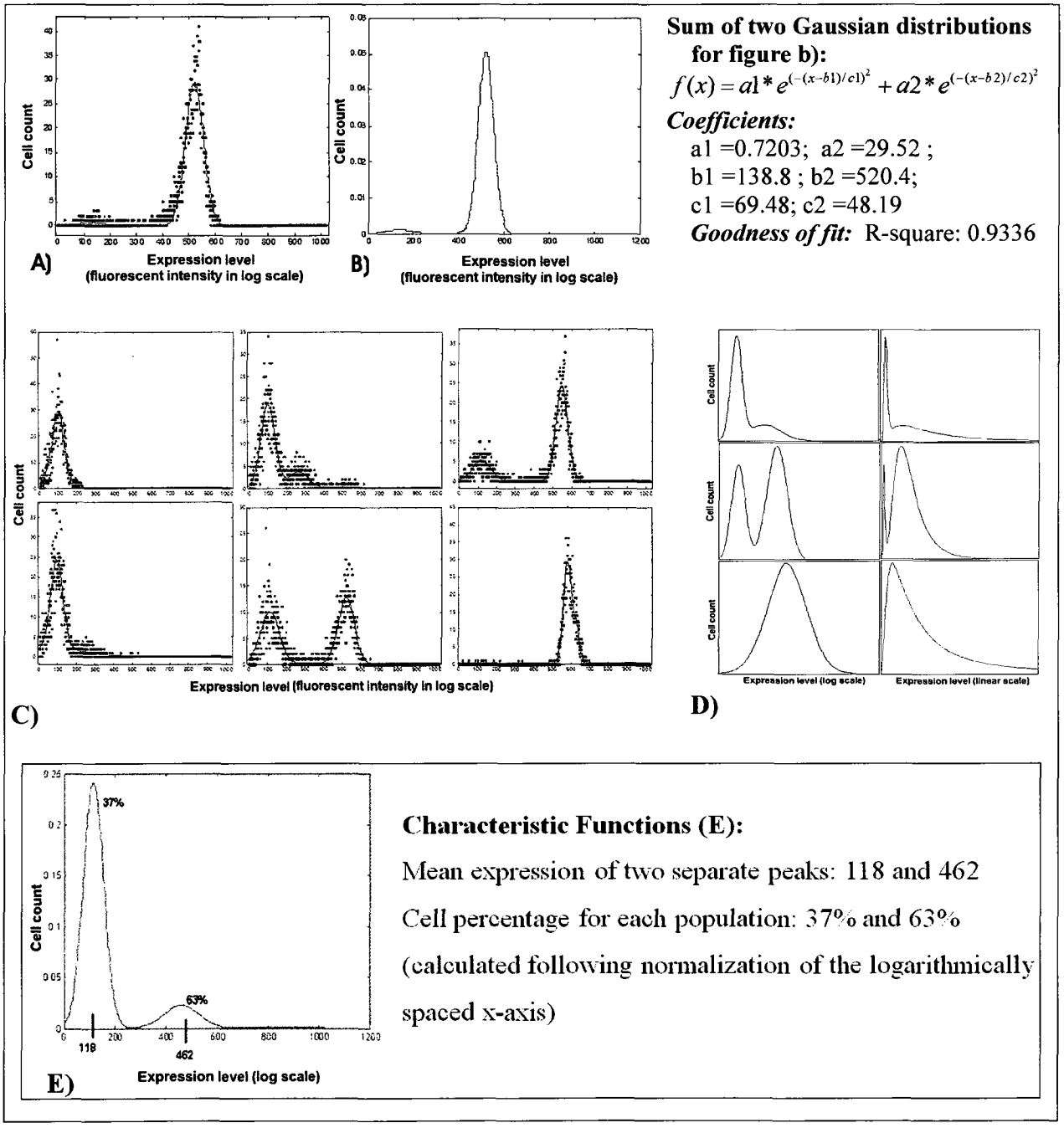

A)
Cell count
Expression level
(fluorescent intensity in log scale)
B)
Cell count
Expression level
(fluorescent intensity in log scale)
Sum of two Gaussian distributions for figure b):
$f(x) = a1 * e^{(-(x-b1)/c1)^2} + a2 * e^{(-(x-b2)/c2)^2}$
Coefficients:
a1 =0.7203; a2 =29.52 ;
b1 =138.8 ; b2 =520.4;
c1 =69.48; c2 =48.19
Goodness of fit: R-square: 0.9336
C)
Cell count
Expression level (fluorescent intensity in log scale)
D)
Cell count
Expression level (log scale)
Expression level (linear scale)
E)
37%
63%
Cell count
118
462
Expression level (log scale)
Characteristic Functions (E):
Mean expression of two separate peaks: 118 and 462
Cell percentage for each population: 37% and 63%
(calculated following normalization of the logarithmically spaced x-axis)

Figure 11

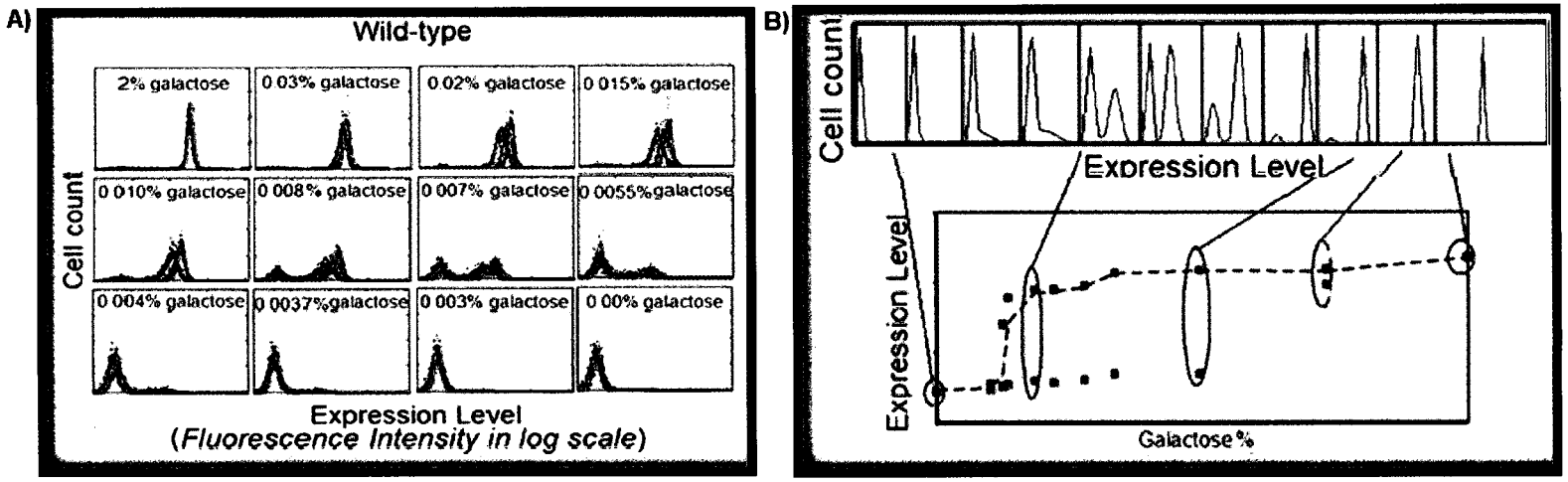


Figure 12: Dose response curve obtained following data analysis. Dose response curves will be utilized to compare different strains and for the system-identification process. A) Single cell expression level for the wild-type strain in 12 different galactose concentrations (the three colors indicate three independent experimental data). B) Single cell data is normalized (red curves) and mean expression values are calculated and plotted to generate the dose response curve. Pink squares represent the mean expression values for each population; the dashed line represents the average expression of the total population.

Comparison of Expression Profiles with respect to Wild-type

In order to compare expression profiles of perturbed strains to the expression profile of the wild-type, an integration method was developed. The integration method consisted of calculating the difference between the area under the dose response curves of perturbed strain and wild-type (see figure 13). This measure is not affected by irregular concentrations of galactose; in fact there are more data points at very low galactose concentrations. In addition to that, adding more data points at a later time, will not affect the fitness score.

The strains were ranked based on their respective p-values with respect to the 34 other perturbed strains. P-values was calculated using the **ZTEST function in Matlab7.1** *(MatWorks, Statistics Toolbox)* where the Null Hypothesis is as follows: "data in the vector X are a random sample from a normal distribution with mean M and standard deviation STD, against the alternative that the mean is not M". Vector X contains the fitness scores for all the perturbed strains; mean value and standard deviation of vector X are utilized for p-value evaluation. More detailed explanation is provided in the next section (Statistical test).

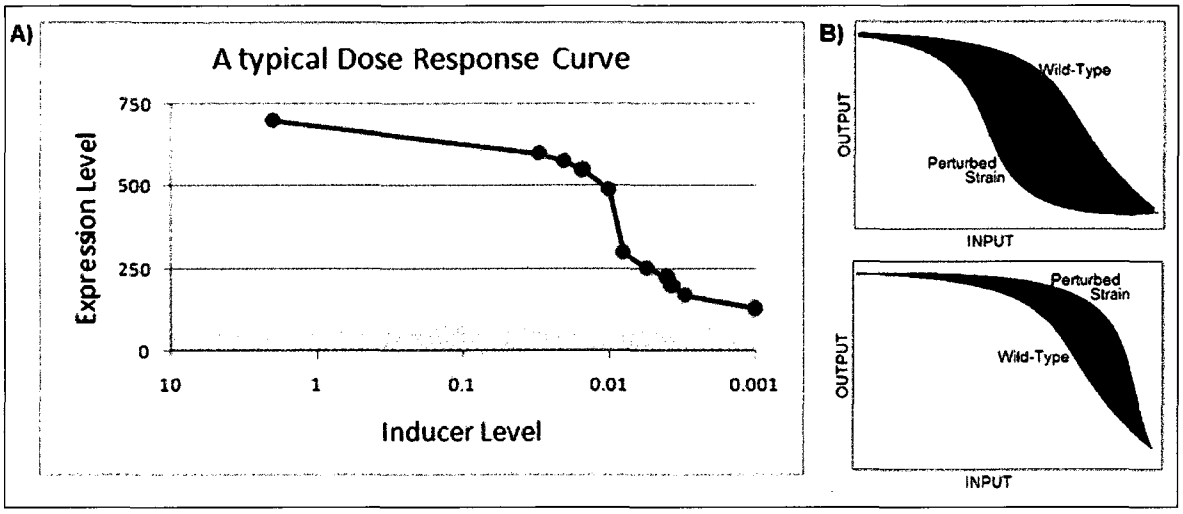


Figure 13: Integration method for comparison of dose response curves of perturbed strains to the wild-type. **A)** A typical dose response curve. There are more data points at low galactose concentrations. **B)** The area under the curve is a measure of the difference in dose response curves between wild-type and the indicated strain. For strains that have higher expression level than wild-type, the values are positive; for strains that have lower expression level than wild-type, the values are negative.

Statistical Test

To compare different expression profiles, statistical Z-test is used, where sample and population means are compared to determine if there is a significant difference. The null hypothesis tests if the difference is purely due to chance. The z-score is indicative of the difference between the sample and the population mean by standard deviation units. P-values are then obtained from the Z table. This statistical test is done in Matlab using the command [H,P,CI] = ZTEST(X,M,SIGMA,ALPHA), which performs a Z-test of the hypothesis that the data (X) come from a distribution with mean M and standard deviation SIGMA. ALPHA is set to 0.05. The returned values are i) H: the test results (H=0 indicates that the null hypothesis ("mean is M") cannot be rejected at the 5% significance level, H=1 indicates that the null hypothesis can be rejected at the 5% level); ii) P: p-value associated with the data and iii) CI: confidence interval.

Mathematical Model Based on Input/Output Relationship

We model the **G**ene **R**egulatory **N**etwork (GRN) of the GAL system using i) seven genes, ii) input to the system and iii) output of the system. The GRN is modeled using a black box approach. A given input will produce a specific output and based on the input/output relationship the regulatory elements are studied and analyzed. Input to the system is modeled by external galactose and output of the system is a reporter gene. The model does not assume any prior knowledge. The change in concentration of gene regulatory elements over time depends on their regulation by other regulatory elements, their decay rates and, the gene copy number (see equation 1). This model is based on a recent model used (Perkins, T.J. *et al.* 2006) to determine the structure and dynamics of genetic networks based on expression data of early *Drosophila melanogaster* development.

$$\frac{dX_i^a}{dt} = \underbrace{N_a}_{\text{Gene copy number}} \underbrace{R_a}_{\text{Maximum synthesis rate}} g\Big(\underbrace{\sum_{b=1}^{M} T_{ab} X_i^a + h_a}_{\text{Effect of regulatory elements}}\Big) - \underbrace{\lambda_a X_i^a}_{\text{Decay rate}}$$

where...

$$g(u_a) = \frac{1}{2}\left(\frac{u_a}{\sqrt{(u_a)^2 + 1}} + 1\right)$$

Equation 1: T_{ab} represents regulatory coefficient characterizing the regulatory effect of the product of gene b on the expression of gene a. N_a represents the gene copy number of gene a ($N_a \in \{0,1,2\}$). R_a is the maximum synthesis rate of gene a. h_a represents the basal level and λ_a indicates the decay rate of gene a. $g(u_a)$ is a sigmoid regulation-expression function where u_a is the total regulatory input on gene a.

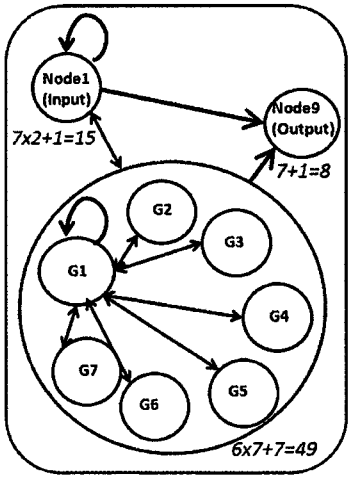
Our modeling framework for the Galactose Utilization Pathway incorporates gene copy number to accommodate various genotypes: a multiplicative factor (N_a) is incorporated to the equations. N_a represents the gene copy number in diploid yeast cells. The ODEs representing all possibilities for the seven-gene network consists of a T matrix (72 parameters) characterizing effect of a gene on any other gene. The regulatory coefficients in the T matrix will give indication of regulatory functions of the genes. The optimized set of coefficients from the T matrix will indicate positive/negative or neutral interactions among the genes in the network. The model is generic and therefore the focus is toward the relative sign of the connectivity values rather than exact values in order to gain insight about functional relationship among regulatory elements in the network. To build the Gene Regulatory Network, the decay rate of the input signal (galactose) is modeled as an indirect measure of the internal galactose concentration. High decay rate is equivalent to low galactose concentration and vice versa. Decay rates are represented by λs (decay rate of galactose is $\lambda 1$); thus for high expressing cells, high galactose is required which translates to low decay rate. This simplification is required in order to keep the equations generic. Table 3 describes the set of ODEs for a seven gene network as well a graphical representation of the network, where connections are obtained from the coefficients of the T matrix.

Seven GENE NETWORK

EQ1: GALACTOSE(INPUT),
EQ2-8: GENE 1 TO GENE 7,
EQ9: REPORTER(OUTPUT)

99 parameters

```
dxdt=[ N(1)*(R1*0.5*((T11*x(1)+T12*x(2)+T13*x(3) +T14*x(4) +T15*x(5) +T16*x(6)
+T17*x(7) +T18*x(8)-B1)/sqrt((T11*x(1)+T12*x(2)+T13*x(3) +T14*x(4) +T15*x(5)
+T16*x(6) +T17*x(7) +T18*x(8)-B1)^2+1)+1)-lamda1*x(1));

N(2)*(R2*0.5*((T21*x(1)+T22*x(2)+T23*x(3) +T24*x(4) +T25*x(5) +T26*x(6)
+T27*x(7) +T28*x(8)-B2)/sqrt((T21*x(1)+T22*x(2)+T23*x(3) +T24*x(4) +T25*x(5)
+T26*x(6) +T27*x(7) +T28*x(8)-B2)^2+1)+1)-lamda2*x(2));

N(3)*(R3*0.5*((T31*x(1)+T32*x(2)+T33*x(3) +T34*x(4) +T35*x(5) +T36*x(6)
+T37*x(7) +T38*x(8)-B3)/sqrt((T31*x(1)+T32*x(2)+T33*x(3) +T34*x(4) +T35*x(5)
+T36*x(6) +T37*x(7) +T38*x(8)-B3)^2+1)+1)-lamda3*x(3));

N(4)*(R4*0.5*((T41*x(1)+T42*x(2)+T43*x(3) +T44*x(4) +T45*x(5) +T46*x(6)
+T47*x(7) +T48*x(8)-B4)/sqrt((T41*x(1)+T42*x(2)+T43*x(3) +T44*x(4) +T45*x(5)
+T46*x(6) +T47*x(7) +T48*x(8)-B4)^2+1)+1)-lamda4*x(4));

N(5)*(R5*0.5*((T51*x(1)+T52*x(2)+T53*x(3) +T54*x(4) +T55*x(5) +T56*x(6)
+T57*x(7) +T58*x(8)-B5)/sqrt((T51*x(1)+T52*x(2)+T53*x(3) +T54*x(4) +T55*x(5)
+T56*x(6) +T57*x(7) +T58*x(8)-B5)^2+1)+1)-lamda5*x(5));

N(6)*(R6*0.5*((T61*x(1)+T62*x(2)+T63*x(3) +T64*x(4) +T65*x(5) +T66*x(6)
+T67*x(7) +T68*x(8)-B6)/sqrt((T61*x(1)+T62*x(2)+T63*x(3) +T64*x(4) +T65*x(5)
+T66*x(6) +T67*x(7) +T68*x(8)-B6)^2+1)+1)-lamda6*x(6));

N(7)*(R7*0.5*((T71*x(1)+T72*x(2)+T73*x(3) +T74*x(4) +T75*x(5) +T76*x(6)
+T77*x(7) +T78*x(8)-B7)/sqrt((T71*x(1)+T72*x(2)+T73*x(3) +T74*x(4) +T75*x(5)
+T76*x(6) +T77*x(7) +T78*x(8)-B7)^2+1)+1)-lamda7*x(7));

N(8)*(R8*0.5*((T81*x(1)+T82*x(2)+T83*x(3) +T84*x(4) +T85*x(5) +T86*x(6)
+T87*x(7) +T88*x(8)-B8)/sqrt((T81*x(1)+T82*x(2)+T83*x(3) +T84*x(4) +T85*x(5)
+T86*x(6) +T87*x(7) +T88*x(8)-B8)^2+1)+1)-lamda8*x(8));

N(9)*(R9*0.5*((T91*x(1)+T92*x(2)+T93*x(3) +T94*x(4) +T95*x(5) +T96*x(6)
+T97*x(7) +T98*x(8)-B9)/sqrt((T91*x(1)+T92*x(2)+T93*x(3) +T94*x(4) +T95*x(5)
+T96*x(6) +T97*x(7) +T98*x(8)-B9)^2+1)+1)-lamda9*x(9));];
```



A seven gene network has:

i) 7 genes (G1 to G7),
ii) input to the system
iii) output of the system

Total number of parameters (complete connectivity):

T_coeff: interaction among the seven genes +
auto-regulation +
interaction between the genes and the input +
auto-regulatory effect of the input +
effect of the seven genes on output +
effect of the input on output

T_coeff: (6x7)+(7)+(7x2)+(1)+(7)+(1)=72
Other param: $R_a + h_a + \lambda_a = 9+9+9=27$
Parameter Space = 72 + 27 = **99**

Table 3: (Top panel) Set of ODEs for a seven gene network. (Bottom panel) Graphical representation of a seven gene network.

This modeling framework eases the comparison between the experimental data and the quantitative ODEs. The experimental method employs combinatorial reduction in gene copy number and the modeling framework allows a systematic alteration of the equations to allow such modifications. In fact a combinatorial array is generated and depending which experimental result is being used for fitting purposes the correct set of N parameters are selected. For instance for the wild-type strain, N(1) and N(9) are set to one and N(2) to N(8) are set to two. There is one copy of the input and one copy of the output, and two copies of all the other seven genes in the network.

Optimization Algorithm

Continuation Analysis

Continuation analysis is the process of finding a series of steady-state solutions to ODEs starting from an initial steady state and varying the continuation (also known as bifurcation) parameter. Here the dose response curve is generated by finding steady-state series for a range of galactose decay rates, representative of galactose concentrations. The galactose concentration is increased to its highest value (111 mM or 2%) and decreased to its minimum value (~0); along this path it will be possible to have one or two solutions for each galactose concentration. There are two main methods used for steady-state evaluation: Newton iteration and ODE solvers in Matlab (specifically ODE45 and ODE15s for stiff equations). The approximation by Newton iteration is faster, and therefore is attempted first. If steady-state cannot be found by the latter, then a Matlab ODE solver is called. The Matlab code for continuation analysis has been kindly provided to me by Dawn Fraser.

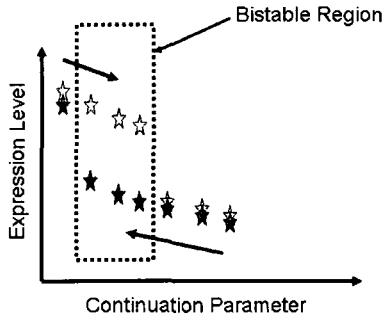


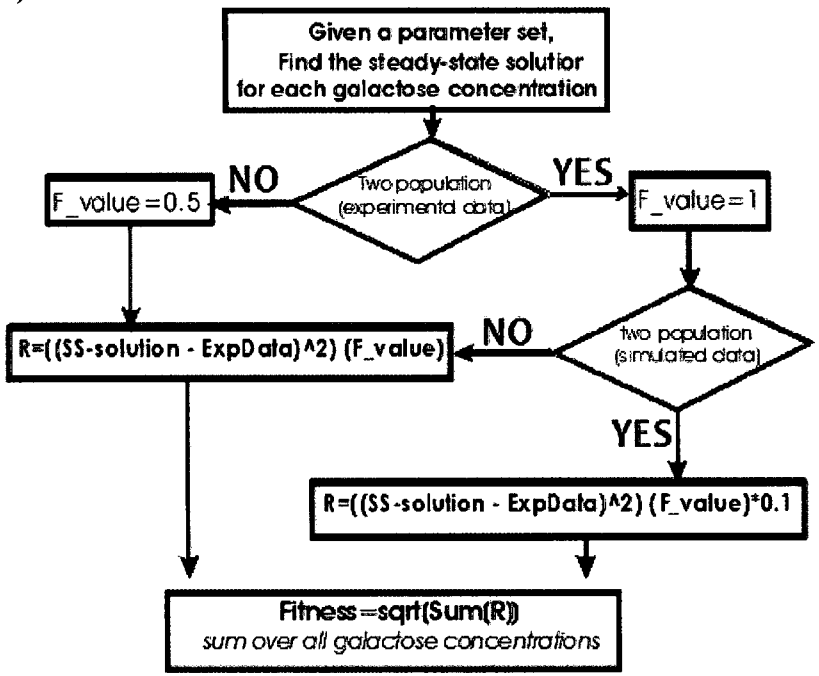
Figure 14: Continuation Analysis. Continuation parameter is represented by inverse of galactose decay rate. Bistability is observed if continuation analysis results in two different dose response curves.

Fitness Function

The main objective of a fitness function is to evaluate the difference between the experimental data (the best solution) and the simulated results. Once this difference is negligible or cannot be reduced further, one could assume the best possible solution is found. This function could be slightly manipulated to give more weight to a particular situation or scenario; for instance the solution to the ODEs has to be bimodal, therefore, if bimodality is found the algorithm should be able to retain the parameter sets producing the bimodal response in the population. The fitness function in the Genetic Algorithm (GA) is implemented as represented by the following flow chart (figure 15). Figure 15 outlines the three different scenarios encountered.

Figure 15: A) Evaluation of fitness score for a genetic algorithm. For each parameter set the expression level for the reporter protein is calculated; this simulated data is compared with the experimental data. The comparison is shown by this flow chart. The idea is to compare the simulated data and experimental data in a way to give more weight to situations where bimodal distributions are observed in the data and detected by the model. Fitness scores are used to rank and select the best parameter sets. **B)** The three different scenarios resulting from fitness score evaluation.

A)



B)

	CASE I
Exp data→	2 populations
Sim data→	2 populations
Weight:	0.1

	CASE II
Exp data→	2 populations
Sim data→	1 population
Weight:	1.0

	CASE III
Exp data→	1 population
Sim data→	1 or 2 populations
Weight:	0.5

SS-solution: steady-state solution
Exp data: experimental data;
Sim data: simulated data.

CASE I: This scenario is optimum. It is important to include this parameter set for the next generation at the end of the inner loop; its fitness score is therefore decreased (by a multiplicative factor of 0.1).
CASE II: This scenario is the most common. The simulated data is unable to produce a bistable response. However, the experimental data is bimodal. Therefore, fitness score will not be modified.
CASE III: If the experimental data is not bistable, regardless of the simulated data, the fitness score will be decreased by 50%. The reason for that is the fact that this situation is not very common (most of the dose response curves are bimodal) and indeed if the simulated data can be bistable, we would therefore prefer to keep the parameter set because it is very likely that the simulation data for other dose response curves are also bistable. In such cases the experimental data is also likely to be bistable (note that only some of the KO strains have lost their bistability).

Figure 15

Heuristic Algorithms

Hill-climbing

Hill-climbing was the first optimization algorithm used for searching the parameter space. The algorithm is outlined in figure 16. The parameter space was large and an improved methodology was required. Genetic Algorithm (discussed below) was then implemented to tackle the optimization problem more efficiently. The hill-climbing was simple to implement and converged fast, yet due to problems of finding local minima instead of global minima; genetic algorithm was further developed and refined.

Figure 16: Flow chart describing the hill climbing algorithm. One parameter at a time is selected for optimization. The order of parameters is randomized; however, each parameter can only be optimized once per round (outer loop). In the first section randperm function returns a random permutation of the integers 1 to (total_number_of_parameters-1); note that last parameter is the continuation parameter and does not require optimization. In the inner loop, each parameter is changed by some random percentage: abs(current_value(1+std(dir)(abs(randn(1))))); std is the standard deviation set arbitrary at the beginning, dir is the direction of change (+1 or -1). If fitness is reduced, the new parameter set replaces the previous parameter set, otherwise direction is reversed and fitness is calculated again. If fitness is not reduced for a second time then next parameter in the array will be analyzed (outer loop).

FORLOOP: 1 to Maximum_Iteration (i.e. 100)

Create an array to hold a list of parameters to select from for optimization: randperm(nb_param-1).

FORLOOP: 1 to nb_param-1

Select parameter to change

FORLOOP: 1 to n_steps (initialized at the beginning, ex: 5)

Change parameter value:
New_value = abs(current_value x (1 + std x dir x abs(randn(1))))

IF new_value ==0
New_value = 1e-07
– In the Aitchison's model, all parameters are >0

Continuation Analysis to find steady-state solution for all galactose concentrations

Calculate Residuals (difference between simulation and experimental results)
IF low peak and high peak are the same (at most 10% different)
THEN residuals is calculated once using average of the data and simulated result
ELSE residuals are calculated for low and high peak separately (therefore twice)
– this will give more weight to the bimodal portion of the data (a more sophisticated procedure is developed for calculating the fitness of genetic algorithm)

Calculated Fitness based on all residuals

IF Fitness < minimum_fitness
Minimum_fitness = Fitness AND par_default = new_parameters
ELSE keep original parameters AND dir=dir x (-1)
IF direction was changed twice for the same parameter, go to the next parameter

Genetic Algorithm

The implementation of Genetic Algorithm (GA) used for optimization of the parameter-set is given as pseudo-code in the following section. The method used is based on both mutation and crossover. The mutation and crossover rates could be adjusted depending on the search and optimization problem. The mutation rate was randomly generated at each round. The crossover rate was set at the beginning of the code and remained fixed throughout the run. A detailed flow chart describes the algorithms in the following figure (see figure 17A and figure 18). The inner loop and outer loop are in place for the GA to find the best possible solution. Once the termination condition is satisfied (see figure 17B) in the inner loop, the GA will restart and make sure the solution found is the best possible solution (see figure 18). The number of maximum iteration is significantly greater for the inner loop than the outer loop.

To test if the above algorithm would be able to find the best solution, a convergence test was performed. The idea is to evaluate the output for a given set of parameters and starting from the initial guess, find a set of parameters that would reproduce that output. It is clear that the parameter set could be different but still produce the same output; however, it is important to see if the desired output is retrievable and the algorithm will eventually find it without getting into sup-optimal solutions. The test for convergence was successful (data not shown).

Figure 17: A) Pseudo-code for the implementation of GA. Number of parents for the crossover is set by default to two: however there is potential for further generalization as this number could be changed from two to any number less or equal to the number of participating parents. **B)** Condition for termination implemented in the inner loop of the GA. The threshold of 0.00005 (0.005%) is chosen arbitrary. This is to assure that the algorithm is progressing and is not trapped in a local minima.

A)

STEP I: INITIALIZATION

Calculate the fitness of the initial parameter set and set this value to Max_Residual

This step is crucial to evaluate the convergence. The termination condition will use this value to evaluate the convergence criterion.

Creating of array of initial parents (P) based on a fairly good guess of parameter-set (b)

Create an array of size 2bxb where each row will differ from b by only one value: section one of the array (from row 1 to b) will contain values slightly higher than b, and section two of the array (from b+1 to 2b) will contain values slightly lower than b. Differences from b are positioned at diagonals of the square arrays (section one and section two).

Calculate the fitness (Fp) value for each parent

Figure 15 explains the fitness function in detail.

Select the best 'number' of fitnesses and their equivalent parameter-sets

Number is given as an input and could be easily varied.

WHILE (TERMINATION CONDITION is not satisfied) ***Repeat***

Termination condition is defined as I) if all the parents selected in Step IV have the same fitness value as the Max_Residual; II) a maximum number of iterations is executed.

Calculate the fitness of best parents and set this value to Max_Residual

This step is crucial to evaluate the convergence. The termination condition will use this value to evaluate the convergence criterion.

STEP II: CROSSOVER

Create a mask array (M) that will be used to generate offsprings (OF)

M is a binary array of size (number x N); Number of columns will correspond to the number of selected parents (i.e. 'number') and number of rows will be equal to the number of offsprings ('N') that will be generated by the crossover. N is given as an input and could be easily varied. Each row contains two 'ones' and the index of the latter will dictate which parents (P) will be used to create the offspring. Number of ones could be increased for creating offsprings from more than two parents at a time.

Calculate the fitness of offsprings (Fof)

STEP III: MUTATION

Select a parameter from each offspring and use the standard deviation given as input to change that parameter by increasing it and decreasing it and creating two children (C)

For each offspring there are two children created, using the admissible change in both directions. This is to introduce reasonable level of population diversity.

Calculate the fitness of children (Fc)

STEP IV: SELECTION

Combine the fitness of initial parental strain (Max_Residual) parents (Fp), offsprings (Fof) and children (Fc)

Select the best 'q' fitnesses: (Fp)

'q' is given as an input and could be easily varied.

Get parameter set equivalent to the best 'q' fitnesses: (P)

After finding the best fitnesses, using the indexes it will be possible to get the corresponding set of parameters that will be used as the parents for the next generation.

END OF WHILE

RETURN the parameter set (P) as the final result

B) Termination condition:

IF $(\sum^{q}_{i=1}$ (Fitness(i) - Max_Residual)/Max_Residual)x100 < 0.00005

THEN Termination_flag=1

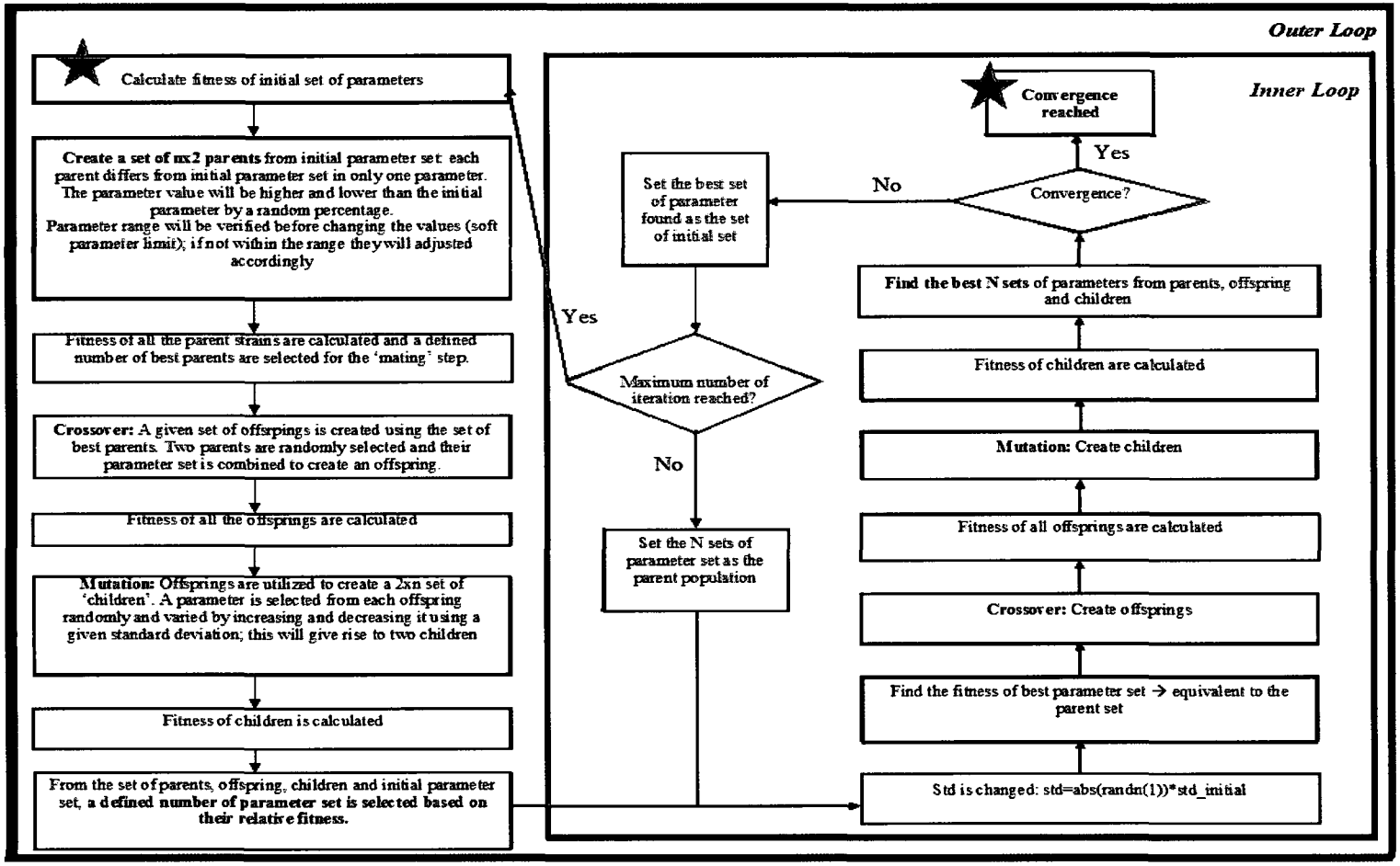


Figure 18: Flow chart describing the GA. The blue star indicates the start and the pink star indicates the end of one run.

Post-optimization

Parameter optimization, as described in detail, will identify sets of parameters that would reproduce the data; a post-optimization is on the other hand, the process of reducing the number of interaction between genes and still producing a good fit to the data. This process will reduce (in some cases) the number of interactions and therefore keep the system as sparse as possible. The threshold is set to 5% (threshold is a measure of acceptable loss in the fitness score). The post-optimization step is important because in the optimization procedure, the number of connectivity among genes had no upper limit. Post-optimization might not be very critical for small networks; however, as the number of genes increases, over-parameterization becomes a significant barrier toward identification of the true network topology. The following algorithm (figure 19) outlines the refinement process used following parameter optimization.

FOR LOOP: 1to Maximum_Iteration (i.e. 10)

Find Original Fitness based on initial parameter set
Create an array to hold a list of parameters to select from for refinement: function randperm

FOR LOOP: 1 to size of T matrix

Select parameter to change from the randomly generated list array

Change parameter value:
New_value = (current_value) x (-1)
Find temp_fitness from the new parameter set
Calculate TH=(temp_fitness-Original_fitness) / temp_fitness

IF TH >0 and TH > threashold
    Keep the parameter's original value
ELSEIF abs(TH)>2xthreashold && TH<0 %fitness was significantly improved! Keep the change
    Optimize new parameter set (using GA*)
ELSE Set the parameter to zero % fitness increase was within the acceptable range
    Optimize new parameter set (using GA*)

The New parameter set is saved for the next round

GA: slightly modified GA. Parameters that are zero will remain zero at all time.*

Figure 19: Post-optimization algorithm. It is important to select parameters randomly at the beginning of the algorithm. This is due to the fact that some of the parameters could be set to 1 (in order to reduce the parameter space), and this simplification causes other parameters to change. Therefore a large T-coefficient will not necessarily correspond to strong binding between two species, and vice-versa.

Generating and Ranking Topologies Based on Their Fitness Score

The model of the gene regulatory network has a set of parameters which define the topology of a network structure. Definition of network topology arises from values of the connectivity matrix: a positive value indicates an activating signal and a negative value is representative of repression. To obtain a set of parameters, parameter optimization is run using experimental data from gene knock-down strains. The latter will focus on relationships among genes without changing the network topology. The accepted parameter set will produce the bimodal response that is observed in the experimental data. This requirement will reduce the number of potential candidates significantly. The reduced set of candidate topologies will then be ranked based on their potential to reproduce the experimental data for gene knock-out strains. This stringent requirement will further reduce the number of successful topologies where interesting hypotheses could be generated toward a better understanding of functional relationship among regulatory elements in the network.

Reproducing experimental data for the knock-out strains will provide stronger evidence and will be of higher weight, because strains carrying gene knock-down do not have necessarily the same reduction in the active protein level. Furthermore, this reduction could be significantly more or less than the assumed 50%.

Chapter 3: Results

Identification of Regulatory Elements of the GAL Network

Fitness profiling is a systematic approach for identification of genetic interactions. It is a simple method that relies on growth phenotype. However, if a gene is important for the dynamic behaviour of a system but does not affect the growth, then fitness profiling will not be able to identify it. On the other hand, in most cases, knocking out essential genes will affect the growth.

Expression profiling by flow Cytometry, can be utilized to measure dynamical stability of a system. By capturing the system's behavior at single cell resolution, it is possible to identify key regulatory elements. These elements could have direct or indirect, and activating or repressing roles.

GAL2, GAL3, GAL4 and GAL7 are identified as key regulatory genes of the GAL network by fitness profiling experiment. A complete library of perturbed strains was first generated using pair-wise gene deletion strategy. Pair-wise gene deletion strategy provides a solid framework for systematic measurement of relative fitness scores of the perturbed strains, leading to identification of key players in a given biological network. Galactose utilization network was analyzed through systematic fitness profiling, and GAL2, GAL3, GAL4 and GAL7 were identified as key elements in the network. Figure 20 and 21 demonstrate this fact. The functional roles of Gal2p, Gal3p and Gal4p

are well established. Yet, the regulatory function of Gal7p in the galactose utilization pathway is still unknown. Our observation emphasizes the importance of GAL7 in the network and provides further evidence (Stagoj, Comino and Komel 2005) for its regulatory role of limiting growth of the cell population. In fact, it is well established that an accumulation of galactose-1-phosphate is toxic and cells lacking Gal7p are unable to transform this by-product to glucose-1- phosphate. Therefore a knock-out strain has a reduced fitness.

Being a coarse grain approach, fitness profiling can lead to discovery of some of the genetic interactions in a regulatory network. However, in some cases, knocking out important genes do not cause reduced growth. In those instances, fitness profiling will not succeed in identifying key elements. For instance, it is well known that GAL80 is a major player in the galactose utilization pathway; however, fitness profiling did not identify the latter as one of the central elements. Therefore, we will use expression profiling by flow Cytometry to complement the fitness measure.

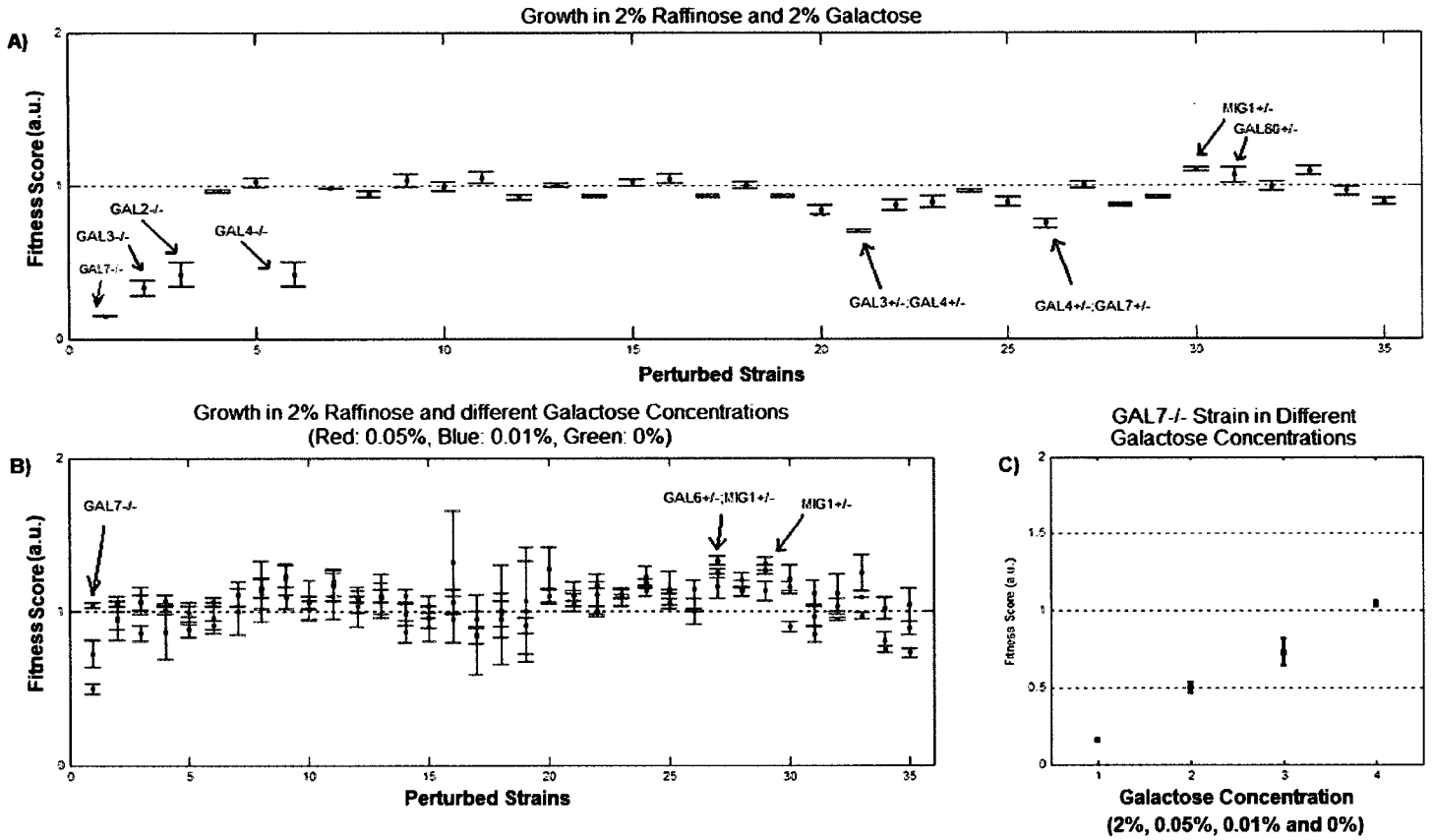


Figure 20: Fitness score of perturbed strains in different conditions. **A)** Media contains 2% raffinose and 2% galactose (full induction); **B)** media contains 2% raffinose and the indicated amount of galactose. **C)** Fitness score of GAL7-/- strain; fitness is inversely proportional to the induction level. Error bars represent standard deviation (n=3).

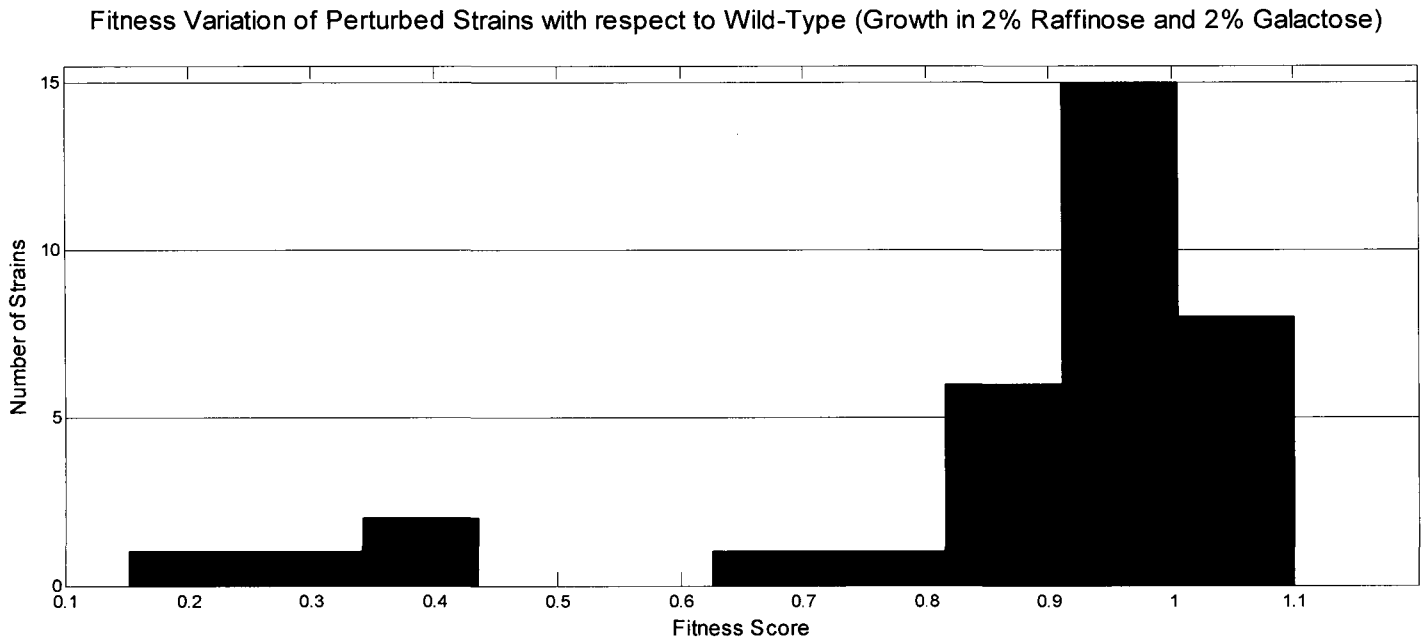


Figure 21: Fitness ratio for 35 strains with respect to wild-type. The four outliers with fitness score of less than 0.4 are GAL2-/-, GAL3-/-, GAL4-/- and GAL7-/- strains.

Expression profiling identifies GAL80, GAL4, GAL2 and GAL3 as key elements of the GAL network. Expression profiling by flow Cytometry led to a clear identification of GAL2, GAL3, GAL4 and GAL80 as key players in the network. Knocking-out GAL7 causes also the expression level to be slightly higher than the wild-type. The next three figures (figure 22, 23 and 24) summarize these findings.

To systematically compare different perturbed strains to the wild-type, over a range of different induction levels, an integration method was developed (see material and methods). This analysis identifies GAL2, GAL3, GAL4 and GAL80 as key elements in the network.

Finally, it is important to use parallel methods for identification of primary regulatory elements in a network. The overlap between different identification methods will further confirm the validity of the findings and their importance for the system-identification method. For instance, expression profiling did identify GAL80 but not GAL7 as a central element in the network. GAL80 is a repressor of the system and therefore could not be easily identified by growth phenotypic assays. On the other hand, GAL7 could function through other indirect means; its deletion was shown to be detrimental for growth, but not for the dynamical stability of the GAL system.

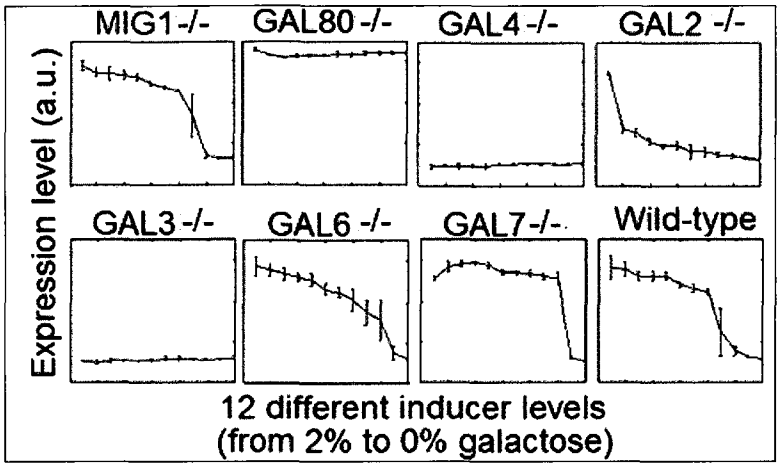


Figure 22: Average expression level of the reporter gene (P_{GAL1}-GFP) at steady-state following induction by galactose at different concentrations (condition 1 to 12: 2, 0.03, 0.02, 0.015, 0.01, 0.008, 0.007, 0.0055, 0.004, 0.0037, 0.003 and 0% galactose respectively). Error bar represents standard deviation (n=3).

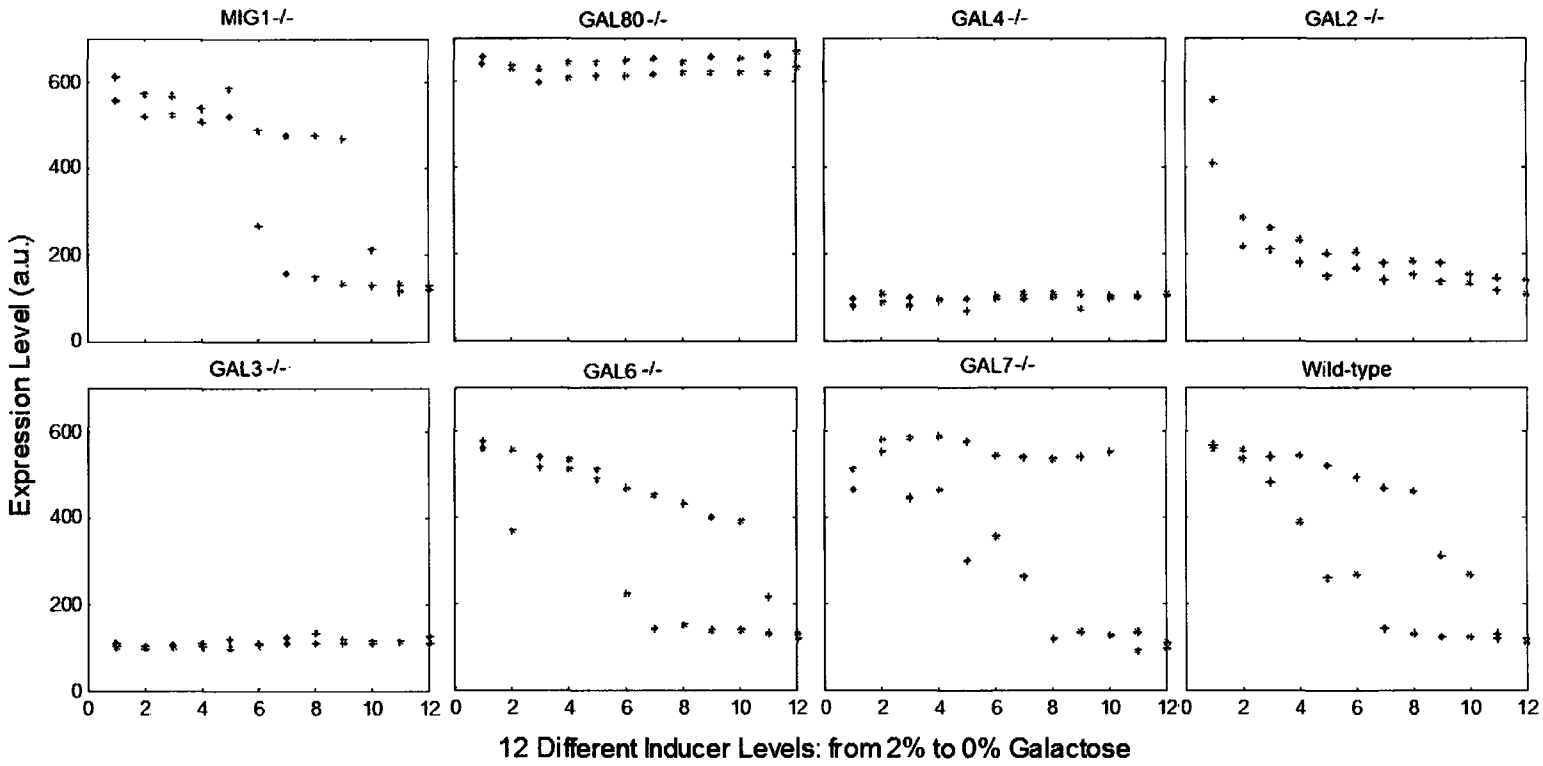


Figure 23: Expression level of the reporter gene (P_{GAL1}-GFP) at steady-state following induction by galactose at different concentrations (condition 1 to 12: 2, 0.03, 0.02, 0.015, 0.01, 0.008, 0.007, 0.0055, 0.004, 0.0037, 0.003 and 0% galactose respectively). Blue represents expressing cells (ON state) and red represents non-expressing cells (OFF state). When expression level is not at its maximum or at its minimum there are usually two populations present, each having their respective mean expression values. The range of inducer level to produce bimodality is different for different strains. In some cases, bimodality is lost due to gene knock-out (GAL2-/-, GAL3-/-, GAL4-/- or GAL80-/-).

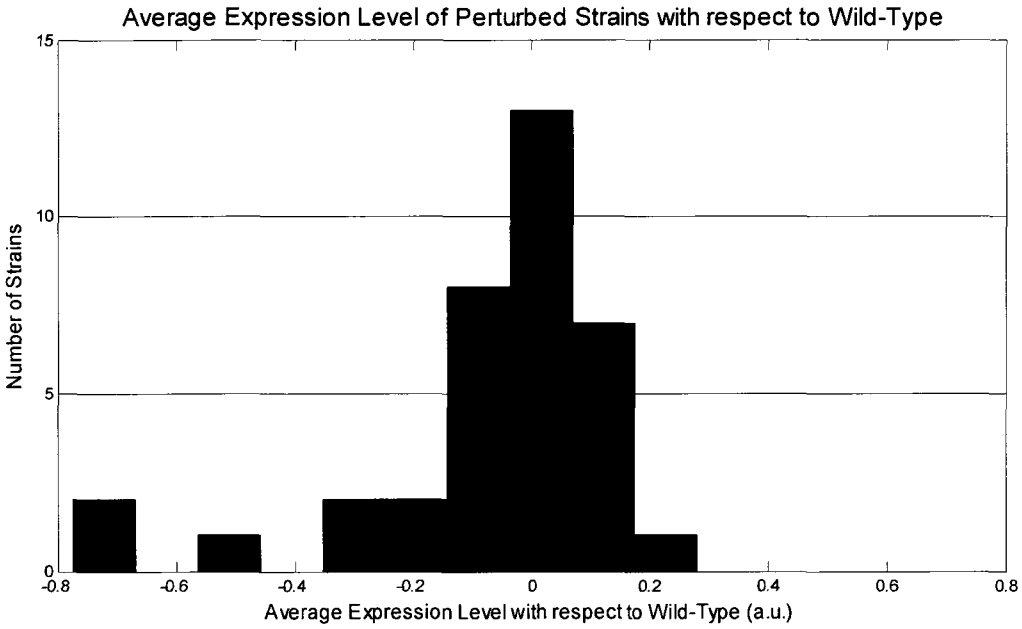


Figure 24: Average expression level of perturbed strains with respect to wild-type strain. The four outliers that result from this analysis are GAL2-/-, GAL3-/-, GAL80-/- and GAL4-/- with a p-value < 0.06. GAL80-/- is the only strain exhibiting significantly higher expression level when compared to wild-type. Table 4 lists the strains with their respective p-values. Interestingly, GAL7-/- does not show a significant difference in dose response curve when compared to the wild-type (p-value < 0.3). Statistical test utilized to select the outliers is based on the ZTEST with the assumption of normal distribution (see Material and Methods for a detailed explanation).

Table 4: Strains along with their respective p-values. The four outliers are GAL2, GAL3, GAL4 and GAL80. The dose response curves of GAL6, GAL7 and MIG1 knock-out strains show no significant difference when compared to the wild-type. The system is robust to gene knock-down perturbations. Statistical test utilized to calculate the p-value is based on the ZTEST with the assumption of normal distribution (see Material and Methods for a detailed explanation).

Strain	P-value
GAL4-/-	**0.0011**
GAL3-/-	**0.0016**
GAL2-/-	**0.0196**
GAL80-/-	**0.0608**
GAL3+/-;GAL6+/-	0.1191
GAL3+/-;MIG1+/-	0.1388
GAL80+/-;GAL6+/-	0.1550
GAL80+/-;MIG1+/-	0.2148
GAL80+/-	0.2188
GAL80+/-;GAL4+/-	0.2295
GAL80+/-;GAL3+/-	0.2429
GAL80+/-;GAL2+/-	0.2466
GAL80+/-;GAL7+/-	0.2555
GAL3+/-;GAL4+/-	0.2705
GAL7-/-	**0.2789**
GAL7+/-	0.2791
GAL4+/-;GAL7+/-	0.2956
GAL6+/-	0.3124
GAL3+/-	0.3486
GAL7+/-;MIG1+/-	0.3505
GAL6+/-;MIG1+/-	0.3546
GAL6+/-;GAL7+/-	0.3598
GAL2+/-;GAL7+/-	0.3719
WILD-TYPE	0.3729
GAL3+/-;GAL7+/-	0.4258
GAL2+/-;GAL3+/-	0.4290
GAL2+/-;GAL4+/-	0.4712
GAL2+/-;MIG1+/-	0.5124
MIG1+/-	0.5262
GAL2+/-;GAL6+/-	0.5274
GAL4+/-;GAL6+/-	0.5510
GAL6-/-	**0.5520**
MIG1-/-	**0.5672**
GAL4+/-;MIG1+/-	0.5681
GAL2+/-	0.5937
GAL4+/-	0.5972

Table 4

Expression Profiling of Pair-Wise Gene Deletion Strains Provides High quality Quantitative Data

One of the challenges in reverse-engineering (RE) and system identification (SI) is the availability of high quality quantitative data from a well established biological network. The Galactose utilization pathway provides us with an opportunity to tackle this simple yet intriguing problem. To be useful, the dataset has to be quantitative, rich and systematically generated. SGPA framework provides this and other advantages: the data is easily generated; it is quantitative and is very rich in nature (*in vivo* single cell measurements).

For the purpose of this investigation, we will concentrate on single cell measurements of combinatorial pair-wise gene deletion strains and we will focus on the input/output relationship of the expressing and non-expressing cells rather than only focusing on population average values. The generated data is highly reproducible. Figure 25 shows the expression profile of the wild-type strain for 12 different galactose concentrations, along with the fitted Gaussian distribution curves. Expression profiles for all 36 strains are included in the supplementary material of this thesis report. Figure 26 provides a typical result of signature profiles from which dose-response curves are generated. Figure 27 represents the dose response curves of all the strains in the library. The average expression values are representative of the total cell population (including expressing and non-expressing cells). Figure 28 represents also the expression profiles

for all the strains; however, mean expression of two populations (cell in ON state and cell in OFF states) are analyzed separately. The process is automated; therefore in some instances the entire cell population is in the ON state but it is represented by two mean values. This situation, particularly common for extreme galactose concentrations, has been taken into account in the analysis. In fact, when two populations are within a 10% margin, they are considered as one.

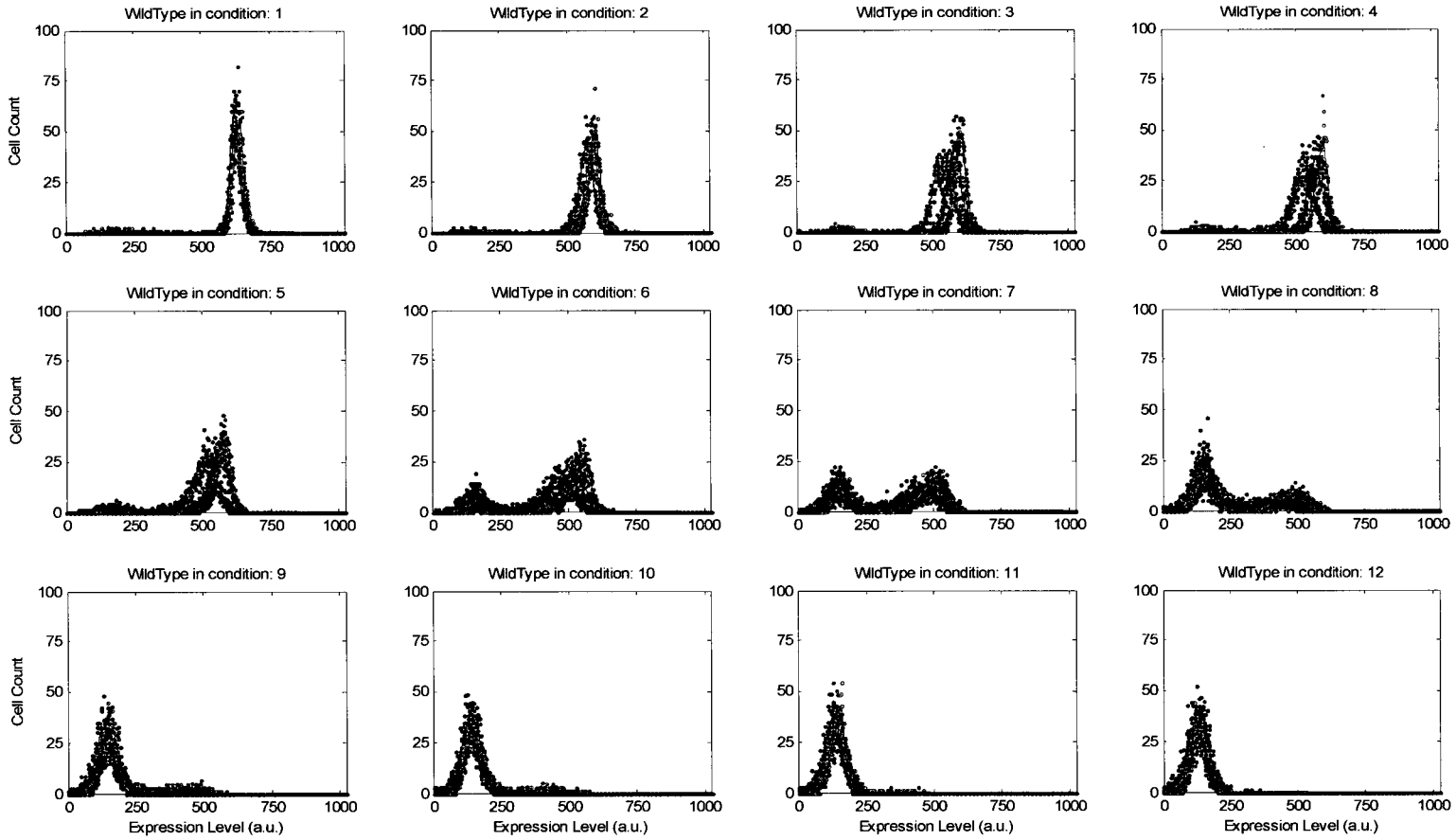


Figure 25: Flow Cytometry data, representing cell count versus expression level of the wild-type strain for 12 galactose concentrations (condition 1 to 12: 2, 0.03, 0.02, 0.015, 0.01, 0.008, 0.007, 0.0055, 0.004, 0.0037, 0.003 and 0% galactose respectively). Red/blue/green dots are three independent experimental data and solid lines are the respective Gaussian distributions used for further analysis.

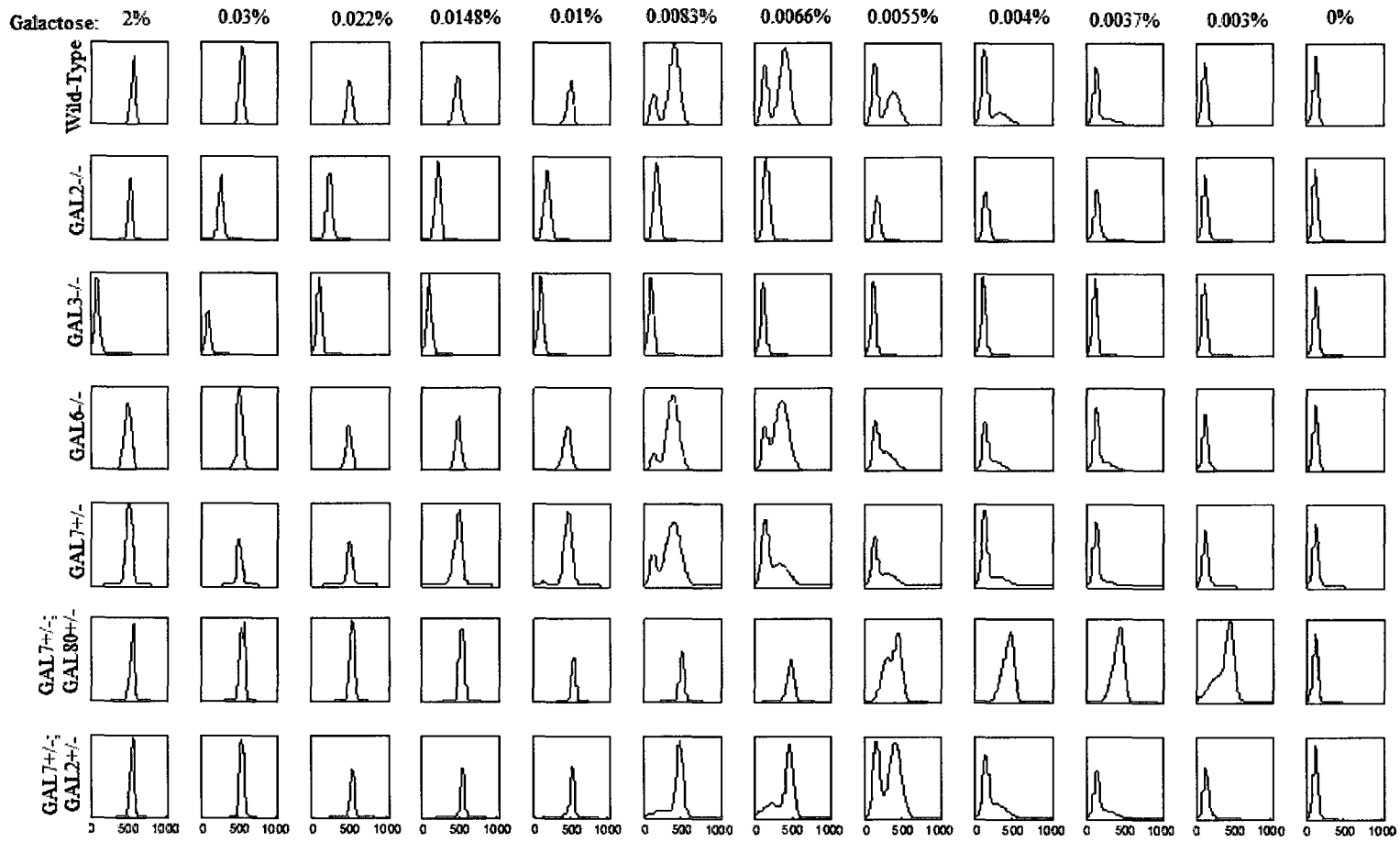


Figure 26: Normalized cell count versus expression level of wild type and six perturbed strains (1.wild-type, 2.GAL2-/-, 3.GAL3-/-, 4.GAL6-/-, 5.GAL7+/-, 6.GAL7+/-GAL80+/-, 7.GAL7+/-GAL2+/-) in 12 different galactose concentrations. These plots represent two Gaussian distributions used to fit the experimental data. The distributions are normalized to a p.d.f. of 1.0. The x-axis, representing the expression level, is logarithmically spaced (FL1-log).

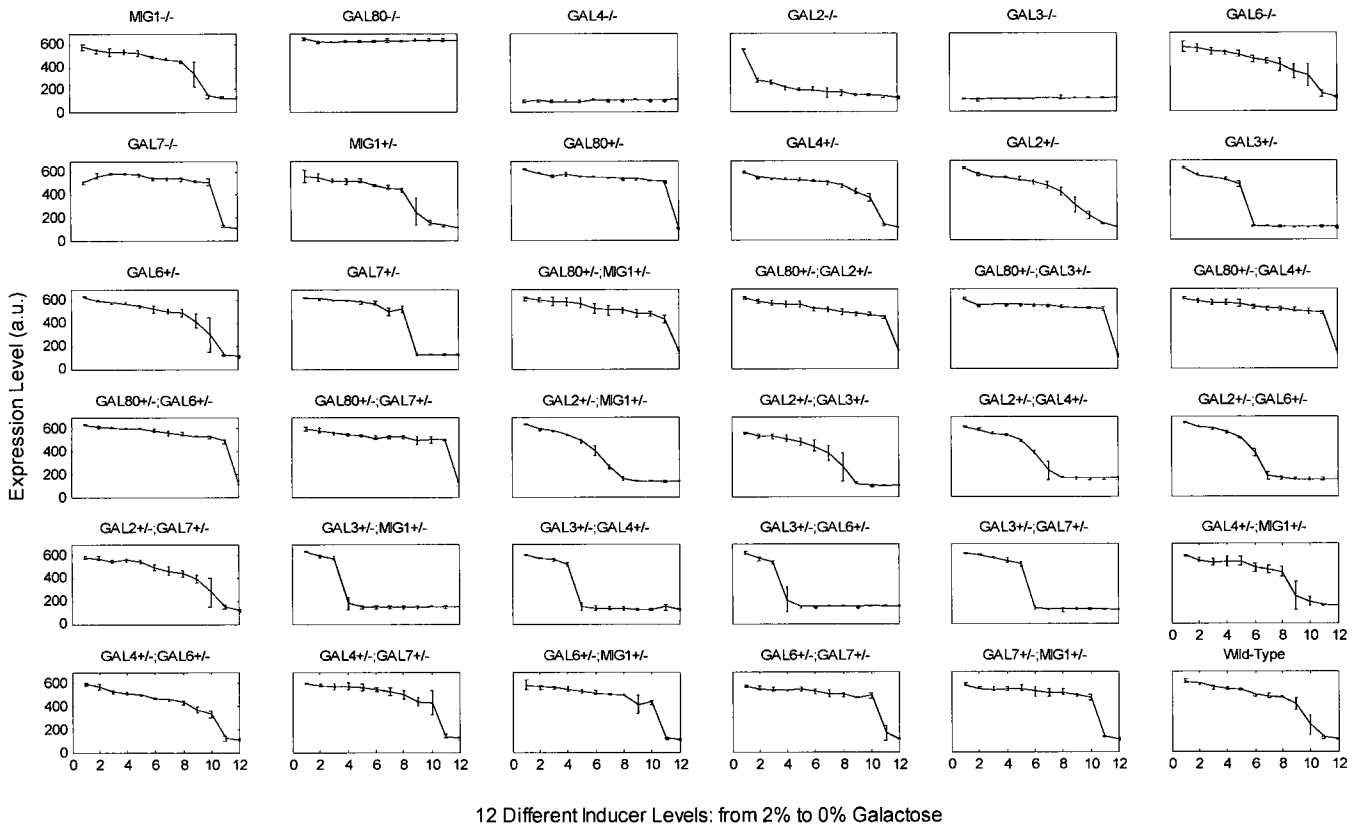


Figure 27: Average expression of the 36 strains in 12 different galactose concentrations (condition 1 to 12: 2, 0.03, 0.02, 0.015, 0.01, 0.008, 0.007, 0.0055, 0.004, 0.0037, 0.003 and 0% galactose respectively). Error bars are the standard deviations (n=3). Note that only GAL3-/-, GAL80-/- and GAL4-/- have an all or none expression level independent of the induction level. The latter characteristic will be of significant importance for the system identification process.

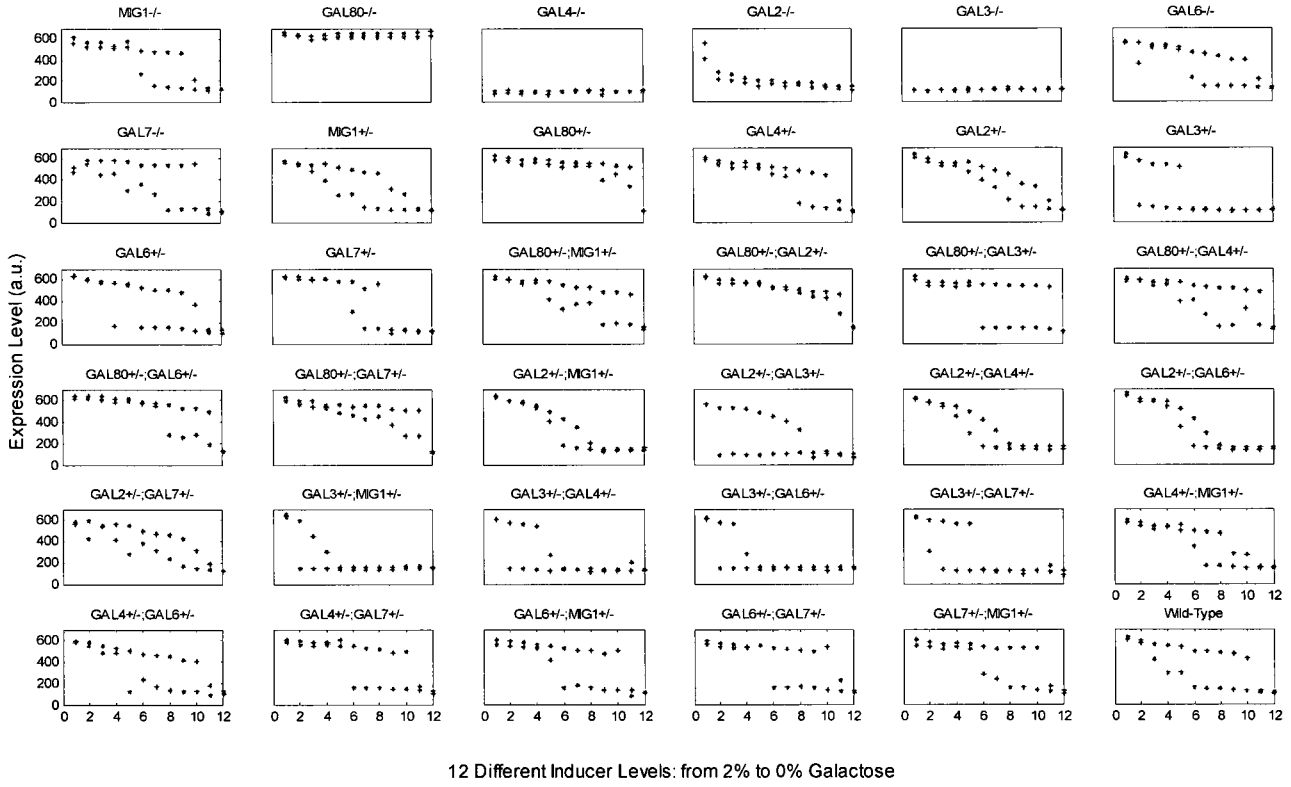


Figure 28: Average expression level for the two populations of each of the 36 strains in 12 different galactose concentrations (average values from three independent experiments). Blue: expressing cells (ON state), red: non-expression cells (OFF state). The 12 galactose concentrations on the x-axis are (from condition 1 to 12): 2, 0.03, 0.02, 0.015, 0.01, 0.008, 0.007, 0.0055, 0.004, 0.0037, 0.003 and 0% galactose respectively. Most of the strains exhibit bistability and have a characteristic expression signature.

A balance of GAL3 and GAL80 gene dosage is essential for dynamic stability of the GAL network. The SI method implemented here is based on pair-wise gene deletion strategy. The library of strains generated will be used to further investigate functional relationships among genes in the network. As a proof of concept, we evaluate a diploid strain in which a copy of GAL3 and a copy of GAL80 are both deleted (GAL3+/-;GAL80+/-). In fact, we will compare four independent expression profiles: 1) diploid wild-type strain, 2) diploid strain with one copy of GAL3 gene, 3) diploid strain with one copy of GAL80 gene and 4) diploid strain with one copy of both GAL3 and GAL80 genes. We show that SGPA can provide insight into functional relationships among regulatory elements. Figures 29 and 30, demonstrate this finding. Strains with one copy of GAL3 show lower expression levels when compared to the wild-type; the inducer level required to activate the network was higher with respect to wild-type. The strain with one copy of GAL80 has a higher expression level with respect to wild-type: the latter required very small amount of inducer to become active. The strain with one copy of both GAL3 and GAL80, however show no significant difference when compared to the wild-type. This observation provides evidence that GAL3 and GAL80 are complementing each other and a balance between the two genes is essential (Verma, Bhat and Venkatesh 2003) for the dynamical stability and robustness of the system.

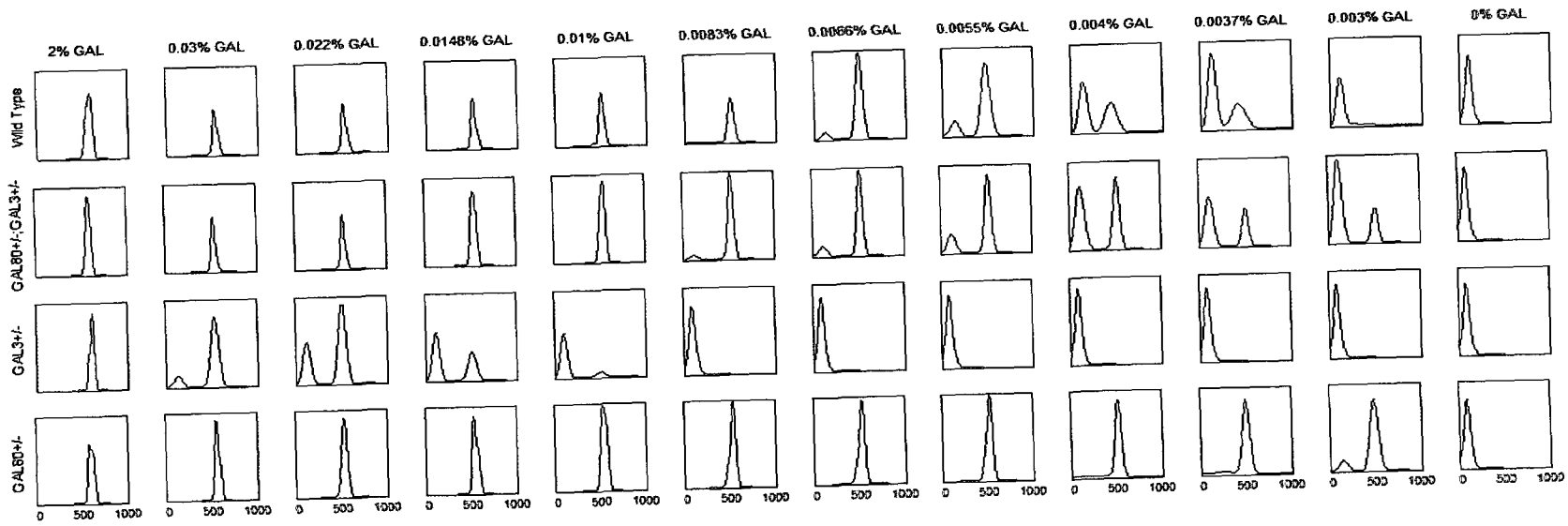


Figure 29: Normalized cell count versus expression level. Single cell fluorescence profile of 1) diploid wild-type, 2) GAL80+/- ;GAL3+/-, 3) GAL3+/- and 4) GAL80+/- strains. Wild-type and GAL80+/-;GAL3+/- strains reach full expression level at similar galactose concentration; GAL3+/- strain needs higher galactose concentration to reach its maximum expression level; GAL80+/- strain requires very small amount of galactose to reach maximum expression level.

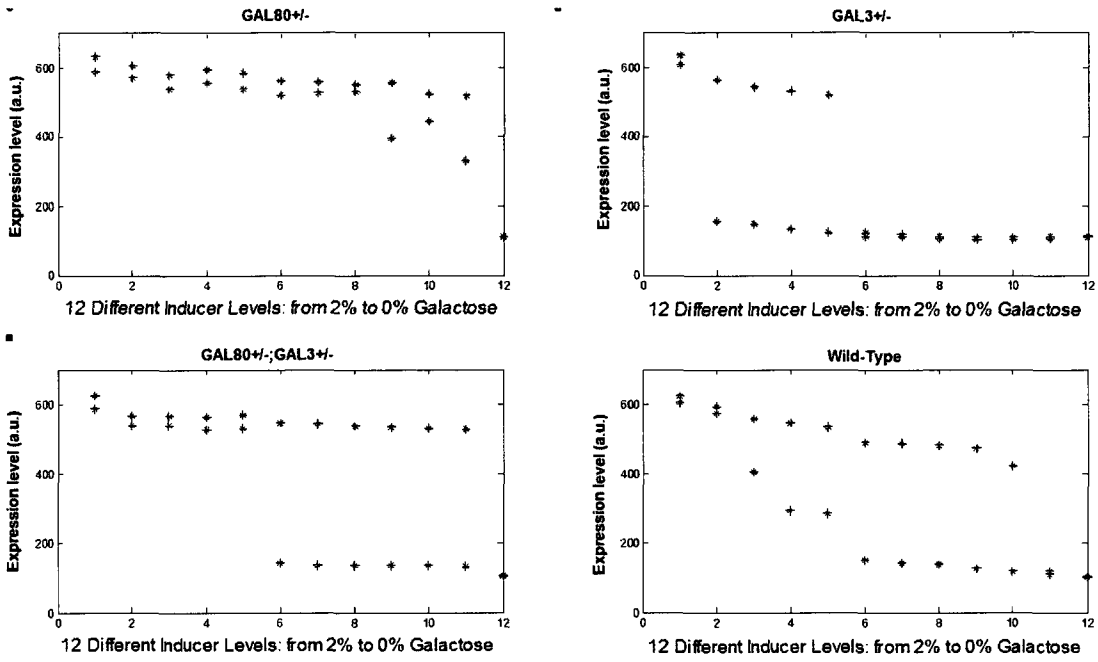


Figure 30: Expression level of the reporter gene at steady-state following induction by galactose at different concentrations (condition 1 to 12: 2, 0.03, 0.02, 0.015, 0.01, 0.008, 0.007, 0.0055, 0.004, 0.0037, 0.003 and 0% galactose respectively). The two populations of cells are analyzed separately: blue represents expressing cells (ON state) and red represents non-expressing cells (OFF state). GAL3 and GAL80 genes are acting in parallel to provide robustness and stability to the galactose utilization pathway. Gal3p is a positive regulatory element and Gal80p is a negative element in the network. A pair-wise gene deletion restores the system's dynamical stability with respect to wild-type strain.

System-Identification based on Genetic Algorithm

To develop and test a data-driven method for extraction of quantitative and mechanistic information, the canonical Galactose Utilization Pathway in yeast is used as a gold standard model system. The model developed and presented for the GAL network, uses i) seven genes, ii) input to the system and iii) output of the system, as described previously. The change of concentration of gene regulatory elements over time depends on their regulation by other regulatory elements, their decay rates and, the gene copy number. Therefore finding the set of parameters that describes regulatory interactions will elucidate functional relationship among the elements in the network. Incorporating gene copy number in the ODEs provides additional means to fit different genotypes into the same model. The task of parameter optimization for different strains simultaneously is therefore simplified. Being a generic model, the number of parameters increases drastically with each added node. It is therefore important to incorporate our knowledge based on the experimental data presented here. The identification of key elements provides a mean to restrict the number of nodes in the model to only those providing important information; this reduction in complexity and parameter space will decrease the computational time significantly. GAL2, GAL3, GAL4, GAL7 and GAL80 have been identified as the key elements in the regulatory network; however, only GAL3, GAL4 and GAL80 were found to be essential for dynamic stability of the system. GAL3, GAL4 and GAL80 are the most important genes and their full knock-out provides decisive

information (all or none) toward identification of the best candidate topologies. GAL7 does not provide additional information; its dose response curve is very similar to that of the wild-type. GAL2, even though important for the system, did not show an all or none expression level when completely knocked-out. Therefore, we will use three elements for the system-identification process: GAL3, GAL4 and GAL80. These three genes characterize the core regulatory elements of the network. The modeling framework will have three nodes, an input (Galactose), and an output (Reporter gene). To reduce the number the parameter sets even more, all λ values and R values are set to 1.0; this will decrease the computational time for parameter fitting purposes. However, because of this simplification, it will not be possible to predict strength of interactions among the genes based on the absolute parameter values.

The set of ODEs are presented in table 6. The parameters of the model are therefore optimized to fit the experimental data for the following genotypes: 1) GAL3+/-GAL4+/-, 2) GAL3+/-GAL80+/-, 3) GAL4+/-GAL80+/-, 4) GAL3+/-, 5) GAL4+/-, 6) GAL80+/- and 7) Wild-Type strain. Table 5, on the other hand, provides the genotype table utilized for parameter optimization process. This table is used to evaluate gene copy number for different genotypes – the values of N(1) to N(5) – when running the optimization algorithm.

The seven different genotypes used for fitting the data, will decrease the dimensionality of the system. In fact, the number of possible solutions will decrease

significantly by utilizing this strategy.

Genetic algorithm succeeds to reproduce the experimental data. Different topologies arise from the parameter optimization procedure (see table 7). These topologies were able to reproduce the bimodal response observed in the experimental data for the seven genotypes; however their respective p-value is based on their potential to reproduce the experimental data for the three knock-out strains. P-values are obtained by comparing fitness scores from the knock-out strains. The statistical test utilized to calculate the p-value is based on the ZTEST with the assumption of normal distribution. Indeed the fitness score, for each network structure, is compared against the fitness scores of the other 40 structures, from which it is possible to calculate mean, M, and standard deviation, STD (see Material and Methods for a detailed explanation).

Genotype ID	Input	Gene 1	Gene 2	Gene 3	Output	Genotype
1	1	1	1	2	1	Gene1+/-Gene2+/-
2	1	1	2	1	1	Gene1+/-Gene3+/-
3	1	1	2	2	1	Gene1+/-
4	1	2	1	1	1	Gene2+/-Gene3+/-
5	1	2	1	2	1	Gene2+/-
6	1	2	2	1	1	Gene3+/-
7	1	2	2	2	1	**Wild-Type**
8	1	0	2	2	1	Gene1-/-
9	1	2	0	2	1	Gene2-/-
10	1	2	2	0	1	Gene3-/-

Table 5: Genotype Table for a three-gene network. Input and output both have one copy; gene 1 to 3 could have 1, 2 or 0 copies. Last column indicates the genotypes of the strains. For instance, wild-type strain has one copy of pGAL1-GFP (output) and one copy of the input. It also has two copies of GAL3 (Gene1), GAL4 (Gene2) and GAL80 (Gene3).

<table>
<tr>
<td>Three GENE NETWORK

Reduction of parameter space: Input has one outgoing link and output has only one ingoing link. The three genes are fully connected. λ and R values are set to 1.
EQ1: GALACTOSE(INPUT),
EQ2-4: GENE 1 TO GENE 3,
EQ5: REPORTER(OUTPUT)</td>
<td>dxdt=[
N(1)*(1*0.5*((-B1)/sqrt((-B1)^2+1)+1)-lamda1*x(1));
N(2)*(1*0.5*((T21*x(1)+T23*x(3)+T24*x(4)-B2)/sqrt((T21*x(1)+T23*x(3)+T24*x(4)-B2)^2+1)+1)-1*x(2));
N(3)*(1*0.5*((T34*x(4)+T32*x(2)-B3)/sqrt((T34*x(4)+T32*x(2)-B3)^2+1)+1)-1*x(3));
N(4)*(1*0.5*((T42*x(2)+T43*x(3)-B4)/sqrt((T42*x(2)+T43*x(3)-B4)^2+1)+1)-1*x(4));
N(5)*(1*0.5*((T53*x(3)-B5)/sqrt((T53*x(3)-B5)^2+1)+1)-1*x(5));];</td>
</tr>
<tr>
<td>The three genes (nodes) in the network could be fully connected. Input and output are connected to only one node.

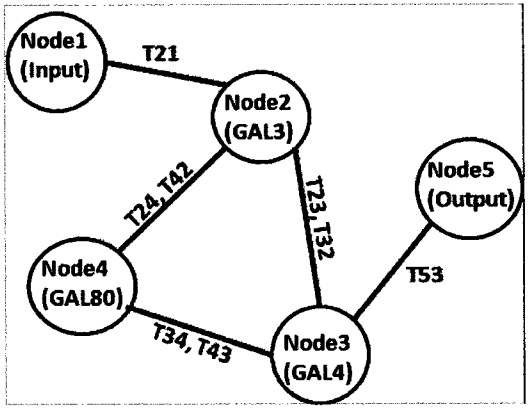
T_coefficients representing the network structure are indicated on each edge.
For instance, input→ node2 is identified by T21;
• if T21>0 then input activates node2,
• if T21<0 then input represses node2,
• if T21=0 then input has no effect on node2.</td>
<td> 
</td>
</tr>
</table>

Table 6: Set of ODEs for a three gene network. Bottom panel represents a graphical view of a three-gene network with one input and one output. Note that decay rates (λs) and maximum expression levels (Rs) are set to 1 in order to reduce the parameter space. Therefore, it will be incorrect to assume a correlation between strength of interaction and absolute values of the T coefficients. In other words, a high T coefficient does not necessarily mean a strong interaction or vice versa.

SGPA Succeeds to Reverse-Engineer a Small Gene Regulatory Network

Due to time constraint we were able to run the optimization algorithm for a total of 186 runs. Among those, 56 sets of parameters were found to be able to reproduce the bimodality of the data. The 56 sets of parameters were illustrative of 41 unique topologies (table 7). For a three gene network, this number is very large. However, by using the single cell expression data from full-knock out strains, it was possible to reduce the 41 topologies to only one, which is the best representation of the network.

The best topology (p-value <0.05) was able to produce bimodality for a given parameter set, and reproduce a relatively good fit to the data (for the seven genotypes) in addition of being consistent with knock-out strains. The network structure that arises from this topology is a very close representation of the GAL network. The second topology, with a p-value <0.1 does provide some insight into regulation mechanism of the GAL network.

Table 7: Network structures identified by the SGPA framework. These structures are capable of capturing bimodality and reproducing the experimental data. These sets are obtained by running the optimization algorithm and fitting with respect to combinatorial gene deletion strains. The initial parameter sets are chosen randomly. A total of 186 optimized sets were analyzed to obtain the 41 structures presented here. The p-values are indicative of the strength of the parameter set to reproduce the data for full deletion strength (GAL3-/-, GAL4-/- and GAL80-/-). The best topology found by this method is the first in the list (Structure 1: 1 1 0 -1 **1** -1 1 1) with a p-value <0.05; the statistical test utilized to calculate the p-value is based on the ZTEST with the assumption of normal distribution (see Material and Methods for a detailed explanation). The correct network structure is: 1 1 0 -1 **0** -1 1 1. The fitness scores with respect to the knock-down and knock-out strains are provided in the supplemental material section.

ID	Network Structure Tcoefficients: T21,	T23,	T24,	T34,	T32,	T42,	T43,	T53	p-value
1	1	1	0	-1	1	-1	1	1	0.037
2	-1	1	1	1	-1	1	-1	1	0.051
3	1	-1	-1	-1	1	-1	1	1	0.079
4	1	-1	-1	-1	-1	1	0	-1	0.097
5	1	-1	-1	1	-1	1	-1	-1	0.097
6	1	-1	-1	1	-1	1	1	-1	0.097
7	1	-1	1	-1	-1	-1	-1	-1	0.097
8	1	-1	1	-1	1	1	1	-1	0.097
9	1	-1	1	1	-1	1	-1	-1	0.097
10	1	-1	1	1	-1	-1	-1	-1	0.113
11	1	-1	-1	-1	-1	-1	-1	-1	0.187
12	1	-1	1	1	-1	-1	1	-1	0.201
13	1	1	-1	1	1	-1	-1	1	0.204
14	1	-1	-1	-1	-1	-1	1	-1	0.240
15	1	1	-1	1	1	1	-1	1	0.253
16	-1	1	1	1	1	-1	-1	-1	0.256
17	-1	-1	1	1	1	1	-1	-1	0.271
18	-1	1	1	0	1	-1	1	-1	0.276
19	1	1	1	-1	1	-1	-1	1	0.278
20	1	1	1	1	-1	-1	-1	-1	0.307
21	1	-1	-1	1	1	1	-1	1	0.322
22	1	1	0	-1	0	-1	1	1	0.342
23	1	1	0	-1	0	-1	0	1	0.366
24	1	-1	-1	-1	-1	1	-1	-1	0.371
25	1	-1	1	1	1	-1	-1	-1	0.397
26	1	-1	-1	1	-1	-1	-1	-1	0.397
27	-1	-1	-1	1	-1	-1	-1	1	0.461
28	1	-1	1	1	-1	1	1	-1	0.602
29	-1	1	1	0	1	-1	-1	-1	0.618
30	-1	1	-1	1	1	1	-1	-1	0.627
31	1	1	-1	1	-1	-1	-1	-1	0.640
32	-1	1	1	-1	1	1	-1	-1	0.663
33	-1	-1	1	1	-1	1	-1	1	0.683
34	-1	1	1	-1	1	-1	-1	-1	0.695
35	1	-1	-1	1	-1	-1	1	-1	0.699
36	-1	1	1	1	1	0	-1	-1	0.710
37	-1	1	1	-1	1	-1	0	-1	0.757
38	-1	-1	1	-1	1	1	-1	-1	0.850
39	-1	1	-1	-1	1	1	-1	-1	0.919
40	-1	1	1	-1	1	-1	1	-1	0.979
41	-1	1	1	1	1	1	-1	-1	0.984

Table 7

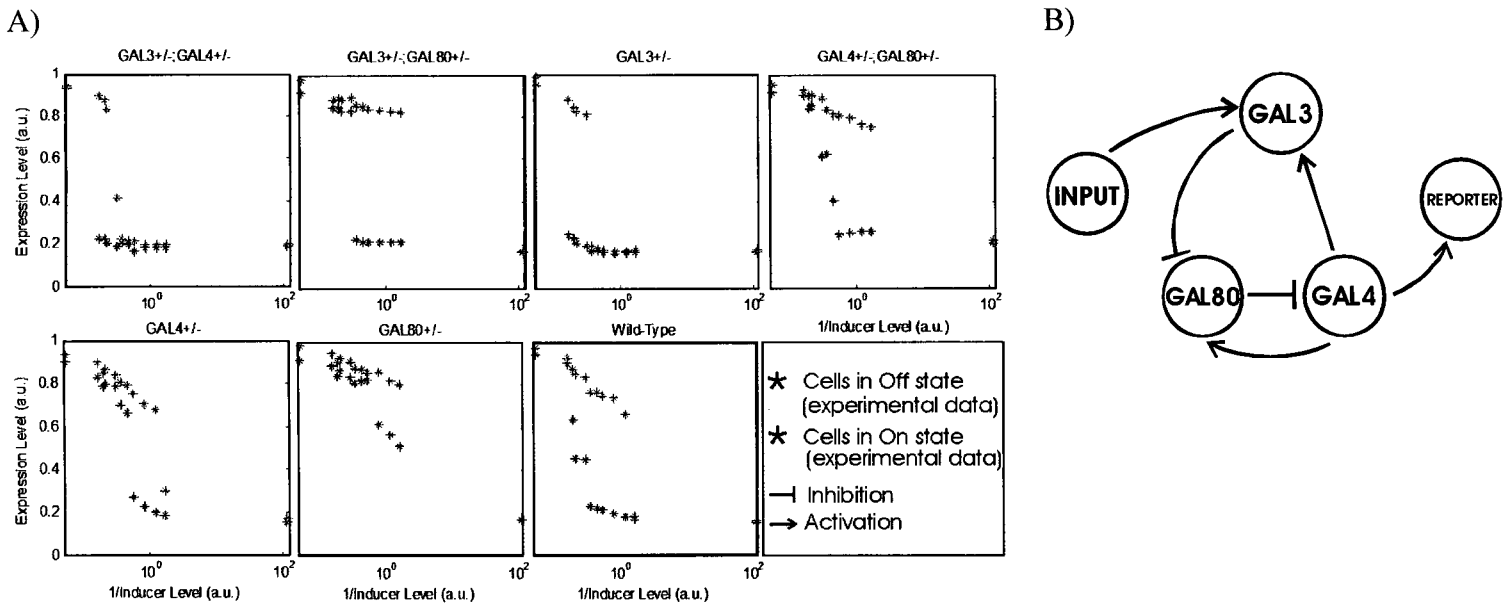


Figure 31: A) Experimental data for the core regulatory elements of the GAL network. The genes involved are GAL4, GAL3 and GAL80; B) The known structure of the network consists of four main steps: 1) induction of Gal3p by galactose (input), repression of Gal80p by Gal3p; 3) repression of Gal4p by Gal80p and 4) activation of the GAL genes by Gal4p.

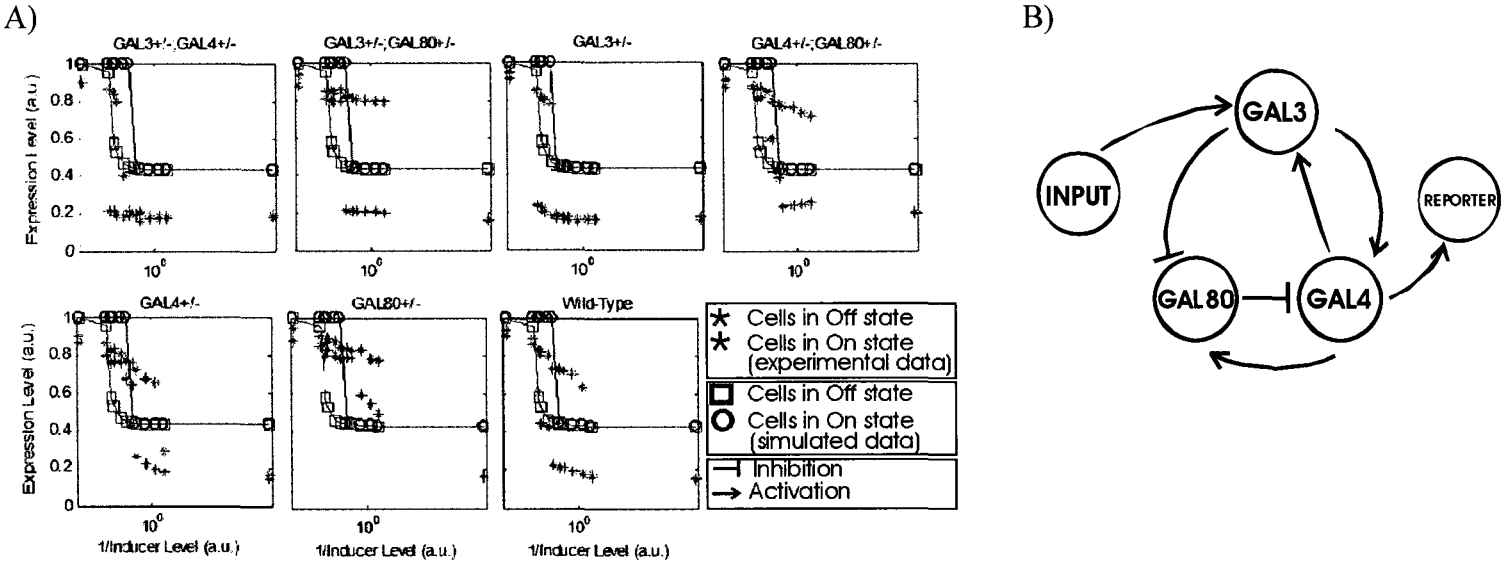


Figure 32: A) Experimental and simulated data following Parameter Optimization for the core regulatory elements of the GAL network. Bimodality is observed for all the genotypes; B) respective network architecture based on the given T coefficients (T-coefficients>0 represent activation, T-coefficients <0 represent repression, T-coefficients =0 represent no interaction). Sum of fitness score for the seven genotypes is 7.0; sum of fitness score for the gene knock-outs (GAL3-/-, GAL4-/- and GAL80-/-) is 1.8, which is significantly low (p-value <0.05) when compared to the other 40 topologies. Fitness scores from knock-out strains are used for p-value evaluation. The statistical test utilized to calculate the p-value is based on the ZTEST with the assumption of normal distribution (see Material and Methods for a detailed explanation).

Chapter 4: Discussion

Identification of Regulatory Genes

To select genes that are important for a given system, it is essential to have a standardized screening method. Synthetic genetic array (SGA) and epistatic microarray profile (E-MAP) technologies have been successful in grouping genes with similar functions. SGA (Tong A.H.Y. *et al.*, 2001) is a qualitative measure of growth phenotype of haploid double deletion strains. E-MAP (Collins S.R. *et al.*, 2006), on the other hand, relies on a quantitative growth rate of double versus single deletion strains to determine aggravating (negative) or alleviating (positive) interactions. Both SGA and E-MAP technologies cannot be used for essential genes, as their deletions are lethal for the organism. We have used quantitative growth phenotype to identify regulatory genes. The method relies on growth phenotype of gene knock-out as well as gene knock-down strains. The technique is successful in detecting genes with positive regulatory roles. The second alternative was to compare galactose dose response curves of perturbed strains to dose response curve of the wild-type. This technique provided quality data and clearly identified core regulatory genes of the network. The advantage of knowing the GAL system and its regulatory components helps us refine and optimize our strategy for further development of the SGPA framework.

Finally, a single universally applicable approach to identify key elements of a

network does not exist. It is a combination of different complementary techniques that will elucidate functional roles and relationships among genes. In fact, it is important to first utilize a fast and reliable method for identification of regulatory elements, such as high throughput growth assay. This could then be followed by a more target oriented experiment, such as dose response curve analysis (or expression profiling).

Generation of Systematically Perturbed Networks

Gene dosage perturbation strategy is based on generating a library of strains in which gene copy number is reduced from two copies to one. This technique produces a large pool of strains even for a small regulatory network. Systematic approaches usually utilize gene knock-outs (Tong A.H.Y. *et al.,* Science 2001; Collins S.R. *et al.,* 2006) or over-expression (Sopko R. *et al.,* 2006; Garder T.S., 2003) to investigate functional roles and interactions among genes. We used combinatorial gene deletion for investigation of functional relationships among genes. This idea is not novel (Collins S.R. *et al.*, 2006; Deans, Cantor and Collins 2007). However, our strategy to use gene knock-down strains with combinatorial gene copy number reduction is innovative and provides additional advantages: 1) the network topologies are not changed, in fact all the nodes in the network will still be present after perturbation; 2) the network under investigation is functional even though its parameters are slightly modified; 3) there will not be compensation through redundancy or rewiring among genes – since all the genes are present, there will be no need for redundant genes to come into play; and 4) generation of

a complete library is practical and can be scaled to more complex systems.

Quantitative Data Generation: Single Cell Measurements at Steady State

To model the system, with no prior knowledge, it is important to have a solid framework where the input to the system and output of the system are well characterized and easily measured. The input to the galactose utilization pathway is the available galactose. Variation of galactose concentration provides a window for investigation of its effect. The output is a downstream gene (GAL1) fused to GFP for single cell measurements – GAL1 and all the other GAL genes are regulated by GAL4, therefore measurement of GAL1-GFP represents relative expression of all the GAL genes in the network. Flow cytometry is used to measure single cell expression level of the reporter gene for a range of input signals. By having access to single cell data, it is possible to analyze the mean of populations in both ON and OFF state separately. Single cell measurements also provide a high quality, quantitative data that can be used to gain insight into dynamic interactions among genes. Growth phenotypes or microarray datasets can provide quantitative data; however, they can only be measured for the total population and therefore not adequate for modeling the dynamics of a system. Single cell datasets are valuable because of their quality and the amount of information they provide. Many modeling frameworks (for example Gardner T.S. *et al.,* 2003) rely on micro-arrays or qPCRs, as they provide high-throughput datasets in a timely manner.

Gardner T.S. *et al.,* (2003) employed over-expression experimental data to model a gene regulatory network (nine-transcript sub-network of the SOS pathway in *E. coli*). Their modeling framework required measurement of all species in the system, for a large number of perturbation experiments. This requirement limits the methodology to smaller systems and becomes infeasible and expensive with larger networks (i.e. for a nine gene network 81 experiments are required). Our framework, on the other hand, required only measurements from a reporter gene for each perturbation experiment. The number of perturbation experiment depends on the number of genes in the network of interest. For example: for a three gene network, we required seven perturbation experiments for model generation and three for model discrimination (therefore 10 dose response curves in total). For a nine gene network, we require a total of 55 dose response curves. In essence, our strategy is more practical and scalable to larger networks. The Gardner T.S. *et al.,* (2003) approach required expression data from the average population. SGPA, on the contrary, relies on high quality, in vivo, single cell datasets, which can be easily generated. Because of the nature of data generated by Gardner T.S. *et al.,* (2003) and their modeling framework, their system had to be linear in order to conform to the developed equations. This assumption was based on the fact that over-expression perturbs the system slightly. We have not assumed a linear system. Indeed our modeling framework embedded a non-linear function (g(u) term in Equation 1) to capture the non-linearity of the system. Finally, our framework requires less data, but higher quality data for system-identification of gene regulatory networks.

Systematic Approach for Generation of a Pool of Potential Topologies

The availability of high quality quantitative data makes the modeling process viable. To model the GAL network a generic model based on ODEs was developed.

To fit the model to the experimental data from the seven genotypes, a Genetic Algorithm (GA) was implemented. The GA was based on mutation and crossover with a variable mutation rate and a constant crossover rate. The GA was utilized to find sets of parameters that would best reproduce the experimental data. The fitness function, favoring bimodality, guided the GA toward finding solutions that would reproduce the bimodality observed. This characteristic was important, as it reduced the number of network topologies significantly. Bistability condition is an inherent property of the GAL system; in fact, in cases where the key regulatory elements were deleted, bistability was lost, thus making a link between dynamical stability and the functional roles of regulatory elements in the network.

We have also attempted to fit our experimental data to a detailed model of the GAL network (developed by Ramsey S. *et al.,* (2006) and implemented here in Matlab). However, we have not been able to find a parameter set that would reproduce our experimental data for the wild-type strain. The model (see Suppl. Material), as also described in the first chapter, is very detailed, with 19 equations and 44 parameters. It is challenging to have a single model that would satisfy all experimental results. It is in fact

important to build simple models for investigation of functional roles and regulatory relationship among genes. Complex and detailed models could nonetheless answer specific biological questions.

In a recent study, Wollman R. *et al.,* (2008) attempted to model force integrating during mitosis in *Drosophila* embryo using experimental data from wild-type and perturbed strains. They were able to generate 1500 models by applying the dataset from wild-type strain. Fitting the data to a total of six different genotypes reduced the number of potential candidates from 1500 to only six. Their model had 39 parameters and was composed of eight different building blocks. The goal was to fit the experimental data to a model and measure how local perturbation experiments can provide additional information to reduce the number of possible solutions. They used stochastic optimization algorithm (based on GA) in addition to clustering analysis and concluded that wild-type dataset is not sufficient to differentiate among models. This observation is in accordance with our findings; in fact, we found that by applying datasets from multiple genotypes, it is possible to find a limited number of solutions which are a good representation of the known network structure.

Nevertheless, there are three fundamental differences between our modeling framework and that of Wollman R *et al.* (2008): 1) to accommodate different genotypes, their model (Wollman R. *et al.,* 2008) had to incorporate a parameter for each new genotype (for a total of 45 parameters); we did not have to increase the number of

parameters: a multiplicative factor, gene copy number, was added to each equation; 2) their approach was based on assumption of a linear model; ours, does not assume linearity and did capture non-linearity in the system; and 3) their model was constructed based on eight building blocks from known literature, whereas ours could be applied in a more systematic manner with no prior knowledge.

Model Discrimination Using Knock-Out Strains

Model discrimination is an integral part of the system identification process. A solid framework must have a rigorous strategy to discriminate among viable topologies. For a three gene network, we have identified 41 different topologies; there is therefore an apparent need for systematic evaluation of these structures.There are different approaches to achieve this goal; parameter sensitivity analysis (PSA) is one of those. PSA measures the range of parameters that would reproduce the experimental results; a wider range is translated to a more robust network structure. This process is computationally expensive and it is based on the assumption that biological systems are evolved to be robust to perturbations. In fact, we have also shown, experimentally for the first time, that the GAL system is relatively robust to gene dosage perturbation. Therefore, sensitivity analysis could be a good approach in this case.

A second approach is to run the optimization algorithm many times to investigate the occurrence of certain topologies. To have a statistically valid view of the potential

candidates, the optimization process has be run for a large number of trials, and with a larger network this process may quickly become computationally infeasible. However, this type of analysis allows ranking of the structures based on their parameter sensitivity and structural robustness. In fact, the optimization algorithm will likely converge to structures that permit a wider range of parameter values.

A third approach that is even more natural and very intuitive is to utilize expression data from gene knock-outs that were initially used to identify regulatory elements. Data from gene knock-outs would represent indeed a test-set. In essence, the viable topologies have to reproduce the experimental data where both copies of a gene are deleted for all the genes in the network. For a three gene network, this implies that the best candidate has to reproduce the data for three extra genotypes. In fact, our method which is based on gene dosage perturbation, cannot guarantee that decreasing gene copy number translates to reduction of all protein levels by half. Therefore, a model discrimination approach that relies on validating the results with experimental data from knock-out strains is very logical. Furthermore, the test-set would increase in size with larger networks, providing a scalable methodology for model discrimination.

Following the last described approach, the 41 candidate topologies were reduced to only one viable topology (p-value <0.05). The latter is a close representation of the GAL network. The top ranked topology represents the known GAL network, but with only one extra connection: a redundant link between GAL3 and GAL4. The second and third best

topologies had a p-value <0.1. The second best topology was different from the known GAL system; however, it did provide some insight into the regulatory structure of the network. It captured GAL3 as the main activator and it also highlighted the main structure of the system in the network. In that topology GAL3 activates GAL80 in two different ways: 1) GAL3 activates GAL80 directly and 2) GAL3 activates GAL80 indirectly, through GAL4. GAL80 then activates GAL3. Obviously this is not correct; however, the main features of the system are embedded in there: direct and indirect activation of a gene (in this case the role of GAL80 and GAL4 are reversed) that is the main activator of the system.

Galactose utilization pathway has been studied for more than 60 years (Reviewed by Barnett J.A., 2007; Winge and Roberts 1948). Its structure and dynamics are well established. However, the function of Gal3p in sequestering Gal80p in the cytoplasm has only been discovered in 2002 (Peng and Hopper 2002). Prior to that, it was believed that Gal3p activates Gal4p directly by forming a complex with Gal80p in the nucleus (Gal80p-Gal3p-Gal4p) which destabilizes Gal4p-Gal80p dimer. This model was accepted because the complex formation (Gal80p-Gal3p-Gal4p) was observed upon induction with galactose (Sil, Xin and Hopper 2000; Platt and Reece 1998). However, that experiment was performed in vitro. A subsequent study found that Gal3p localizes only in the cytoplasm (Peng and Hopper 2002) and therefore the model in which GAL3 activates GAL4 directly was rejected. In fact, the model obtained by our analysis (p-value < 0.05)

provides two activation modes of GAL4 by GAL3: 1) GAL3 activates GAL4 directly, and 2) GAL3 activates GAL4 through GAL80. This demonstrates the strength of our modeling framework and optimization algorithm and their potential to direct the experimentalist to design and implement experiments that can provide useful information toward our understanding of cellular regulation and dynamics.

Future Directions

The goal of this research project was tool development for reverse-engineering gene regulatory networks. The design and implementation of this framework could be optimized and further developed. For instance, the GA algorithm could be further refined by introducing controlled mutation rate rather than random mutation rate. A pre-optimization algorithm could also be implemented and used prior to GA for a faster convergence. In addition to that, a model discrimination approach based on parameter sensitivity analysis could be designed, implemented, and compared to the presented methodology.

SGPA provides a solid framework for system-identification and reverse-engineering of gene regulatory networks. The integral part of the data is based on single cell flow Cytometry, therefore this methodology could be applied to larger networks without major limitations. Furthermore, the strain generation is very simple, fast and practical, providing a sustainable scheme for system identification based on steady state single cell measurements. Moreover, data analysis could be expanded to incorporate level of noise in expressing and non-expressing cells. This additional knowledge may provide added information about the system. However, the implementation of this research strategy requires a system where input output relationships are well identified and measurable.

Using a systems approach, it was possible to develop and test this tool for

extraction of quantitative and mechanistic information to gain deeper insight into cellular regulation and network structure. The methodology presented here could be easily implemented and applied to higher organisms with RNAi technology, where reduction in protein level could give insights into network topologies. Full knock-out of genes in the network would eventually reduce the set of viable structures, leading to new testable hypotheses. The galactose utilization pathway was used for implementation and validation of this tool development because its structure is well defined and therefore allows us to benchmark the developed technique and improve upon it. The GAL network identified by our systematic analysis is a very close representation of the known structure.

Considering a limited number of iterations, the presented approach for dynamical modeling was successful in identifying network structures based on in vivo single cell dataset. The computing limitations and the available time frame did not allow for a deeper search for parameter space. This project leads the way to a better application of available resources and provides a scalable framework for system identification and reverse-engineering of biological networks based on in vivo systematic data generation. The approach could potentially be applied to larger networks in a systematic manner.

Appendix

PCR validation

I: PCR validation in ADE2 gene integration

Primers for Reaction A: Int pRS 5 - R1 and Int ADE2 5? F1

Primers for Reaction B: Int pRS 3 - F1 and Int ADE2 3? R1

Primers for Reaction C: Int ADE2 5 - F1 and Int ADE2 3? R1

For positive clones	Validation for Ade2 gene disruption
Reaction A	1.6kbp
Reaction B	0.6kbp
Reaction C	too long for the PCR
For negative clones	
Reaction C	>2kbp
Reaction A and B should not produce any fragment	

II: PCR validation for ADE2del with integration of HIS3 or URA3 cassette

Primers for Reaction A:

FW primer for Ade2 : Int ADE2 5' - F1 and Rev primer for specific cassette (URA3 or HIS3)

Primers for Reaction B:

FW primer for specific cassette (URA3 or HIS3) and Rev primer for Ade2: Int ADE2 3' - R1

Primers for Reaction C:

(not used, since the PCR reaction is not informative, since this will produce 2kbp band in positive and negative clones)

FW and Rev primers for Ade2 gene: Int ADE2 5' - F1 and Int ADE2 3' - R1

For positive clones	Validation for HIS3 cassette	Validation for URA3 cassette
Reaction A	1kbp	1kbp
Reaction B	0.8kbp	0.7kbp
Reaction C	2kbp	2kbp
For negative clones		
Reaction C	2kbp	2kbp
Reaction A and B should not produce any fragment		

Table S.8: Product size for PCR validation reactions

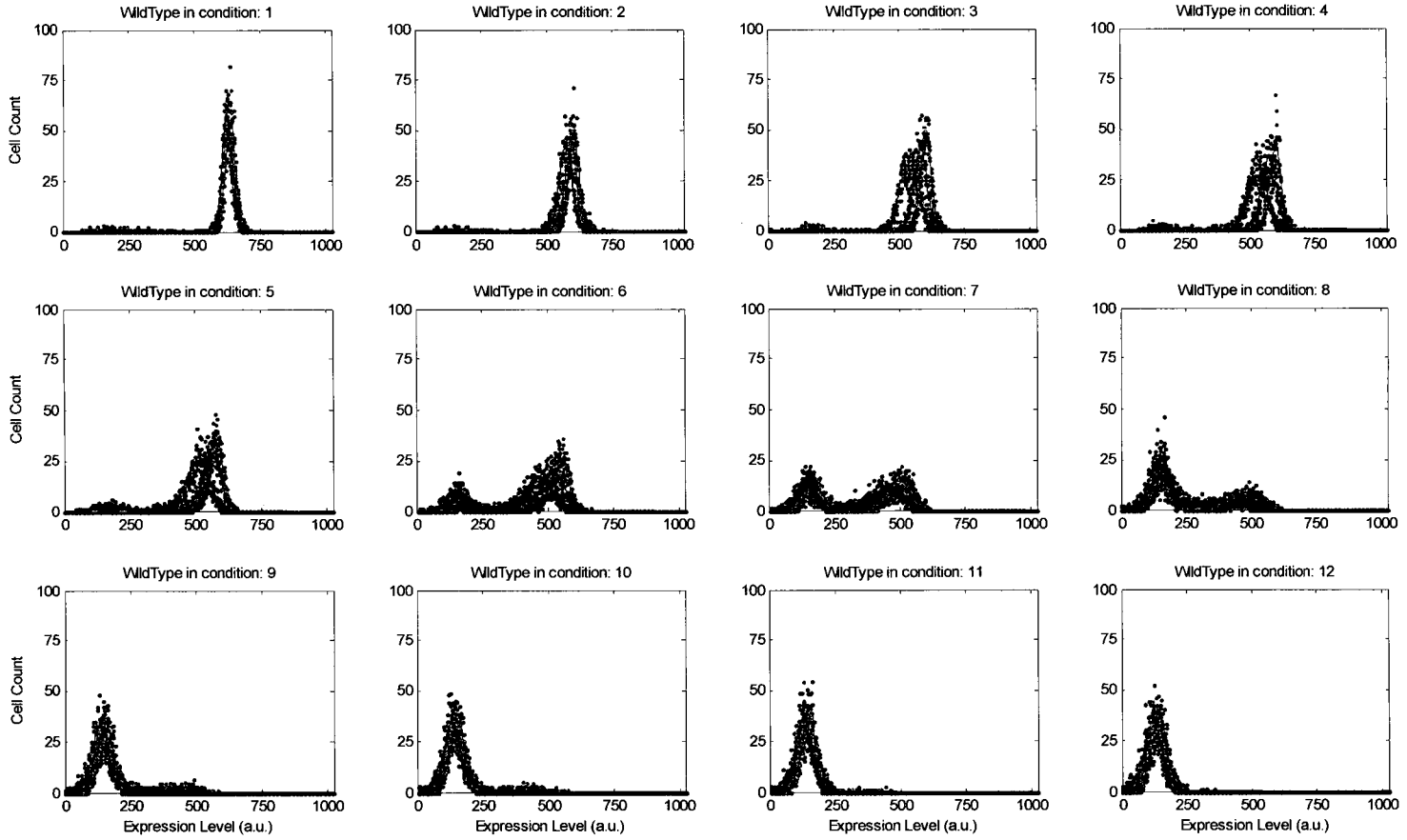
Primer's use	Primer's name	Primer's sequence
Used for PCR-mediated integration of plasmid template sequence into the ADE2 site	B27: ADE2delta0 F1	CTGAGAAGCCGAGAATTTTGT AACACCAACATAACACTGAAG ATTGTACTGAGAGTGCAC
Used for PCR-mediated integration of isotrophic marker cassette into the ADE2 site	B28: ADE2delta0 IntR1	CTTGTTTTCTAGATAAGCTTCG TAACCGACAGTTTCTAACCTGT GCGGTATTTCACACCG
Used for PCR-mediated integration of plasmid template sequence into the ADE2 site	B29: ADE2delta0 IntR2	CTTGTTTTCTAGATAAGCTTCG TAACCGACAGTTTCTAACCCA ATACGCAAACCGCCTCT
Used to confirm PCR-mediated gene disruption of pRS vectors	B22: Int pRS 5' - R1	GGTAtTTTCTCCTTACGCATCT
	B23: Int pRS 3' - F1	CTTGAGTAACTCTTTCCTGTAG GTCA
Used to confirm PCR-mediated gene disruption at the ADE2 locus	B25: Int ADE2 5' - F1	GTTACAGGATTCATGCTTATG
	B26: Int ADE2 3' - R1	GAAAAGGTTACACGGTACAGT CACT
Used to confirm the integration of URA3 cassette	URA3_INTValid_R	TGC CCA TTC TGC TAT TCT
	URA3_INTValid_F	GAC GCA TTG GGT CAA CAG TAT AGA
Used to confirm the integration of HIS3 cassette	HIS3_INTValid_R	TGT GTG GAC GTT AAT CAC
	HIS3_INTValid_F	GGA GTA AAA AGG TTT GGA TCA GGA

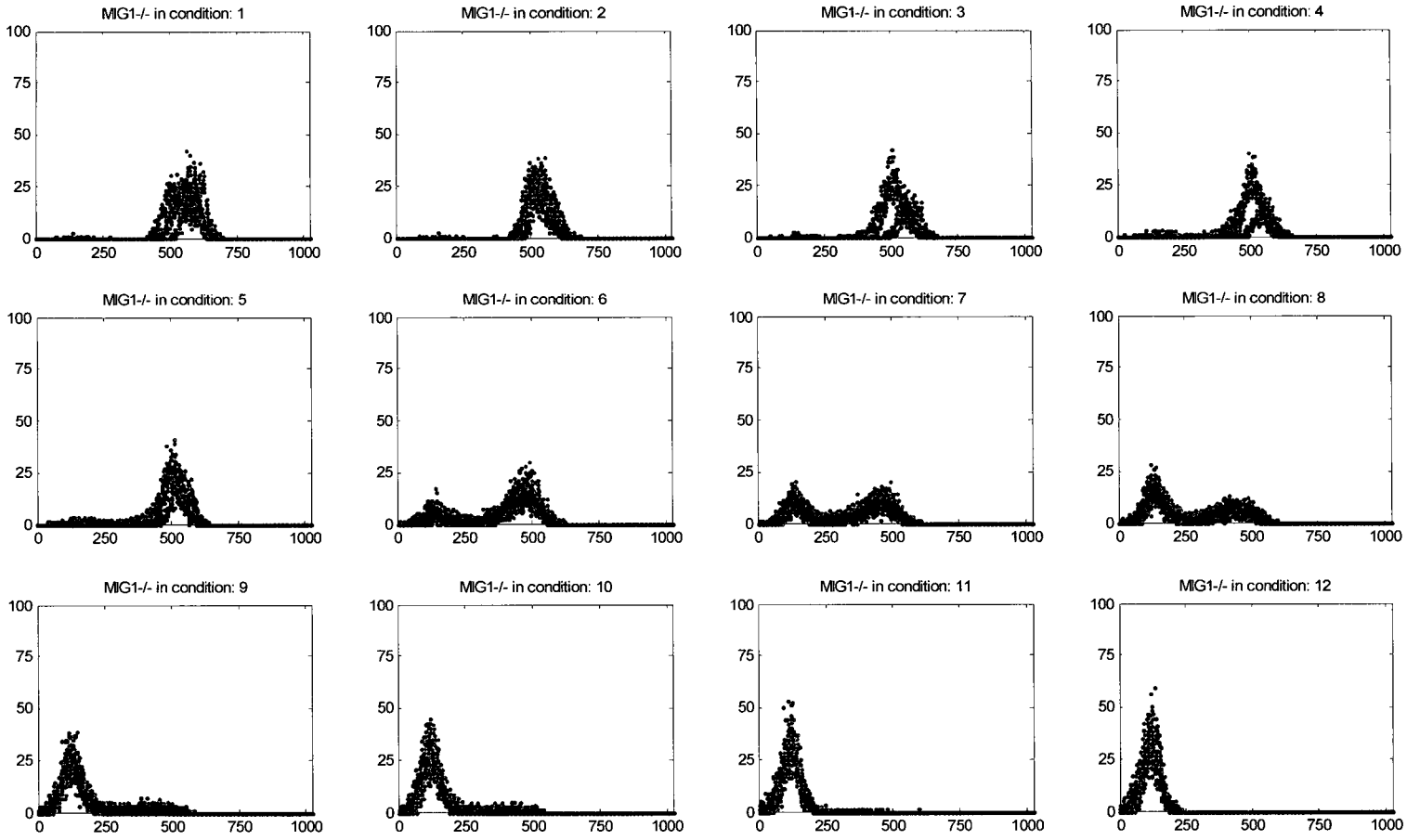
Table S.9: List of primers used for cloning, integration and, validation.

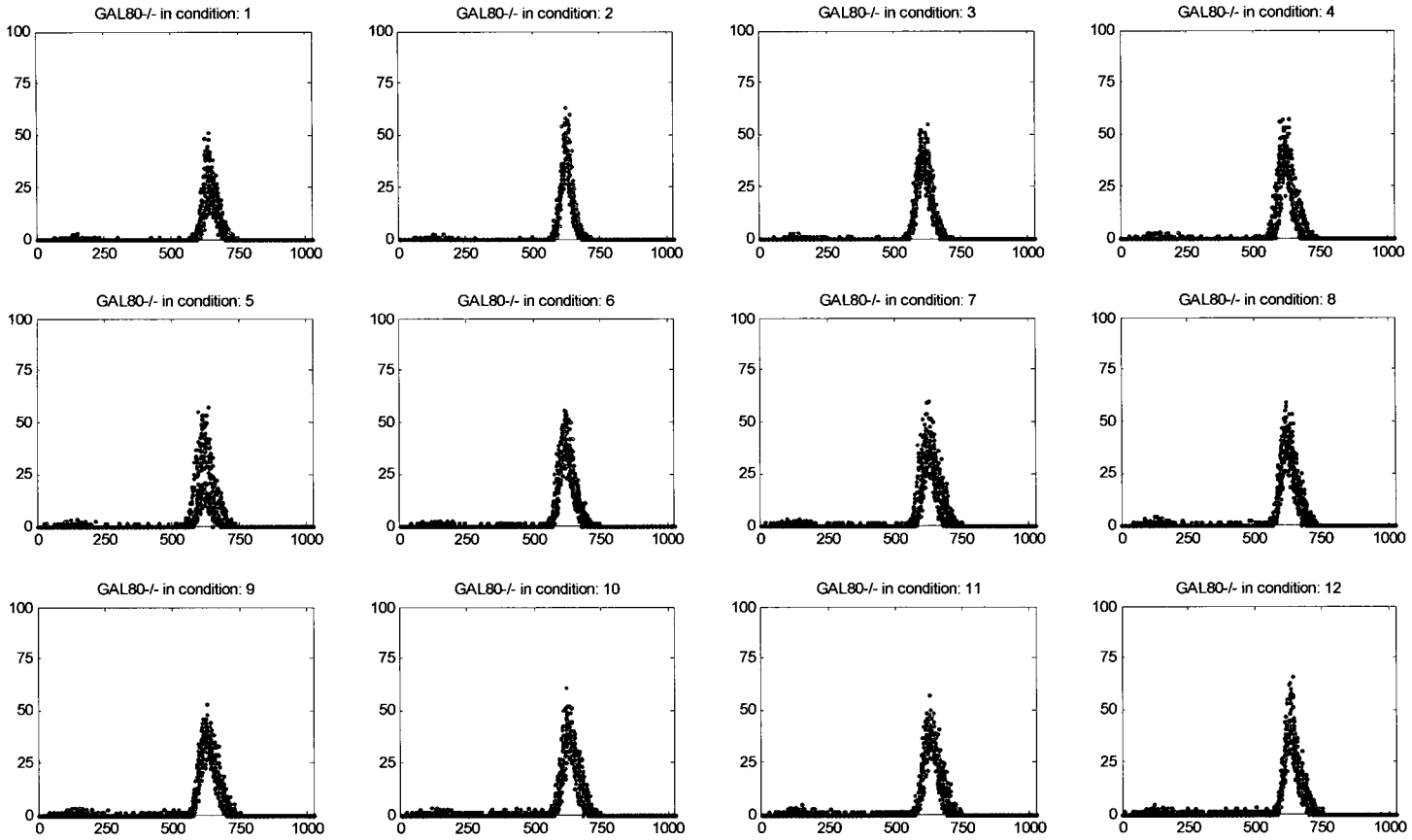
Supplementary Material

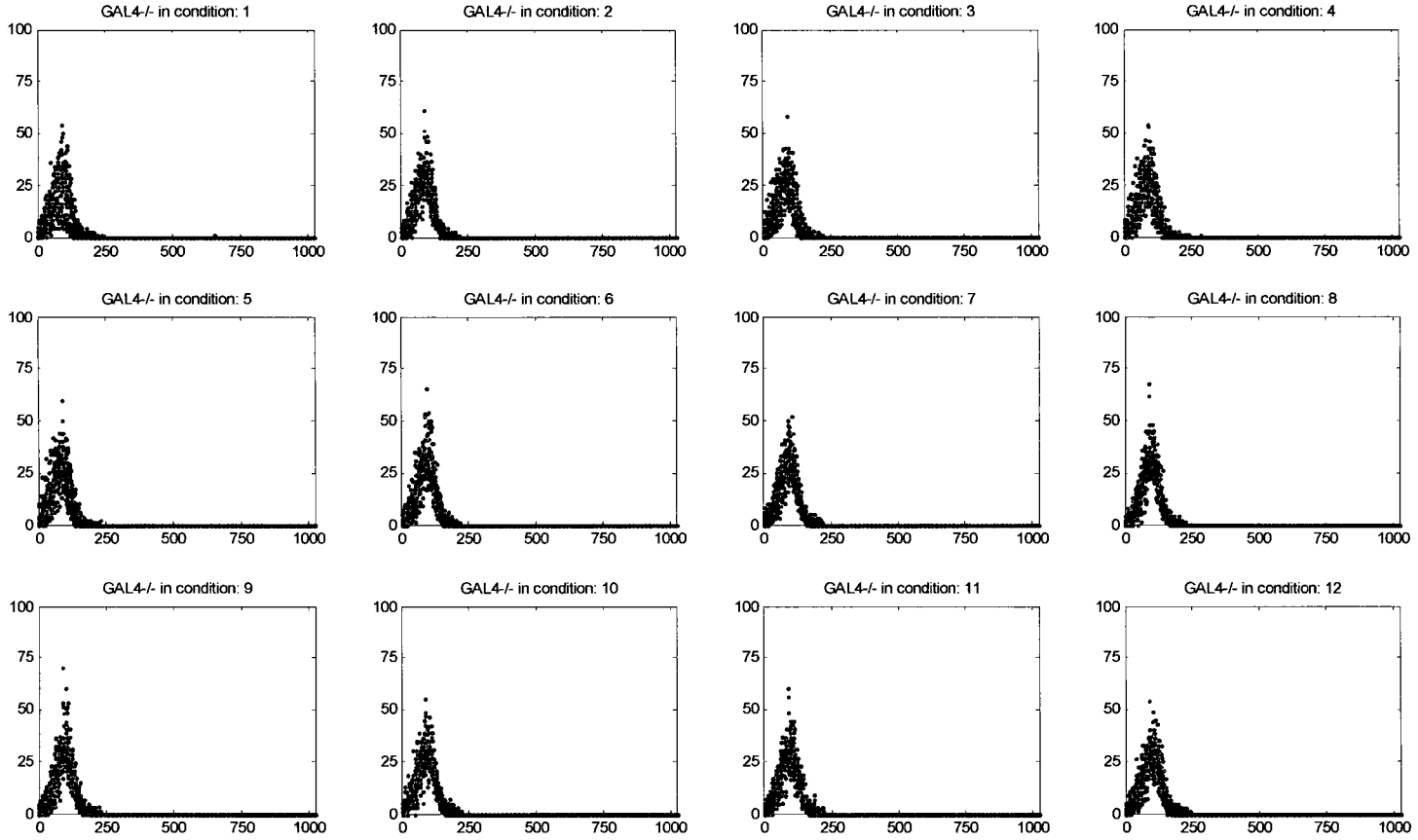
Experimental Data

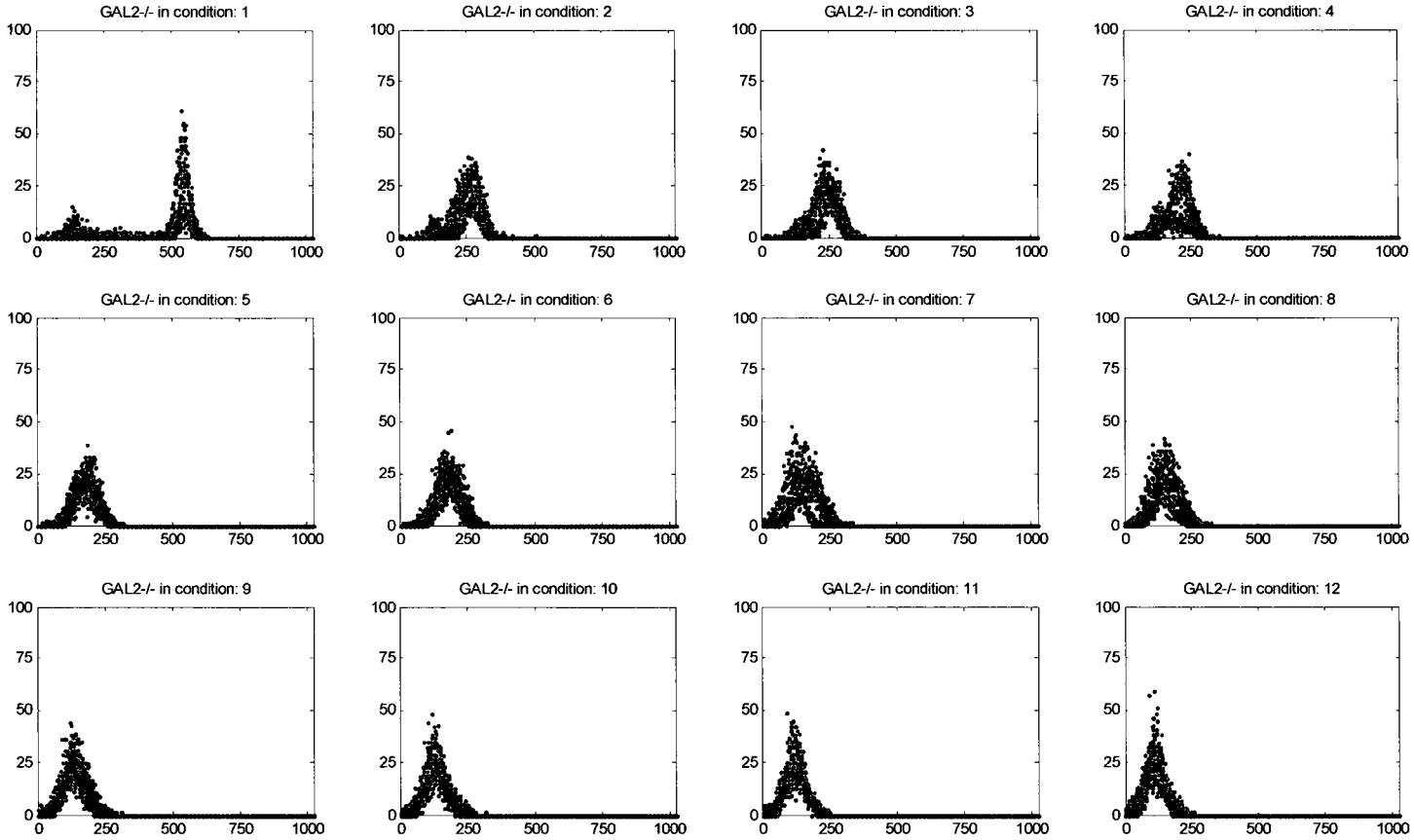
Figure S.33: Triplicate flow Cytometry results for all the diploid strains. Flow Cytometry data for the wild-type and the perturbed strains. The three colors indicate the three independent measurements for the 12 previously defined conditions. The data is represented by dots and the Gaussian distribution (solid lines) is superimposed on the same graph (only for the wild-type strain).

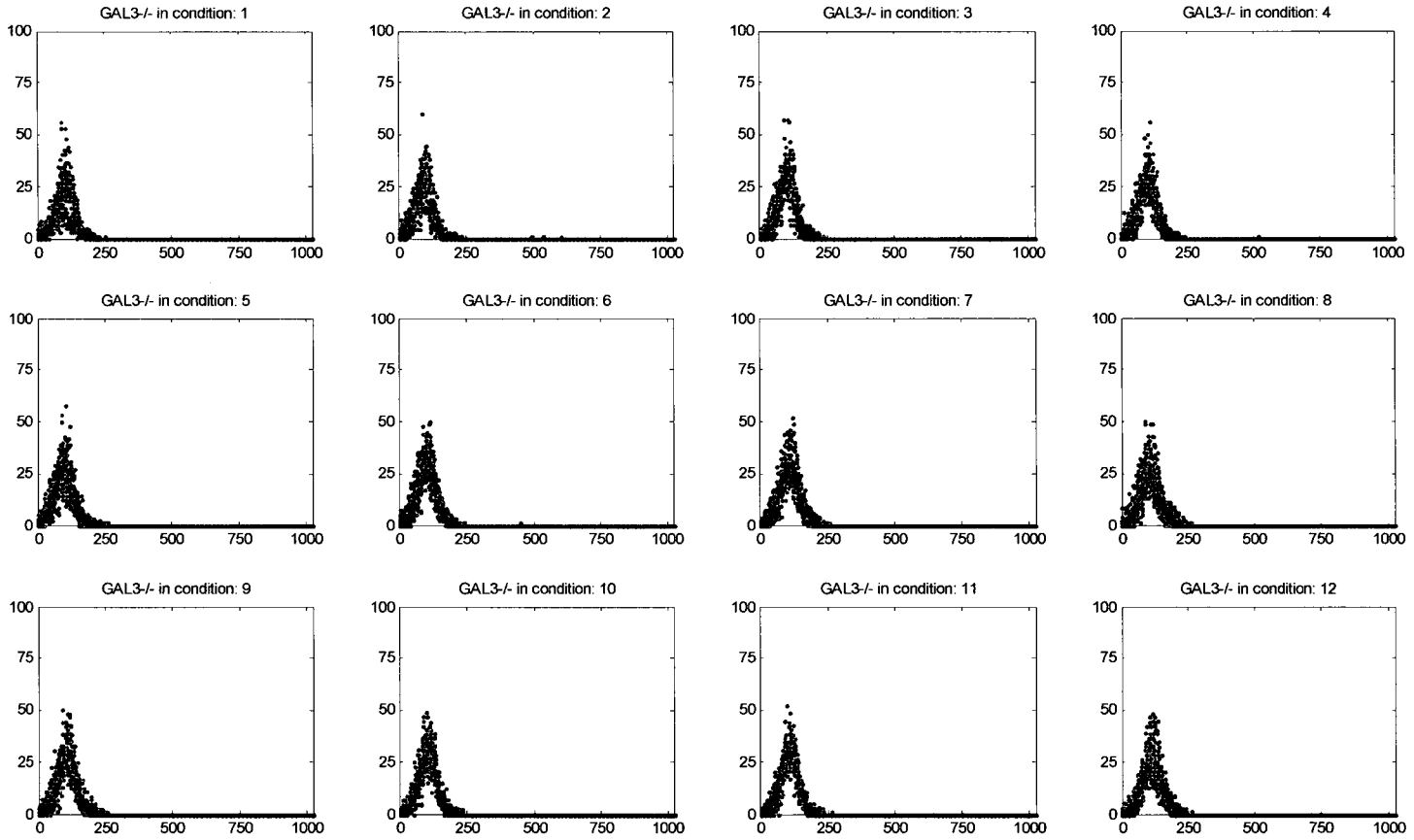

WildType in condition: 1
WildType in condition: 2
WildType in condition: 3
WildType in condition: 4
WildType in condition: 5
WildType in condition: 6
WildType in condition: 7
WildType in condition: 8
WildType in condition: 9
WildType in condition: 10
WildType in condition: 11
WildType in condition: 12
Cell Count
Expression Level (a.u.)
0
25
50
75
100
250
500
750
1000

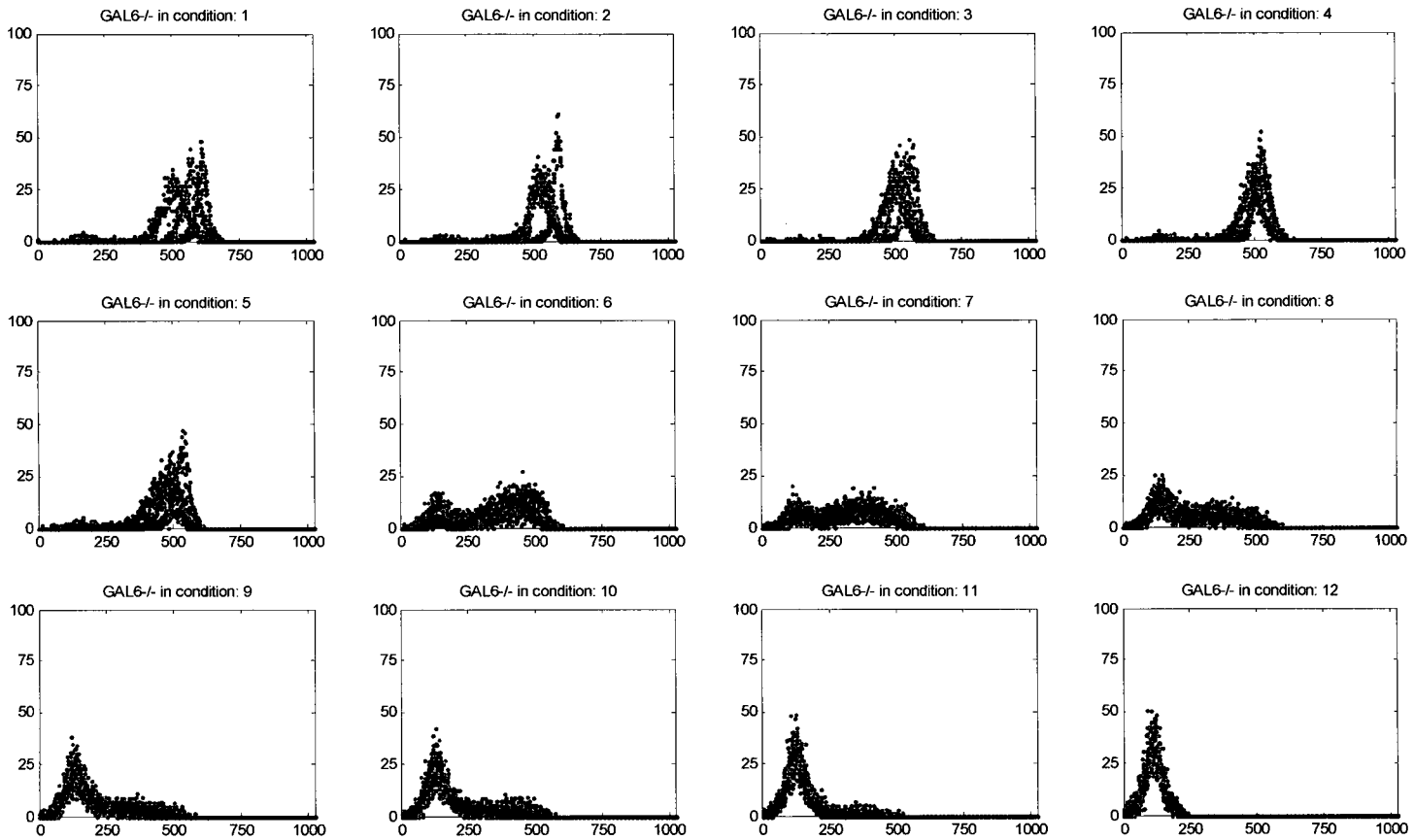

MIG1-/- in condition: 1
MIG1-/- in condition: 2
MIG1-/- in condition: 3
MIG1-/- in condition: 4
MIG1-/- in condition: 5
MIG1-/- in condition: 6
MIG1-/- in condition: 7
MIG1-/- in condition: 8
MIG1-/- in condition: 9
MIG1-/- in condition: 10
MIG1-/- in condition: 11
MIG1-/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

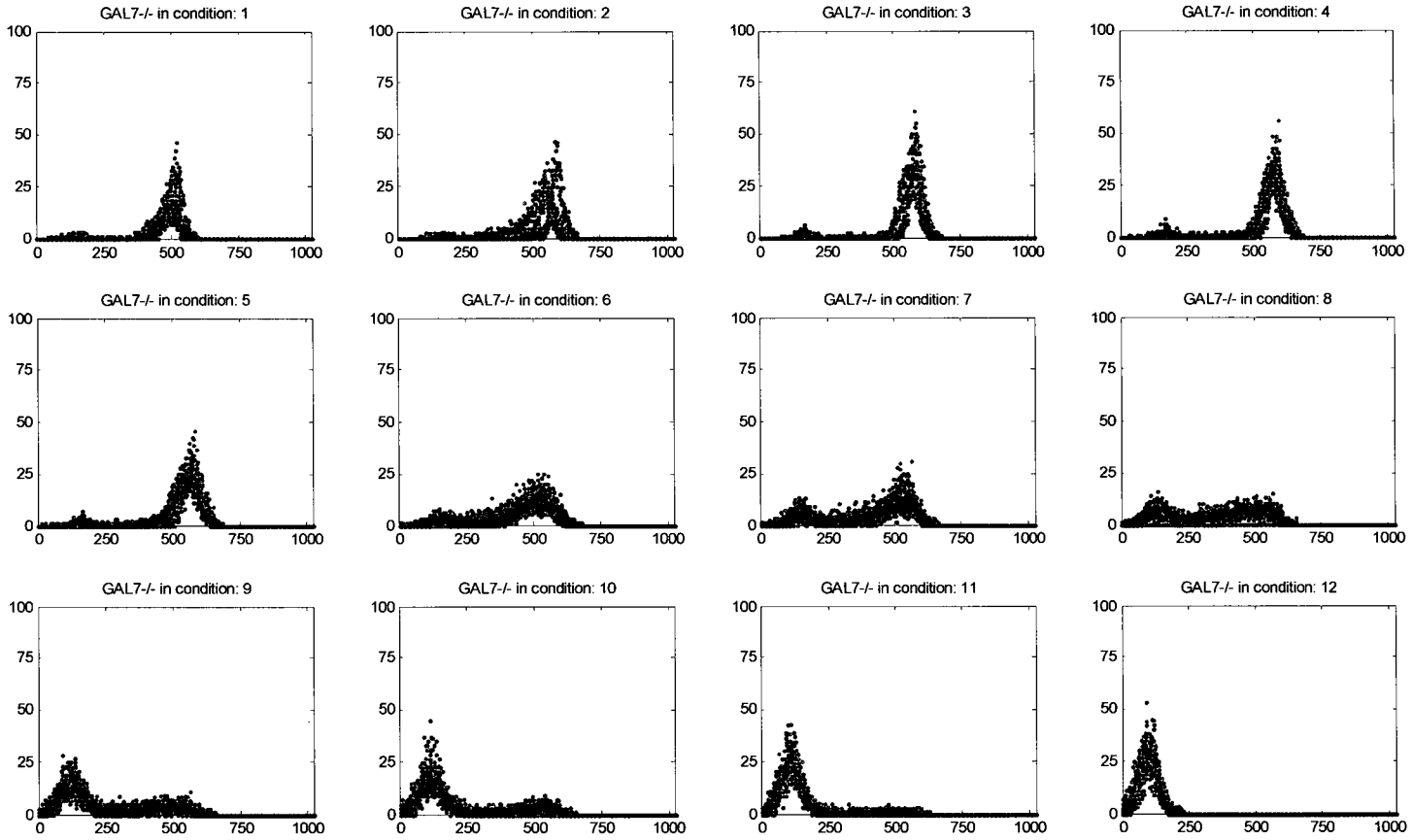

GAL80-/- in condition: 1
GAL80-/- in condition: 2
GAL80-/- in condition: 3
GAL80-/- in condition: 4
GAL80-/- in condition: 5
GAL80-/- in condition: 6
GAL80-/- in condition: 7
GAL80-/- in condition: 8
GAL80-/- in condition: 9
GAL80-/- in condition: 10
GAL80-/- in condition: 11
GAL80-/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

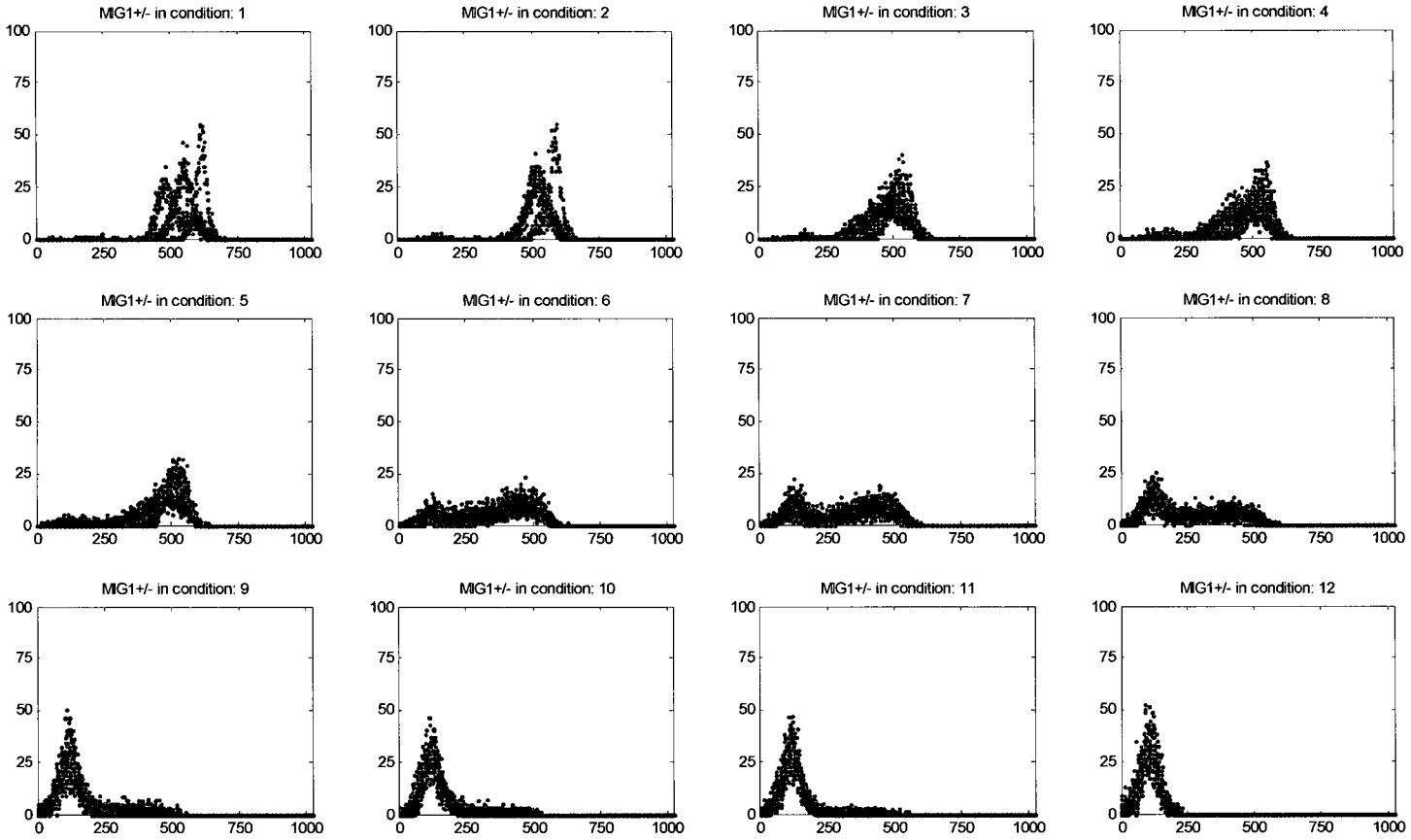

GAL4-/- in condition: 1
GAL4-/- in condition: 2
GAL4-/- in condition: 3
GAL4-/- in condition: 4
GAL4-/- in condition: 5
GAL4-/- in condition: 6
GAL4-/- in condition: 7
GAL4-/- in condition: 8
GAL4-/- in condition: 9
GAL4-/- in condition: 10
GAL4-/- in condition: 11
GAL4-/- in condition: 12
100
75
50
25
0
0
250
500
750
1000

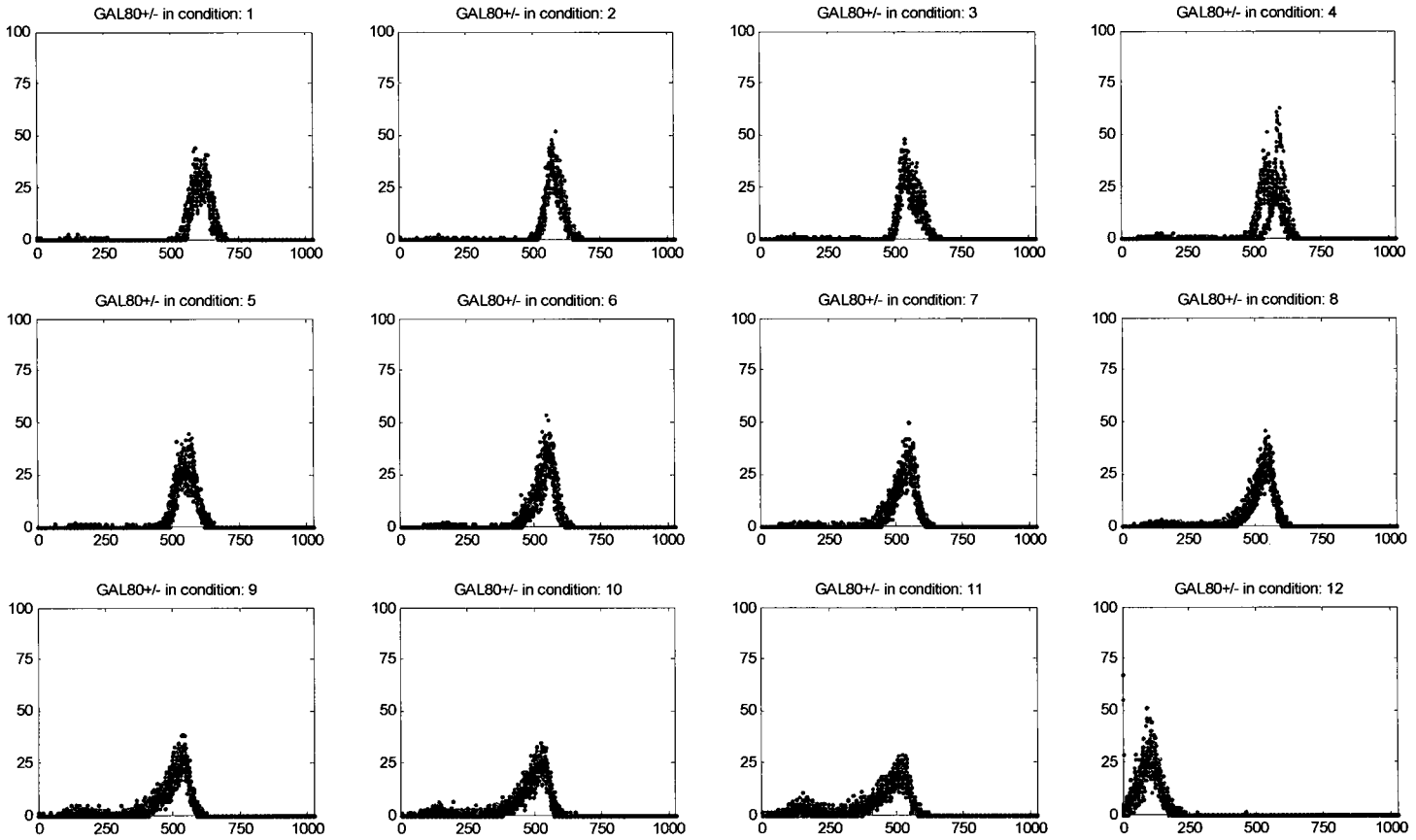

GAL2-/- in condition: 1
GAL2-/- in condition: 2
GAL2-/- in condition: 3
GAL2-/- in condition: 4
GAL2-/- in condition: 5
GAL2-/- in condition: 6
GAL2-/- in condition: 7
GAL2-/- in condition: 8
GAL2-/- in condition: 9
GAL2-/- in condition: 10
GAL2-/- in condition: 11
GAL2-/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

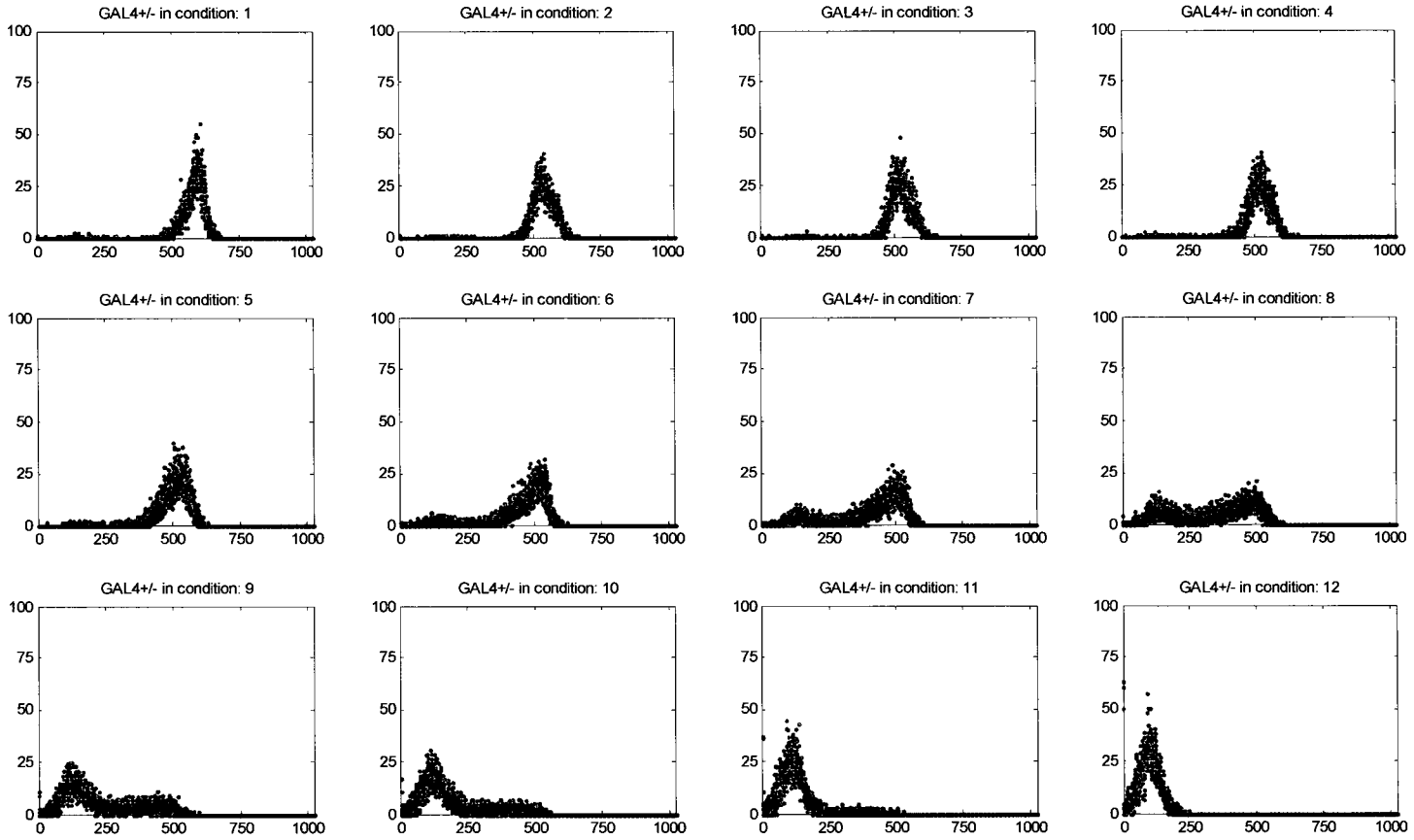

GAL3-/- in condition: 1
GAL3-/- in condition: 2
GAL3-/- in condition: 3
GAL3-/- in condition: 4
GAL3-/- in condition: 5
GAL3-/- in condition: 6
GAL3-/- in condition: 7
GAL3-/- in condition: 8
GAL3-/- in condition: 9
GAL3-/- in condition: 10
GAL3-/- in condition: 11
GAL3-/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

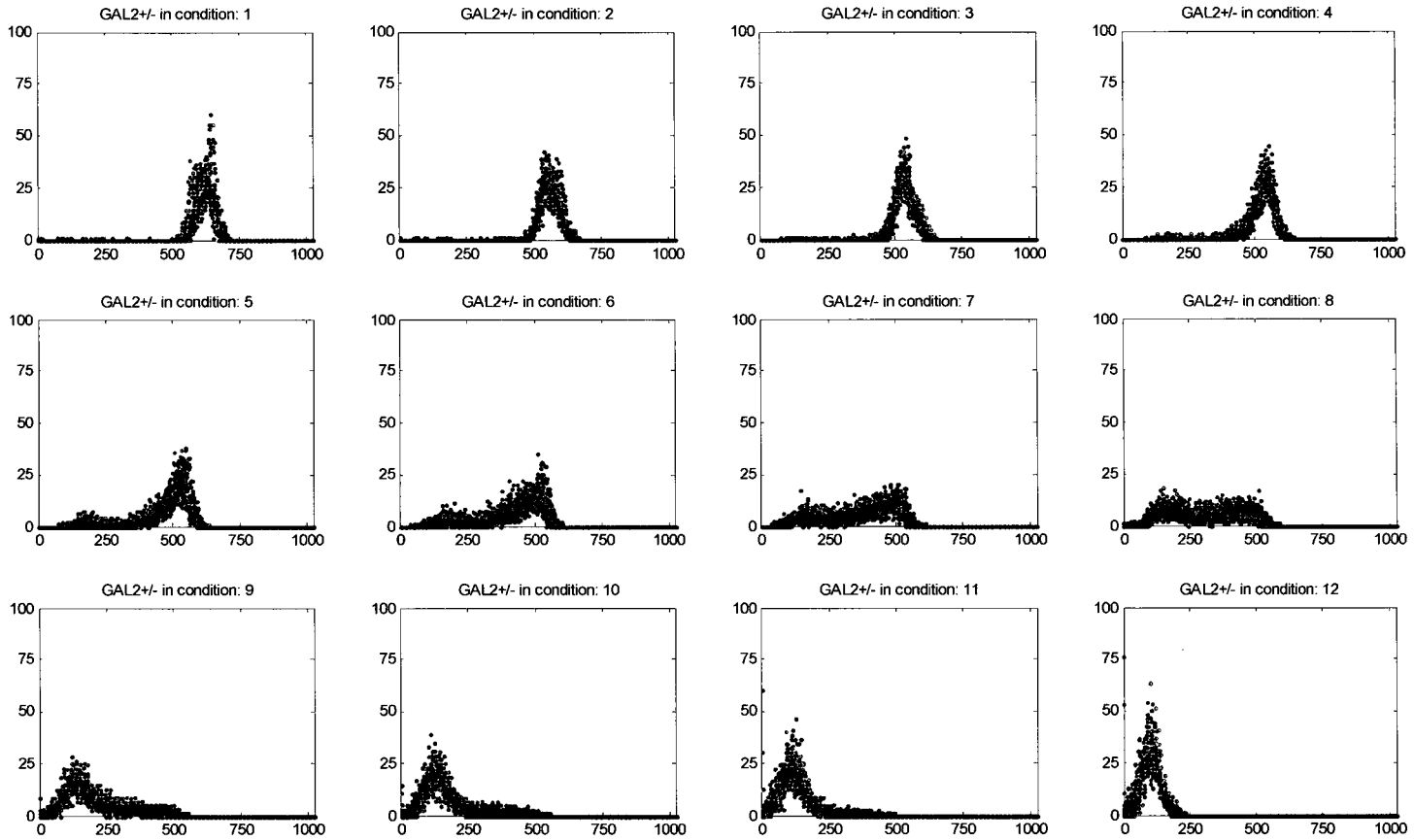

GAL6-/- in condition: 1
GAL6-/- in condition: 2
GAL6-/- in condition: 3
GAL6-/- in condition: 4
GAL6-/- in condition: 5
GAL6-/- in condition: 6
GAL6-/- in condition: 7
GAL6-/- in condition: 8
GAL6-/- in condition: 9
GAL6-/- in condition: 10
GAL6-/- in condition: 11
GAL6-/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

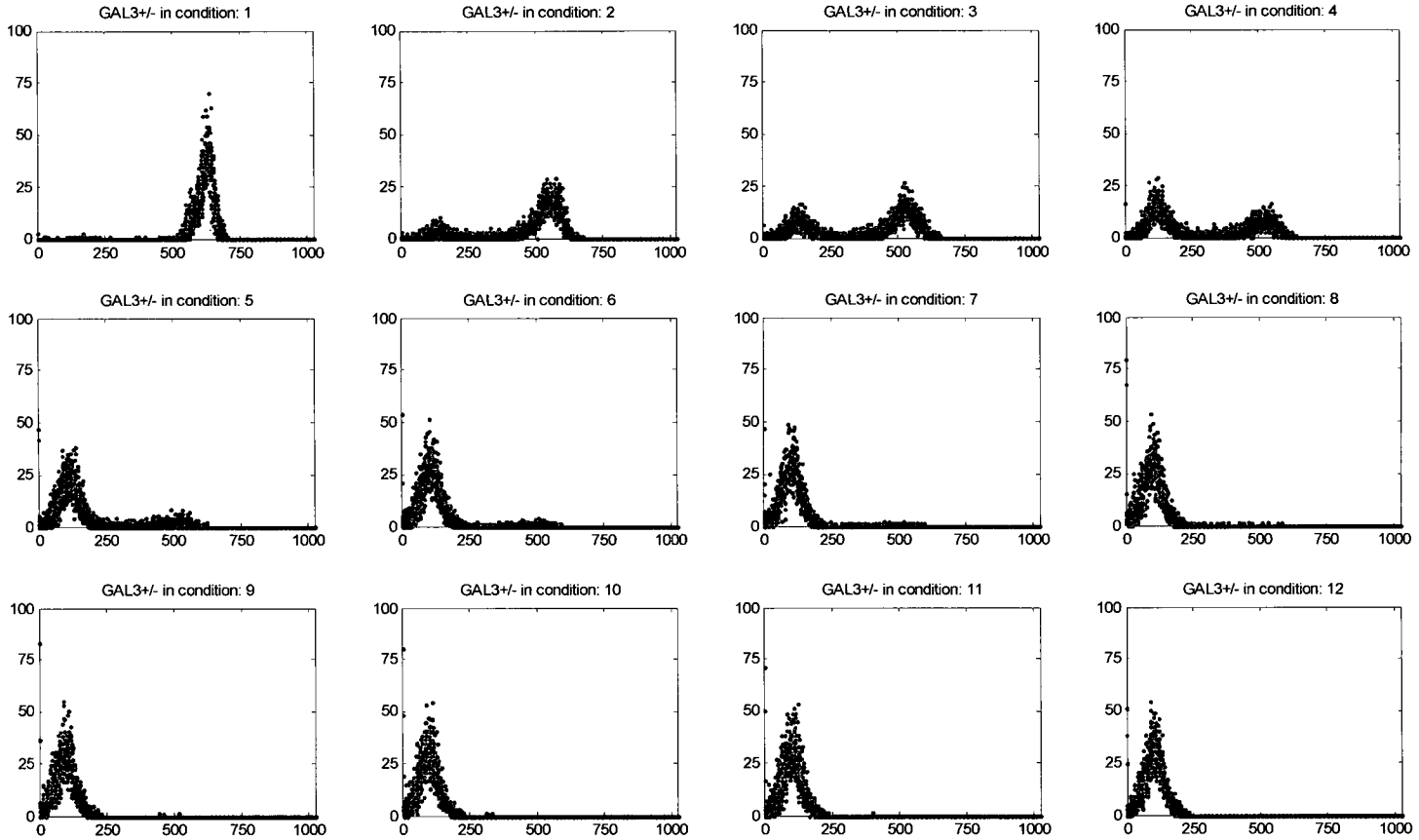

GAL7-/- in condition: 1
GAL7-/- in condition: 2
GAL7-/- in condition: 3
GAL7-/- in condition: 4
GAL7-/- in condition: 5
GAL7-/- in condition: 6
GAL7-/- in condition: 7
GAL7-/- in condition: 8
GAL7-/- in condition: 9
GAL7-/- in condition: 10
GAL7-/- in condition: 11
GAL7-/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

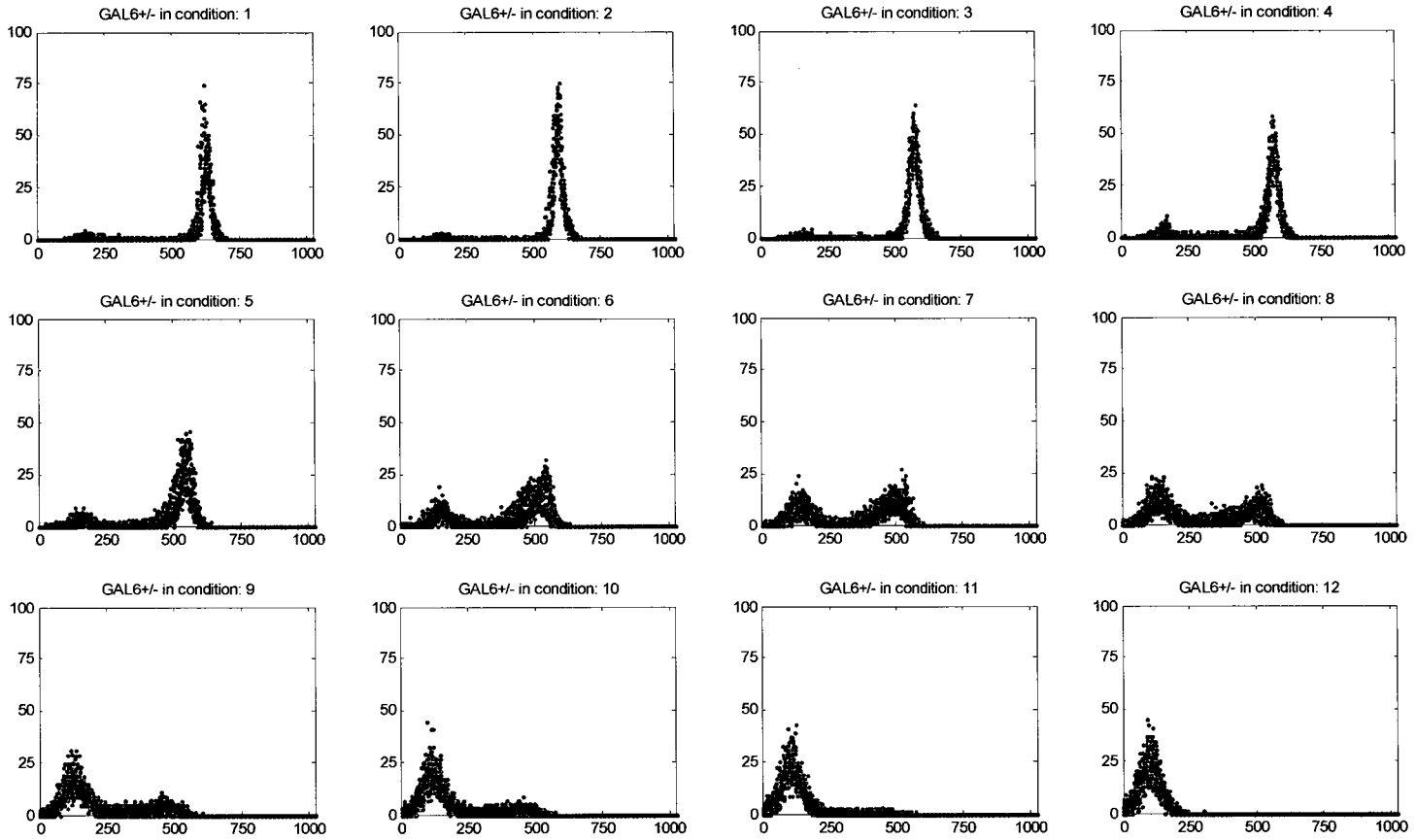

MIG1+/- in condition: 1
MIG1+/- in condition: 2
MIG1+/- in condition: 3
MIG1+/- in condition: 4
MIG1+/- in condition: 5
MIG1+/- in condition: 6
MIG1+/- in condition: 7
MIG1+/- in condition: 8
MIG1+/- in condition: 9
MIG1+/- in condition: 10
MIG1+/- in condition: 11
MIG1+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

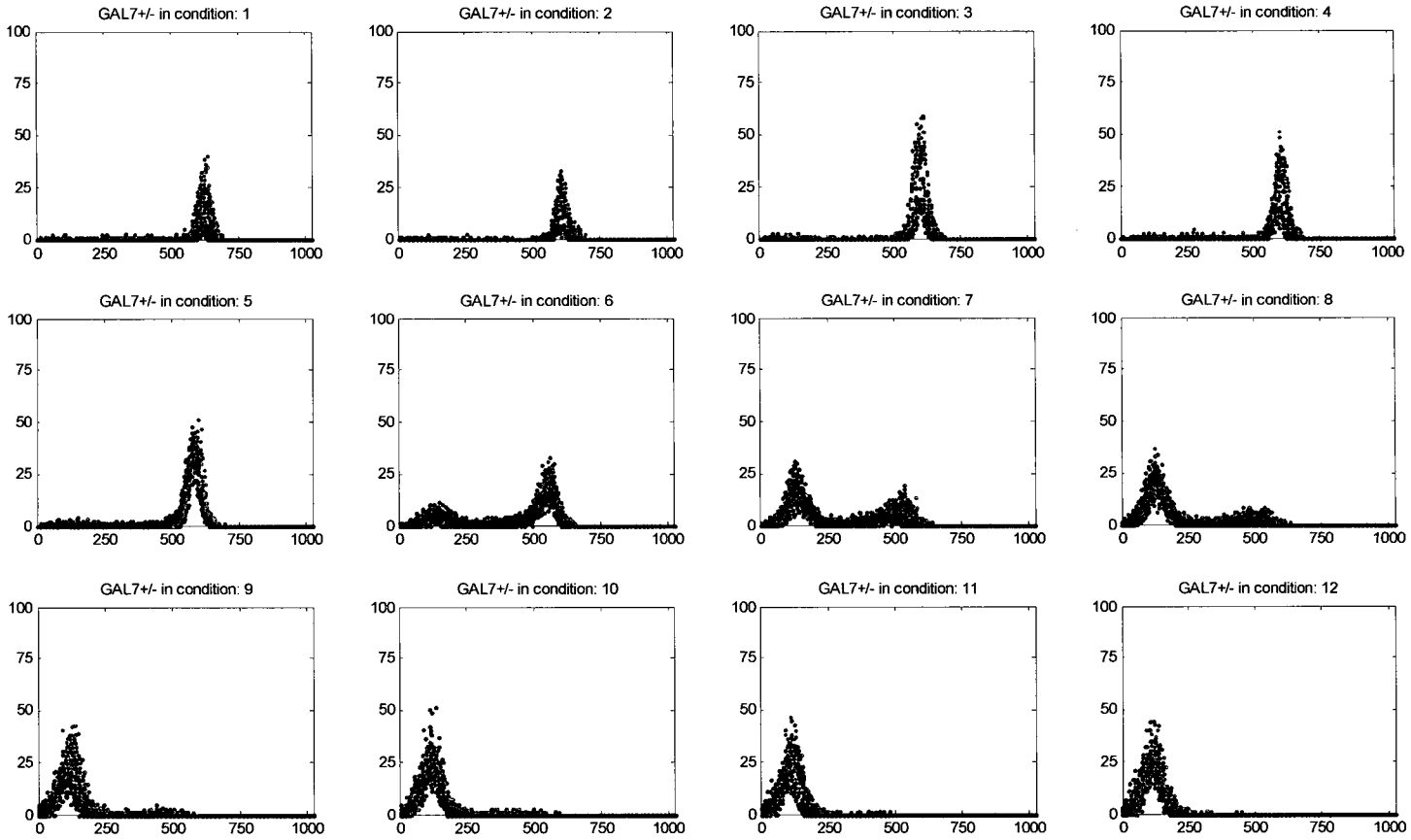

GAL80+/- in condition: 1
GAL80+/- in condition: 2
GAL80+/- in condition: 3
GAL80+/- in condition: 4
GAL80+/- in condition: 5
GAL80+/- in condition: 6
GAL80+/- in condition: 7
GAL80+/- in condition: 8
GAL80+/- in condition: 9
GAL80+/- in condition: 10
GAL80+/- in condition: 11
GAL80+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

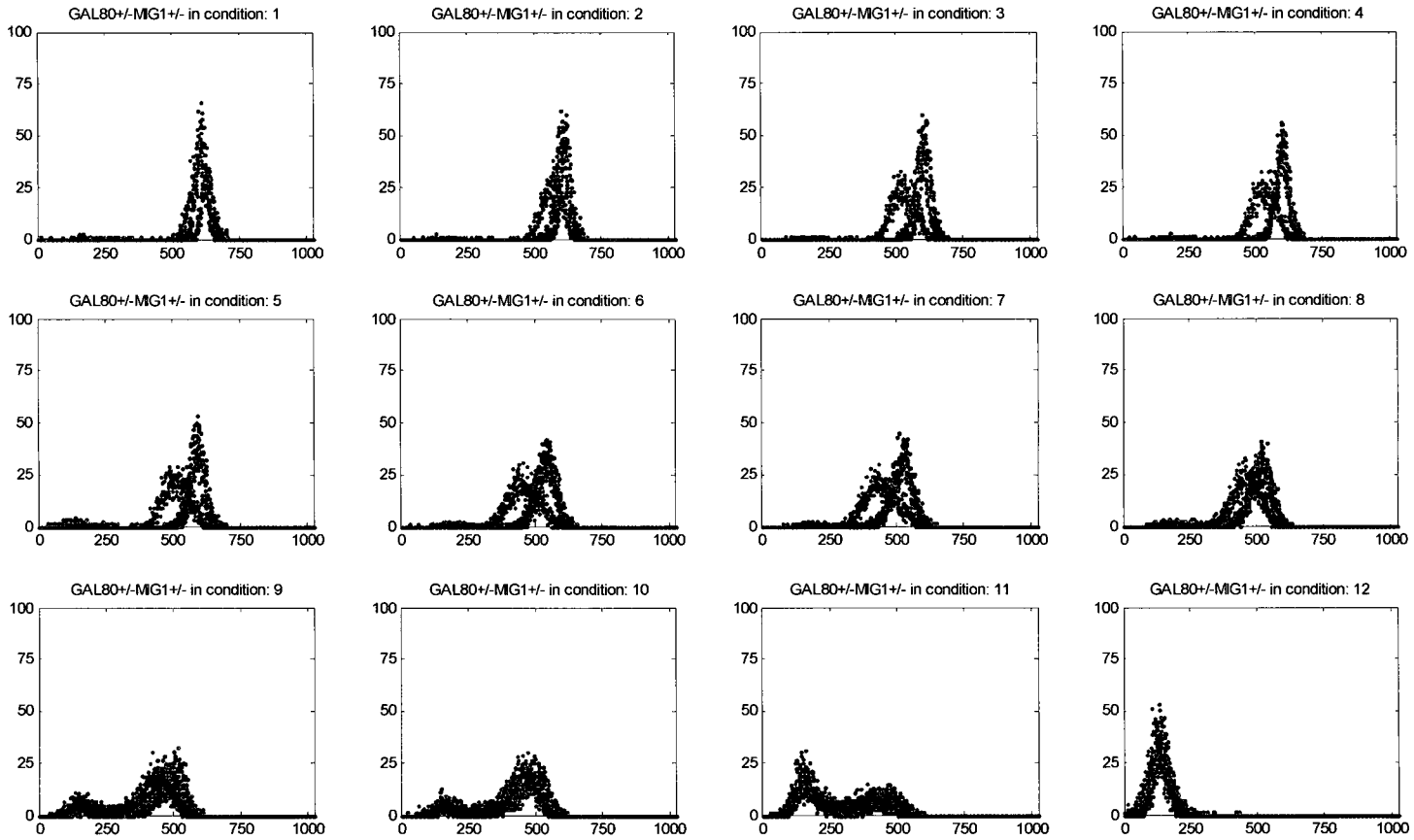

GAL4+/- in condition: 1
GAL4+/- in condition: 2
GAL4+/- in condition: 3
GAL4+/- in condition: 4
GAL4+/- in condition: 5
GAL4+/- in condition: 6
GAL4+/- in condition: 7
GAL4+/- in condition: 8
GAL4+/- in condition: 9
GAL4+/- in condition: 10
GAL4+/- in condition: 11
GAL4+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

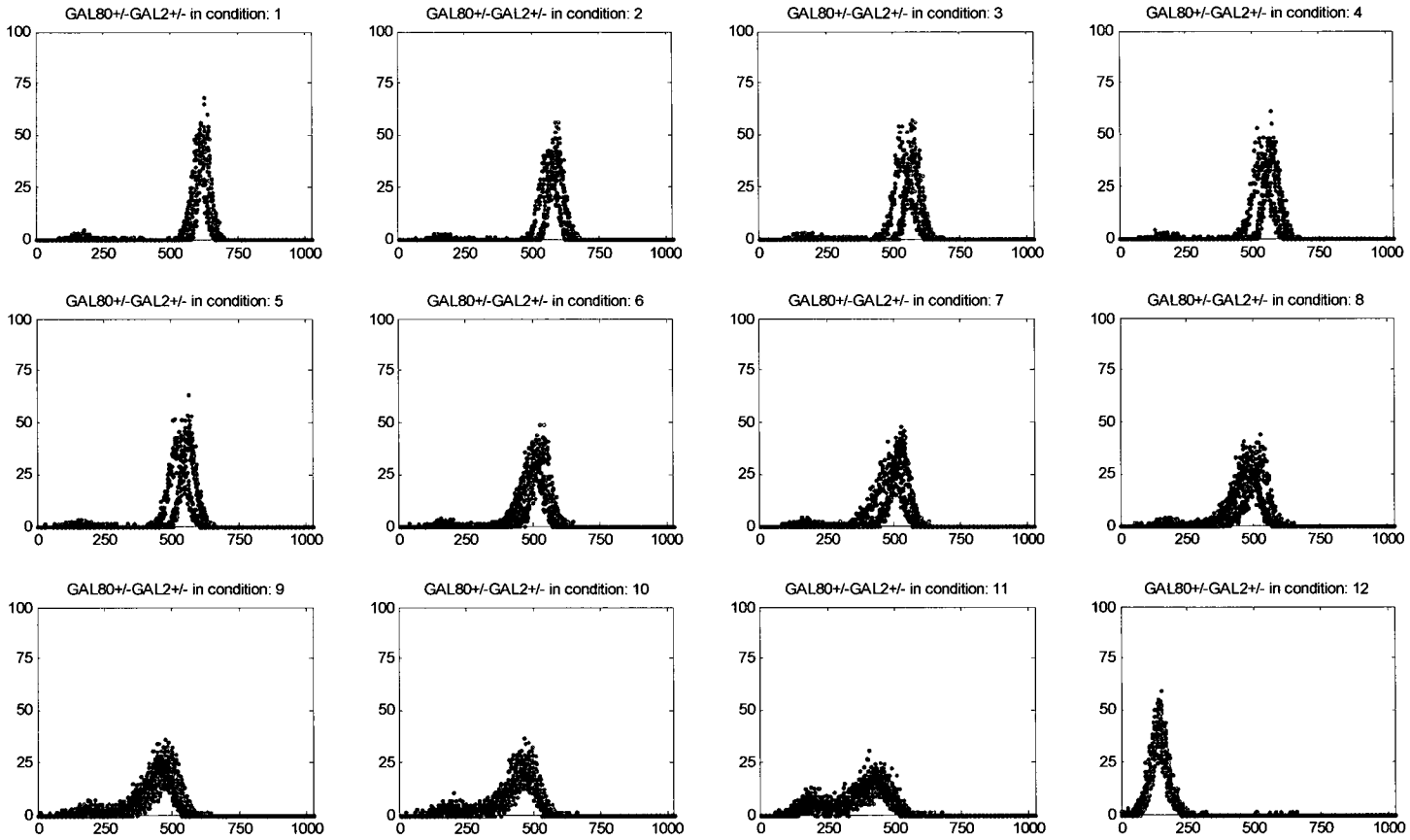

GAL2+/- in condition: 1
GAL2+/- in condition: 2
GAL2+/- in condition: 3
GAL2+/- in condition: 4
GAL2+/- in condition: 5
GAL2+/- in condition: 6
GAL2+/- in condition: 7
GAL2+/- in condition: 8
GAL2+/- in condition: 9
GAL2+/- in condition: 10
GAL2+/- in condition: 11
GAL2+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

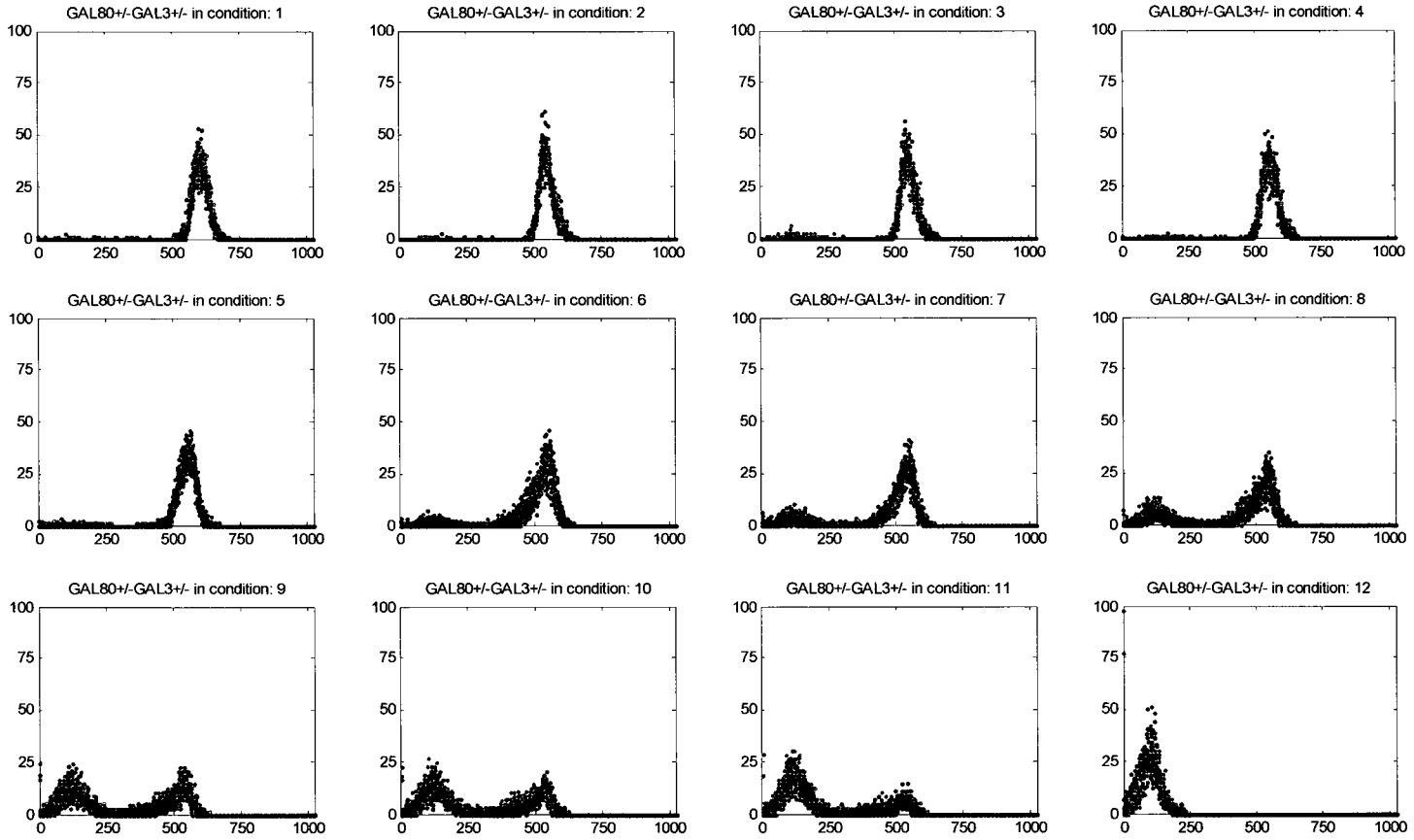

GAL3+/- in condition: 1
GAL3+/- in condition: 2
GAL3+/- in condition: 3
GAL3+/- in condition: 4
GAL3+/- in condition: 5
GAL3+/- in condition: 6
GAL3+/- in condition: 7
GAL3+/- in condition: 8
GAL3+/- in condition: 9
GAL3+/- in condition: 10
GAL3+/- in condition: 11
GAL3+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

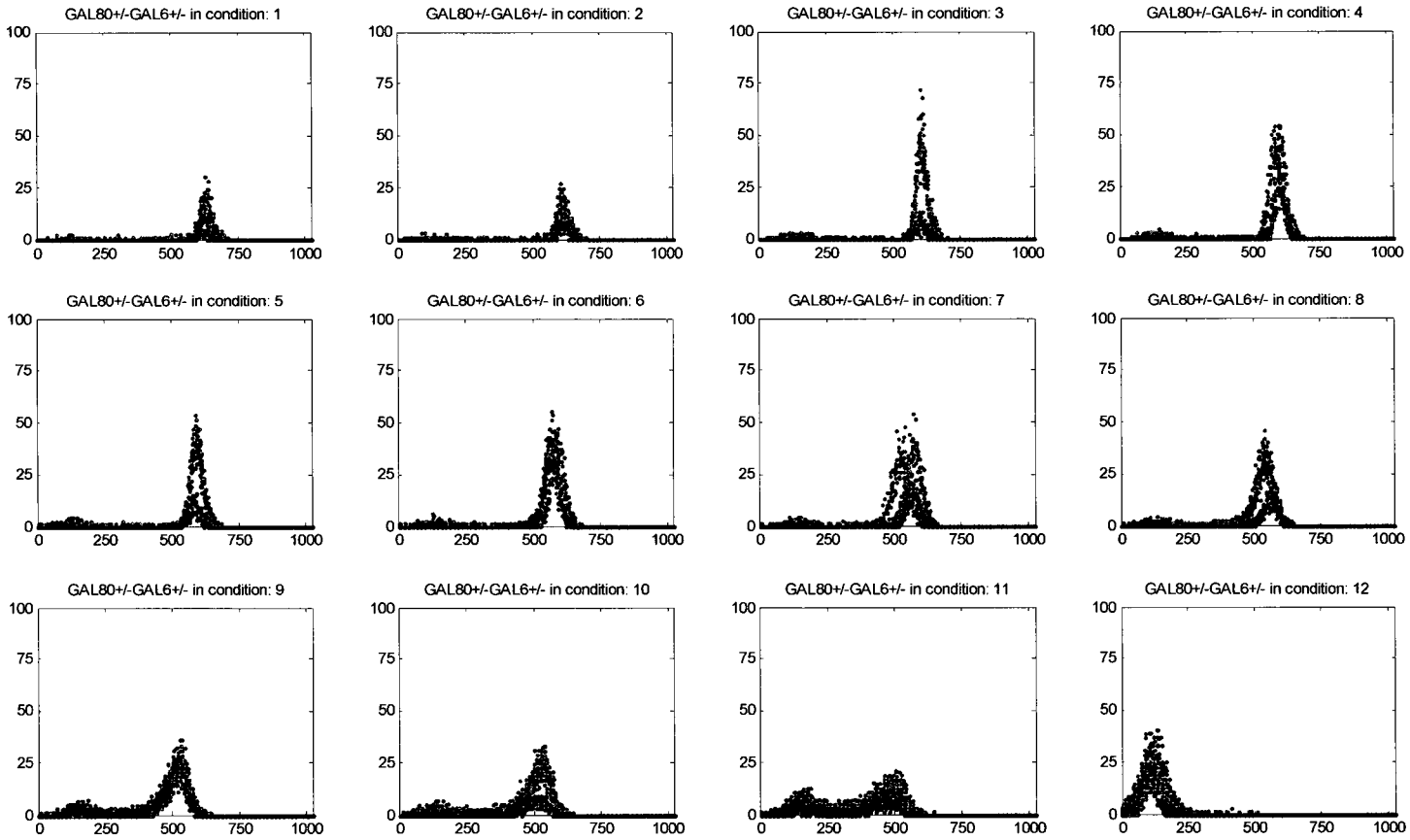

GAL6+/- in condition: 1
GAL6+/- in condition: 2
GAL6+/- in condition: 3
GAL6+/- in condition: 4
GAL6+/- in condition: 5
GAL6+/- in condition: 6
GAL6+/- in condition: 7
GAL6+/- in condition: 8
GAL6+/- in condition: 9
GAL6+/- in condition: 10
GAL6+/- in condition: 11
GAL6+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

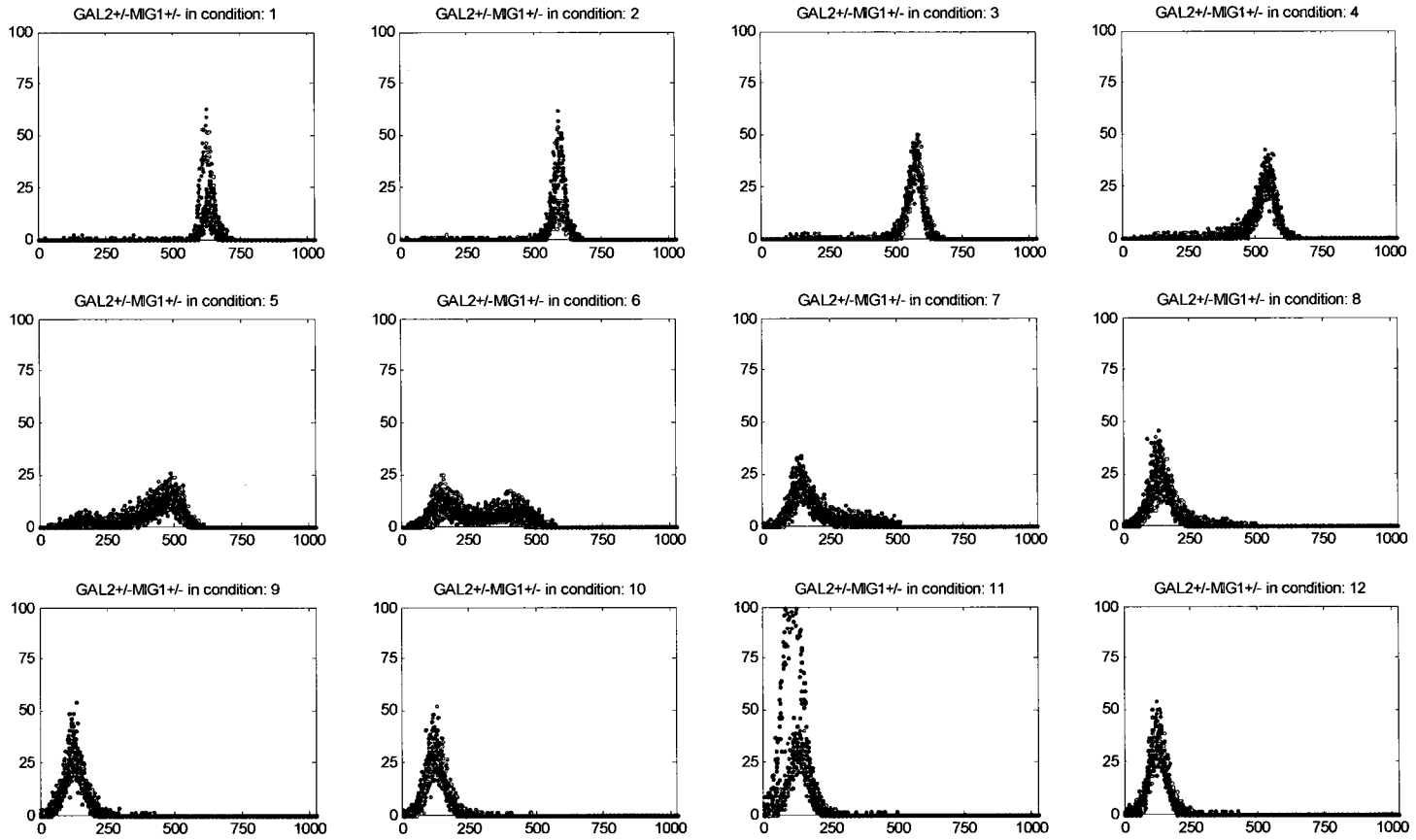

GAL7+/- in condition: 1
GAL7+/- in condition: 2
GAL7+/- in condition: 3
GAL7+/- in condition: 4
GAL7+/- in condition: 5
GAL7+/- in condition: 6
GAL7+/- in condition: 7
GAL7+/- in condition: 8
GAL7+/- in condition: 9
GAL7+/- in condition: 10
GAL7+/- in condition: 11
GAL7+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

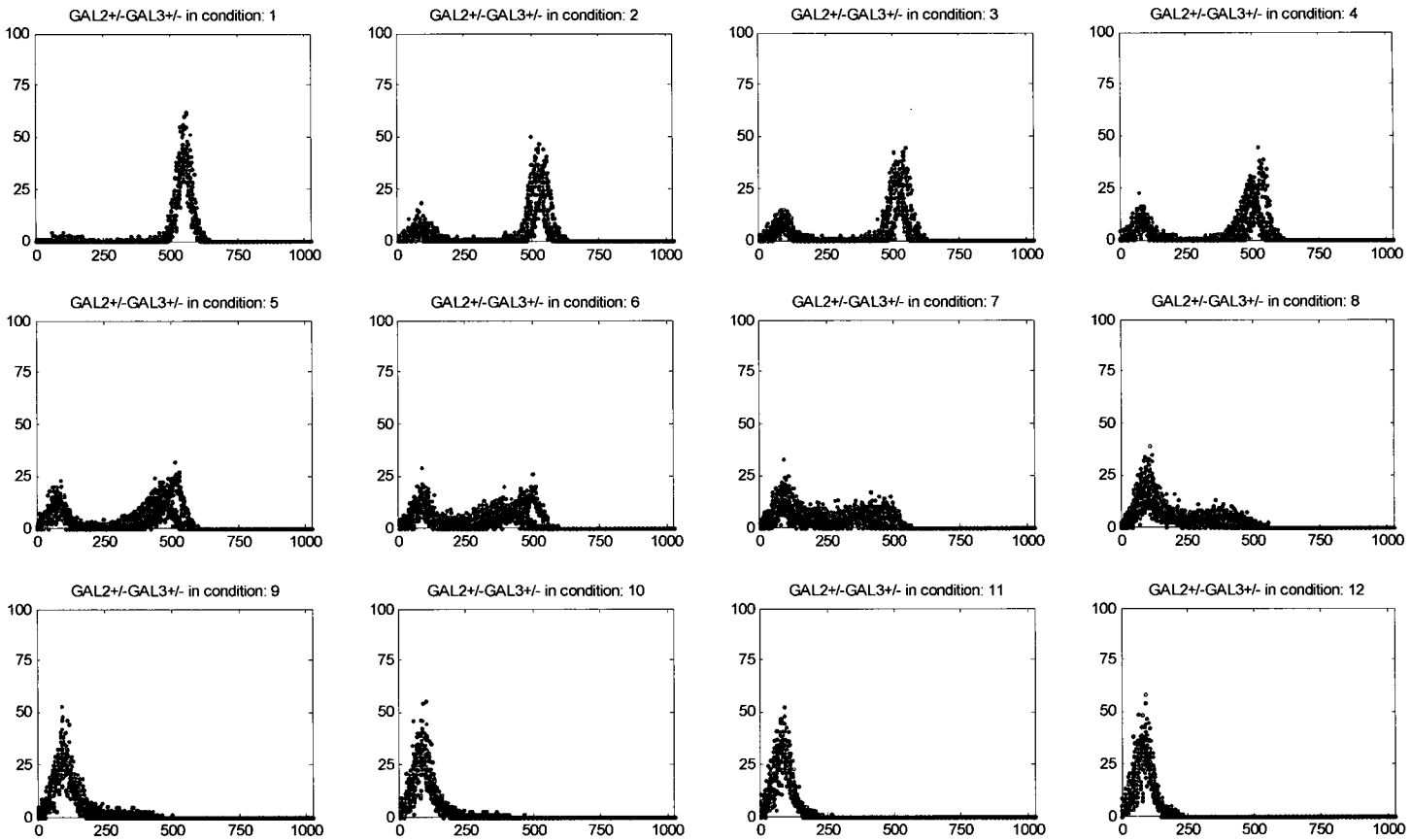

GAL80+/-MIG1+/- in condition: 1
GAL80+/-MIG1+/- in condition: 2
GAL80+/-MIG1+/- in condition: 3
GAL80+/-MIG1+/- in condition: 4
GAL80+/-MIG1+/- in condition: 5
GAL80+/-MIG1+/- in condition: 6
GAL80+/-MIG1+/- in condition: 7
GAL80+/-MIG1+/- in condition: 8
GAL80+/-MIG1+/- in condition: 9
GAL80+/-MIG1+/- in condition: 10
GAL80+/-MIG1+/- in condition: 11
GAL80+/-MIG1+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

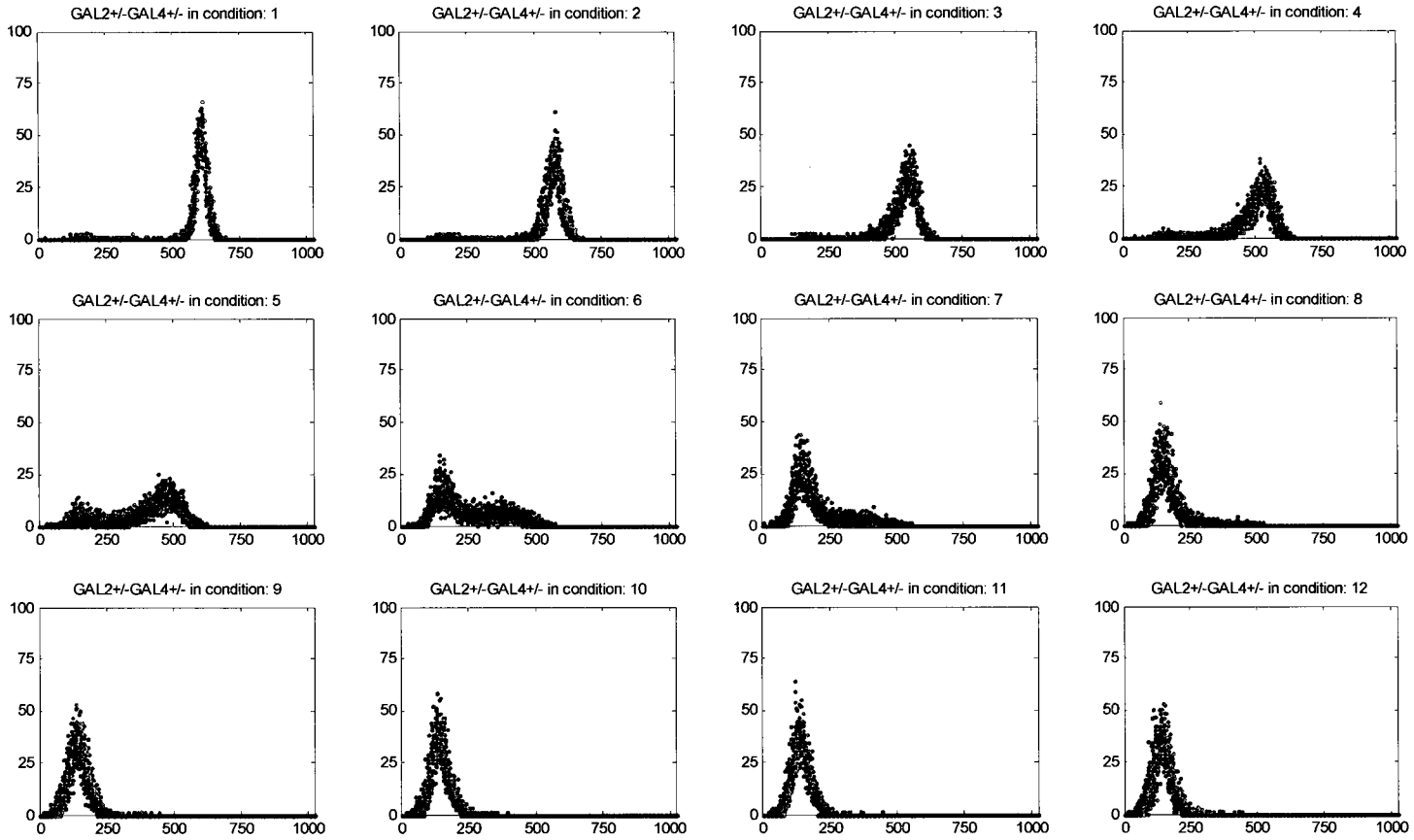

GAL80+/-GAL2+/- in condition: 1
GAL80+/-GAL2+/- in condition: 2
GAL80+/-GAL2+/- in condition: 3
GAL80+/-GAL2+/- in condition: 4
GAL80+/-GAL2+/- in condition: 5
GAL80+/-GAL2+/- in condition: 6
GAL80+/-GAL2+/- in condition: 7
GAL80+/-GAL2+/- in condition: 8
GAL80+/-GAL2+/- in condition: 9
GAL80+/-GAL2+/- in condition: 10
GAL80+/-GAL2+/- in condition: 11
GAL80+/-GAL2+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

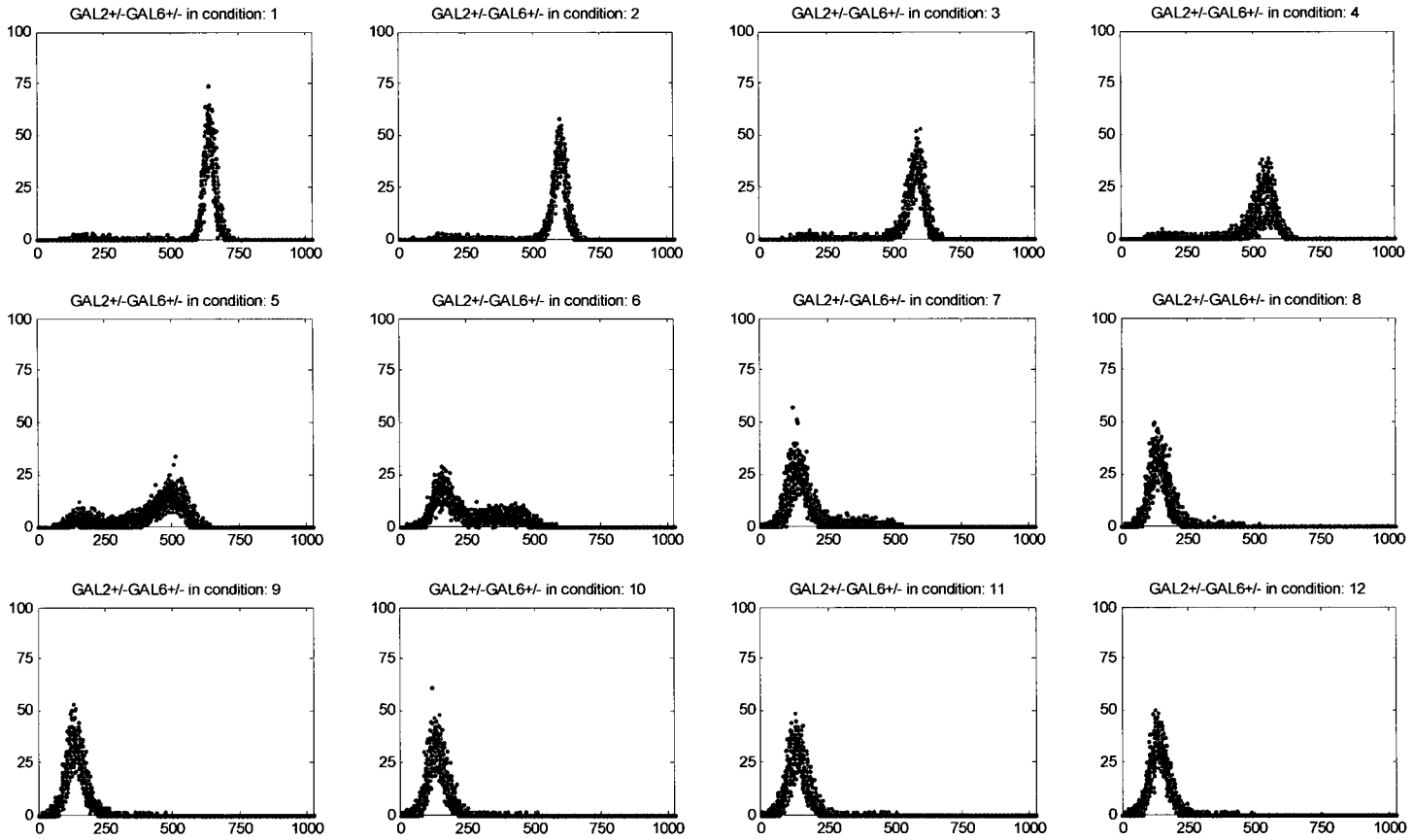

GAL80+/-GAL3+/- in condition: 1
GAL80+/-GAL3+/- in condition: 2
GAL80+/-GAL3+/- in condition: 3
GAL80+/-GAL3+/- in condition: 4
GAL80+/-GAL3+/- in condition: 5
GAL80+/-GAL3+/- in condition: 6
GAL80+/-GAL3+/- in condition: 7
GAL80+/-GAL3+/- in condition: 8
GAL80+/-GAL3+/- in condition: 9
GAL80+/-GAL3+/- in condition: 10
GAL80+/-GAL3+/- in condition: 11
GAL80+/-GAL3+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

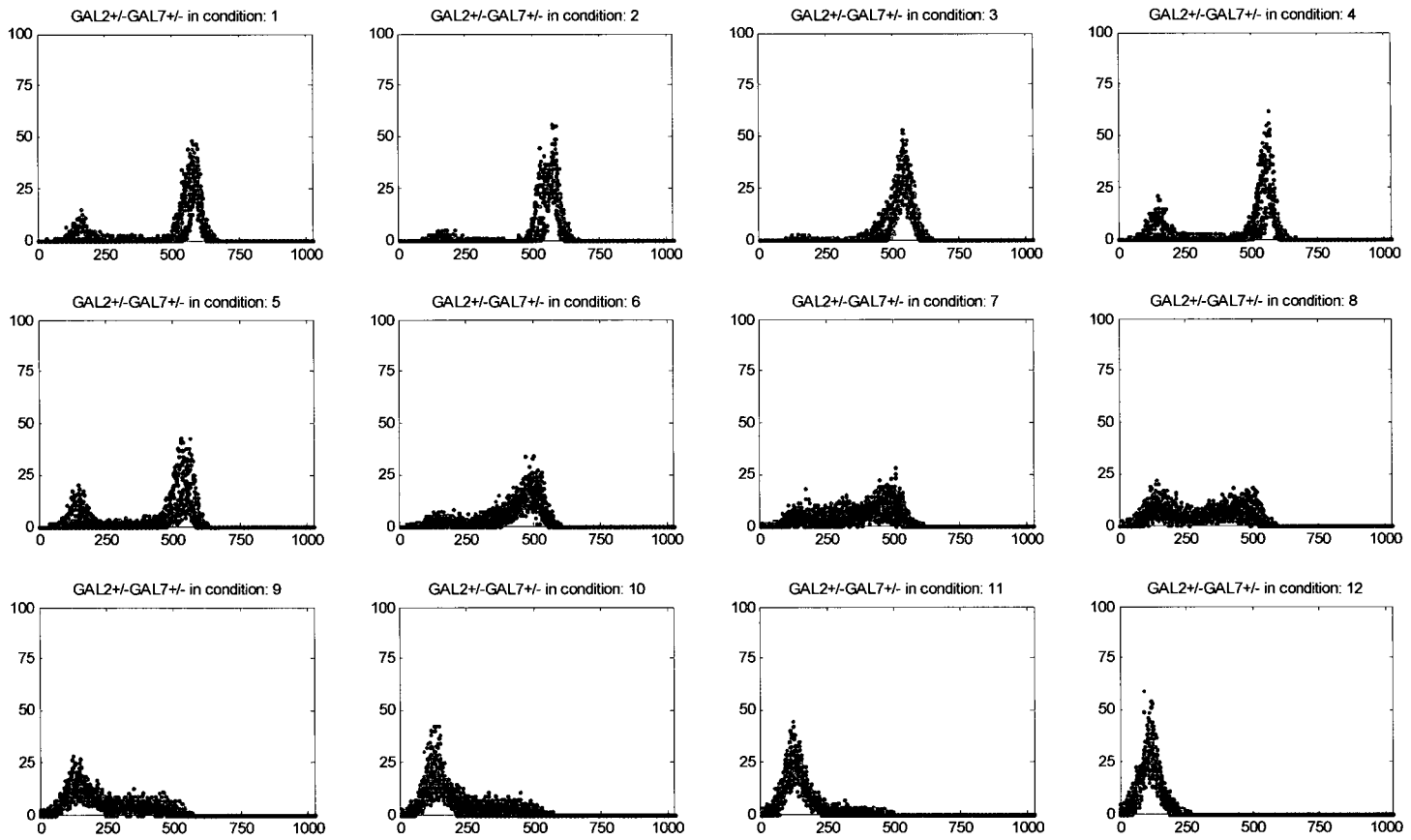

GAL80+/-GAL6+/- in condition: 1
GAL80+/-GAL6+/- in condition: 2
GAL80+/-GAL6+/- in condition: 3
GAL80+/-GAL6+/- in condition: 4
GAL80+/-GAL6+/- in condition: 5
GAL80+/-GAL6+/- in condition: 6
GAL80+/-GAL6+/- in condition: 7
GAL80+/-GAL6+/- in condition: 8
GAL80+/-GAL6+/- in condition: 9
GAL80+/-GAL6+/- in condition: 10
GAL80+/-GAL6+/- in condition: 11
GAL80+/-GAL6+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

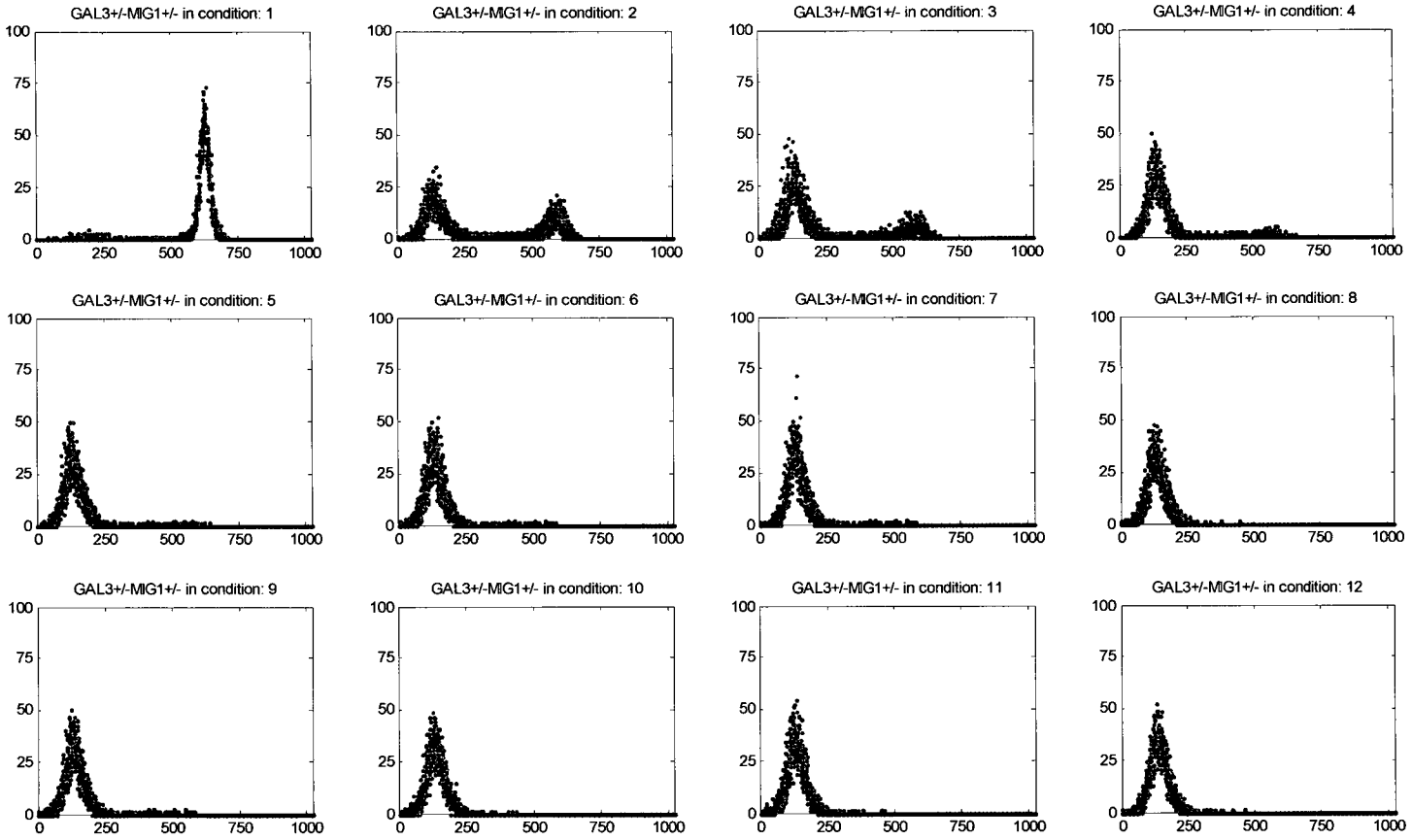

GAL2+/-MIG1+/- in condition: 1
GAL2+/-MIG1+/- in condition: 2
GAL2+/-MIG1+/- in condition: 3
GAL2+/-MIG1+/- in condition: 4
GAL2+/-MIG1+/- in condition: 5
GAL2+/-MIG1+/- in condition: 6
GAL2+/-MIG1+/- in condition: 7
GAL2+/-MIG1+/- in condition: 8
GAL2+/-MIG1+/- in condition: 9
GAL2+/-MIG1+/- in condition: 10
GAL2+/-MIG1+/- in condition: 11
GAL2+/-MIG1+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

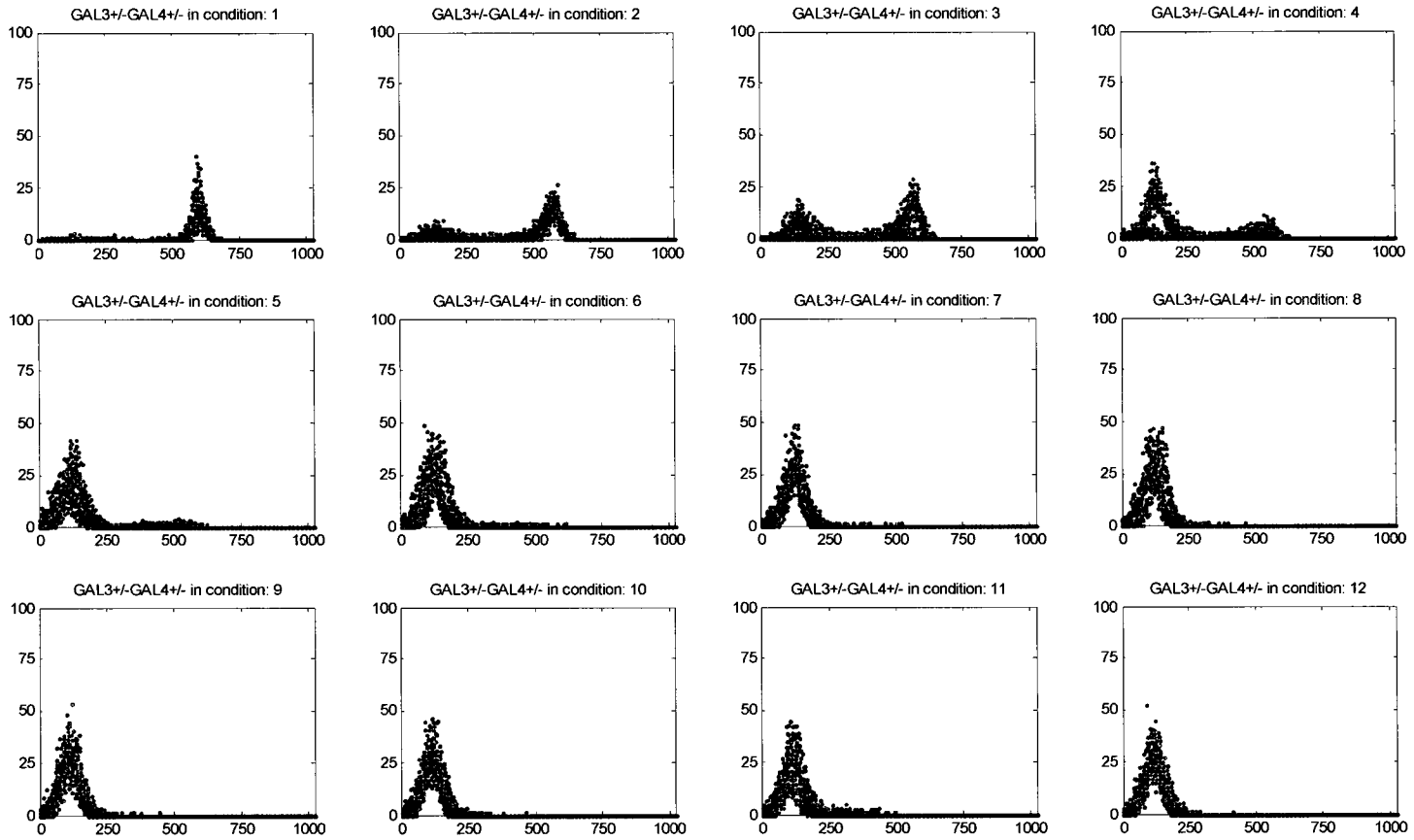

GAL2+/-GAL3+/- in condition: 1
GAL2+/-GAL3+/- in condition: 2
GAL2+/-GAL3+/- in condition: 3
GAL2+/-GAL3+/- in condition: 4
GAL2+/-GAL3+/- in condition: 5
GAL2+/-GAL3+/- in condition: 6
GAL2+/-GAL3+/- in condition: 7
GAL2+/-GAL3+/- in condition: 8
GAL2+/-GAL3+/- in condition: 9
GAL2+/-GAL3+/- in condition: 10
GAL2+/-GAL3+/- in condition: 11
GAL2+/-GAL3+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

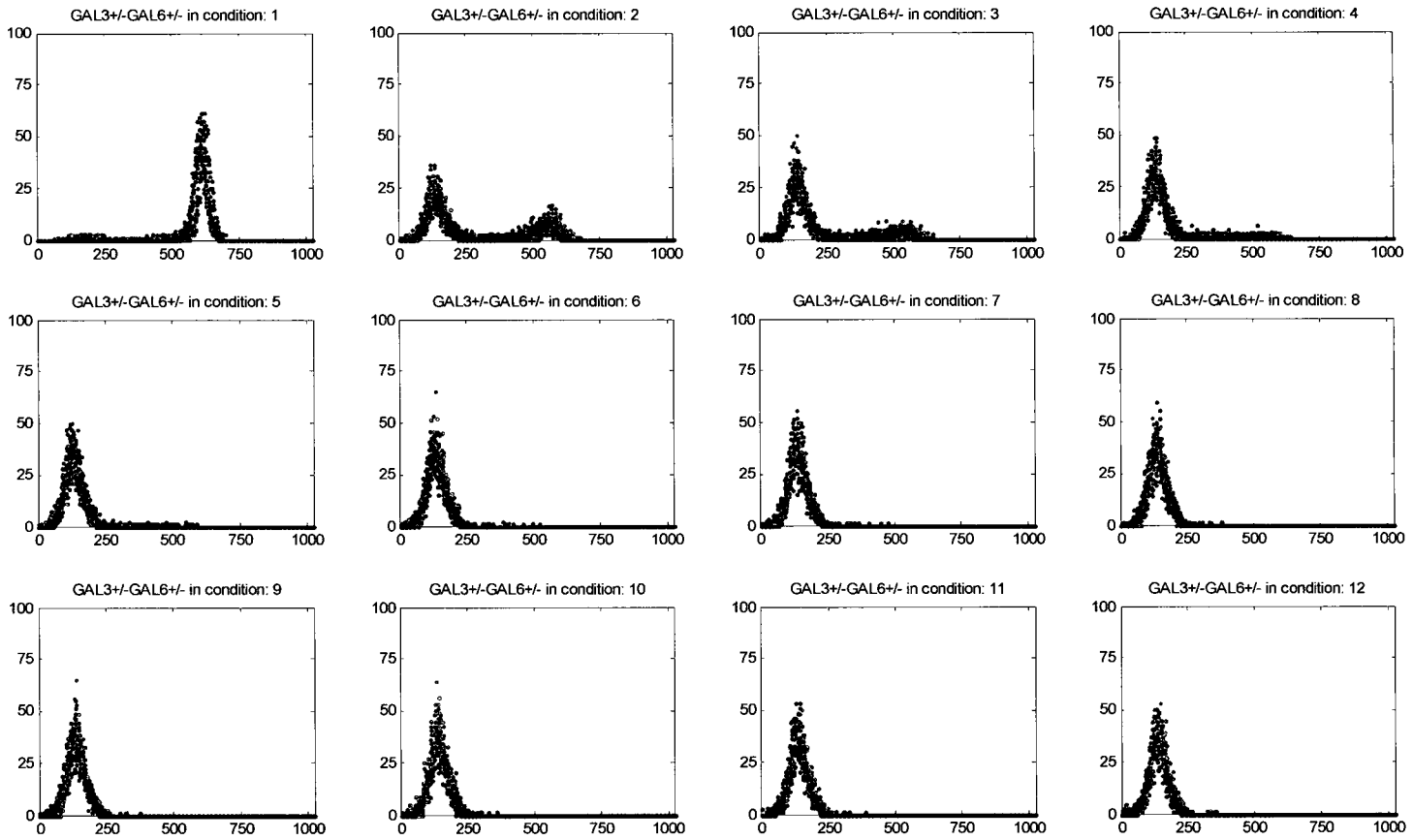

GAL2+/-GAL4+/- in condition: 1
GAL2+/-GAL4+/- in condition: 2
GAL2+/-GAL4+/- in condition: 3
GAL2+/-GAL4+/- in condition: 4
GAL2+/-GAL4+/- in condition: 5
GAL2+/-GAL4+/- in condition: 6
GAL2+/-GAL4+/- in condition: 7
GAL2+/-GAL4+/- in condition: 8
GAL2+/-GAL4+/- in condition: 9
GAL2+/-GAL4+/- in condition: 10
GAL2+/-GAL4+/- in condition: 11
GAL2+/-GAL4+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

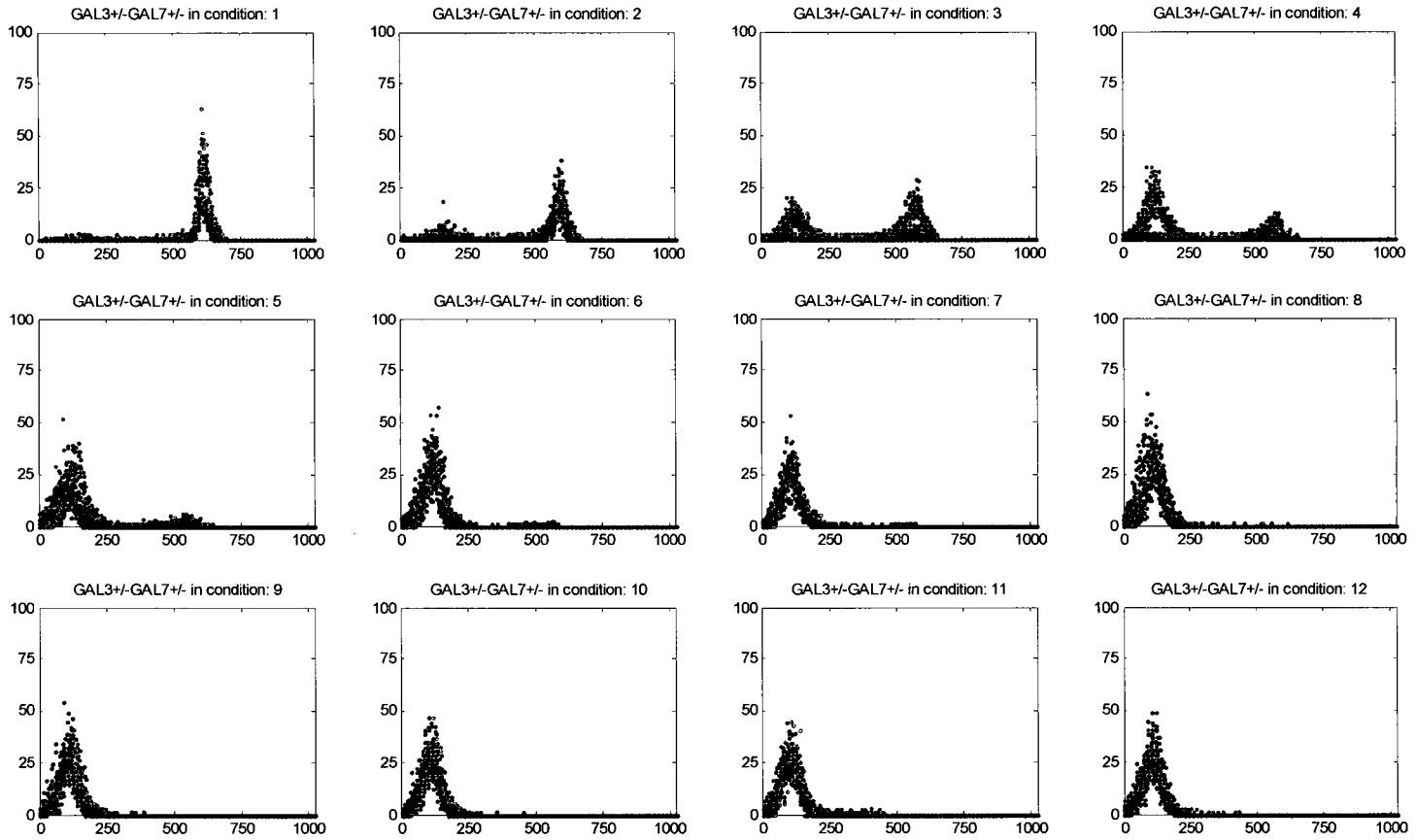

GAL2+/-GAL6+/- in condition: 1
GAL2+/-GAL6+/- in condition: 2
GAL2+/-GAL6+/- in condition: 3
GAL2+/-GAL6+/- in condition: 4
GAL2+/-GAL6+/- in condition: 5
GAL2+/-GAL6+/- in condition: 6
GAL2+/-GAL6+/- in condition: 7
GAL2+/-GAL6+/- in condition: 8
GAL2+/-GAL6+/- in condition: 9
GAL2+/-GAL6+/- in condition: 10
GAL2+/-GAL6+/- in condition: 11
GAL2+/-GAL6+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

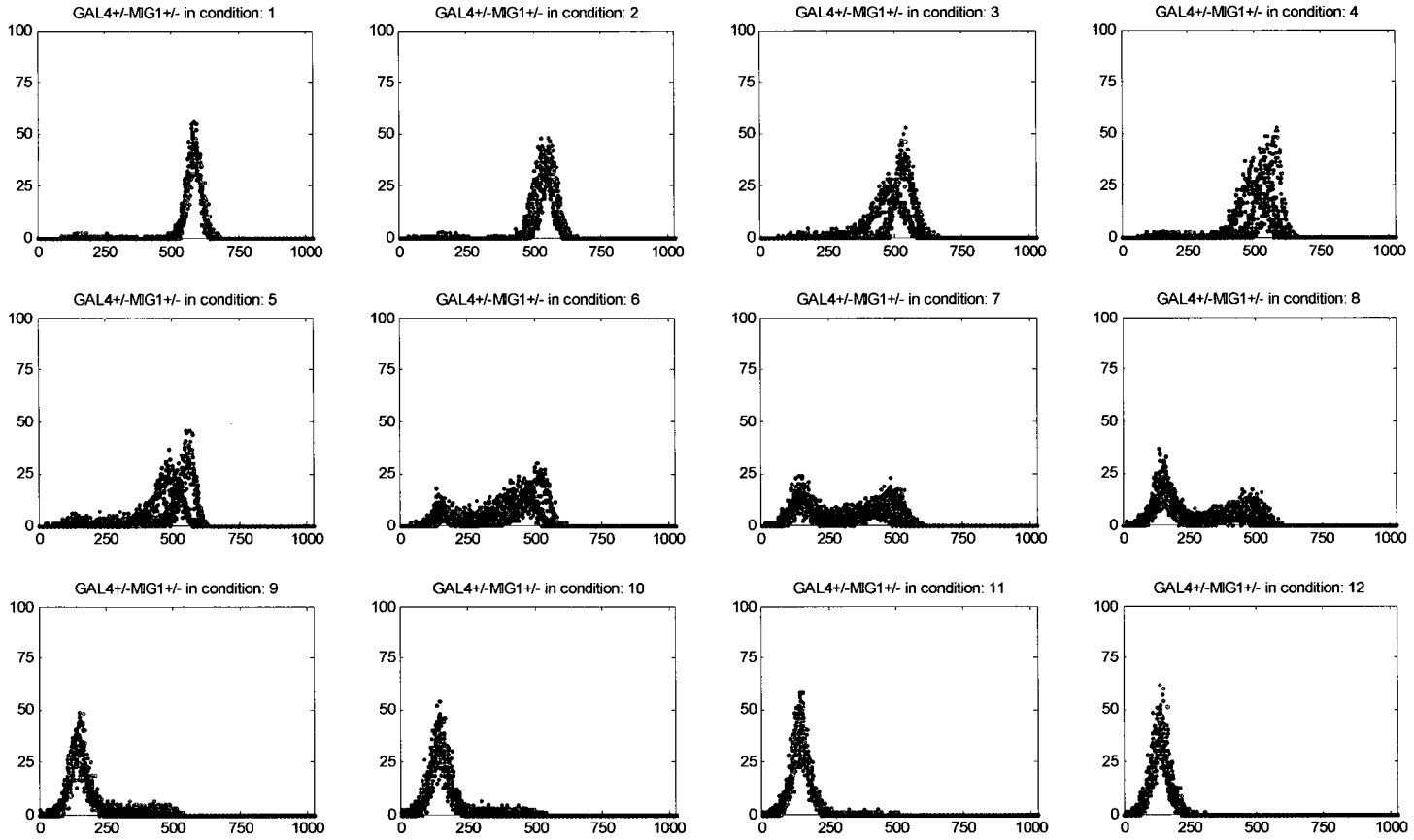

GAL2+/-GAL7+/- in condition: 1
GAL2+/-GAL7+/- in condition: 2
GAL2+/-GAL7+/- in condition: 3
GAL2+/-GAL7+/- in condition: 4
GAL2+/-GAL7+/- in condition: 5
GAL2+/-GAL7+/- in condition: 6
GAL2+/-GAL7+/- in condition: 7
GAL2+/-GAL7+/- in condition: 8
GAL2+/-GAL7+/- in condition: 9
GAL2+/-GAL7+/- in condition: 10
GAL2+/-GAL7+/- in condition: 11
GAL2+/-GAL7+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

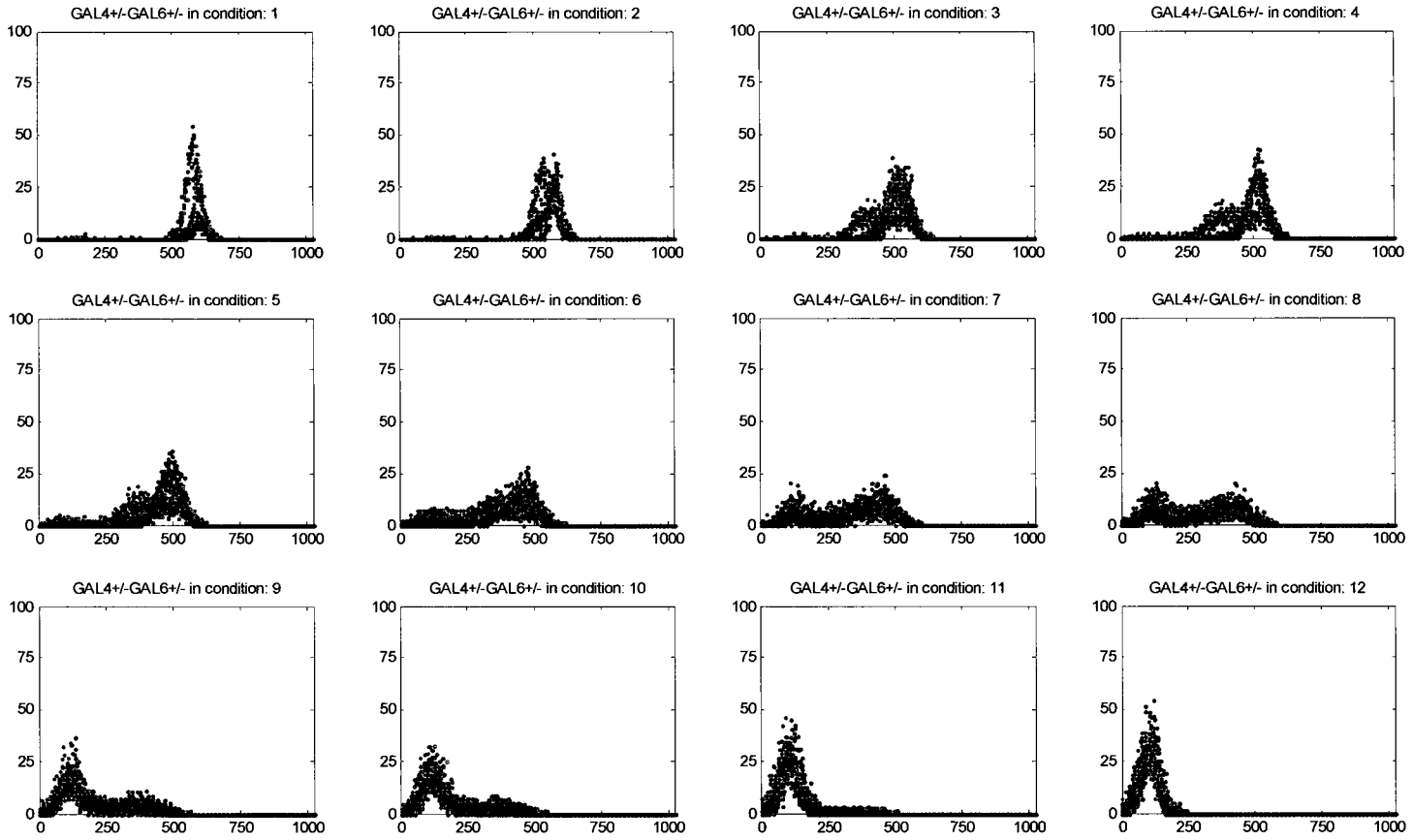

GAL3+/-MIG1+/- in condition: 1
GAL3+/-MIG1+/- in condition: 2
GAL3+/-MIG1+/- in condition: 3
GAL3+/-MIG1+/- in condition: 4
GAL3+/-MIG1+/- in condition: 5
GAL3+/-MIG1+/- in condition: 6
GAL3+/-MIG1+/- in condition: 7
GAL3+/-MIG1+/- in condition: 8
GAL3+/-MIG1+/- in condition: 9
GAL3+/-MIG1+/- in condition: 10
GAL3+/-MIG1+/- in condition: 11
GAL3+/-MIG1+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

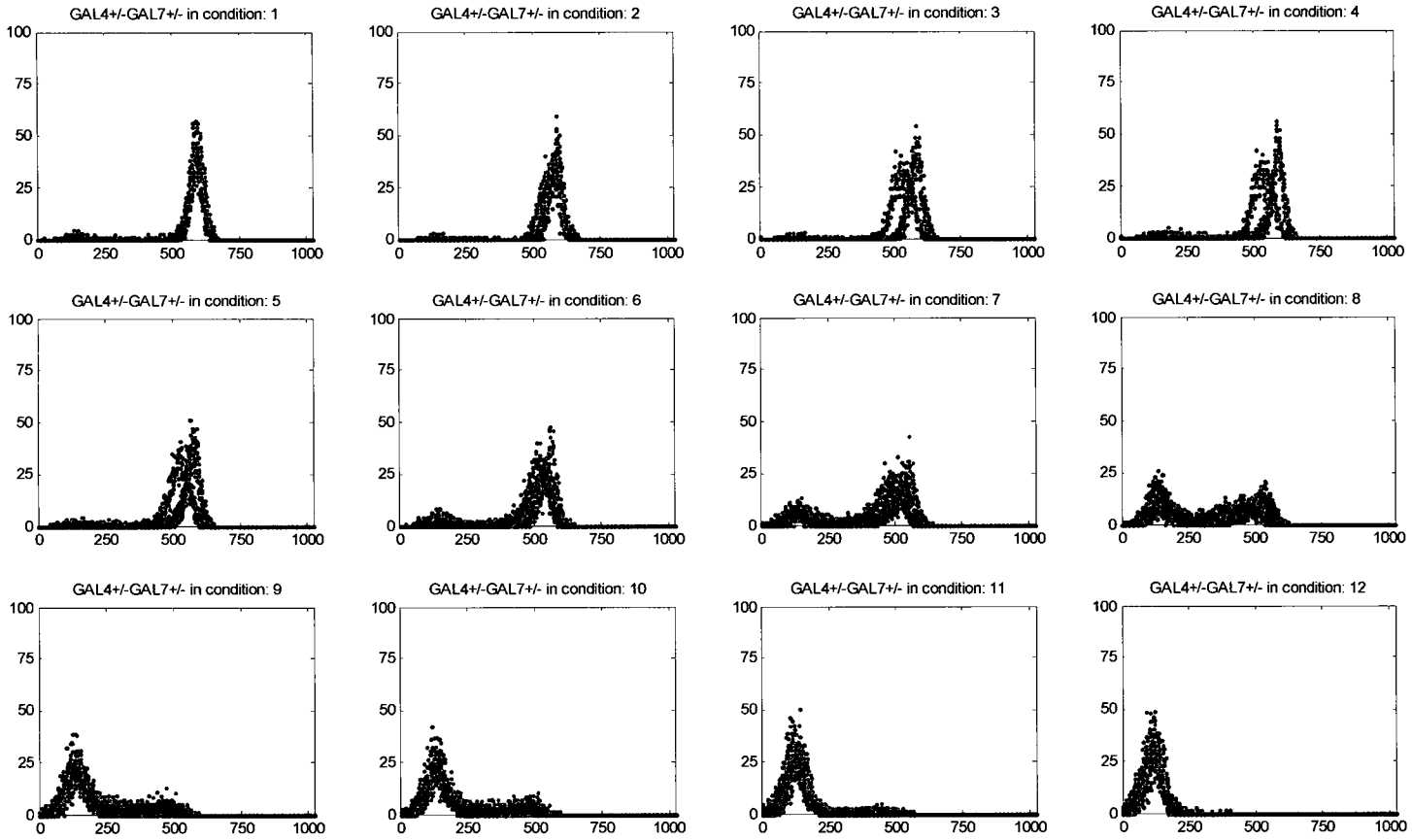

GAL3+/-GAL4+/- in condition: 1
GAL3+/-GAL4+/- in condition: 2
GAL3+/-GAL4+/- in condition: 3
GAL3+/-GAL4+/- in condition: 4
GAL3+/-GAL4+/- in condition: 5
GAL3+/-GAL4+/- in condition: 6
GAL3+/-GAL4+/- in condition: 7
GAL3+/-GAL4+/- in condition: 8
GAL3+/-GAL4+/- in condition: 9
GAL3+/-GAL4+/- in condition: 10
GAL3+/-GAL4+/- in condition: 11
GAL3+/-GAL4+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

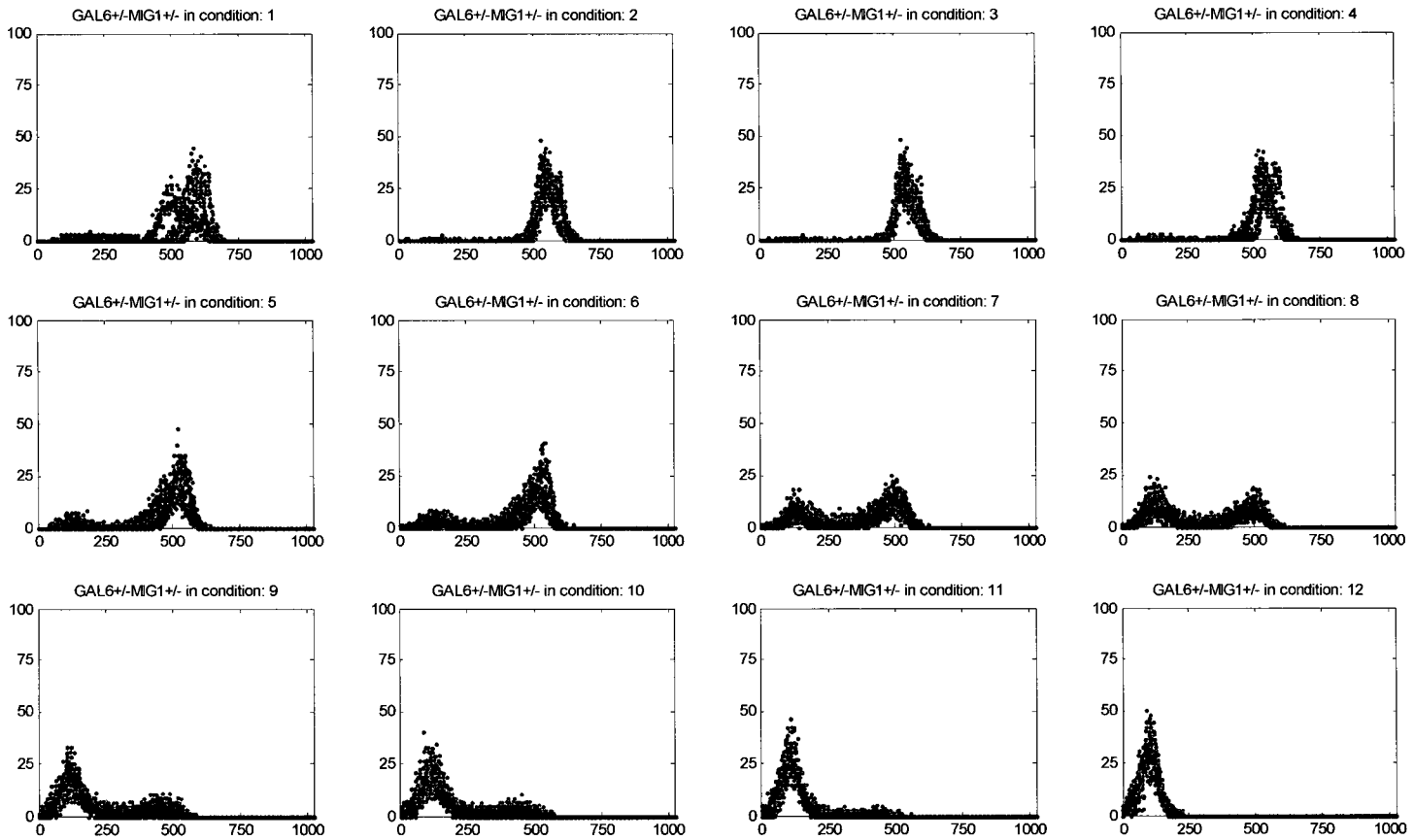

GAL3+/-GAL6+/- in condition: 1
GAL3+/-GAL6+/- in condition: 2
GAL3+/-GAL6+/- in condition: 3
GAL3+/-GAL6+/- in condition: 4
GAL3+/-GAL6+/- in condition: 5
GAL3+/-GAL6+/- in condition: 6
GAL3+/-GAL6+/- in condition: 7
GAL3+/-GAL6+/- in condition: 8
GAL3+/-GAL6+/- in condition: 9
GAL3+/-GAL6+/- in condition: 10
GAL3+/-GAL6+/- in condition: 11
GAL3+/-GAL6+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

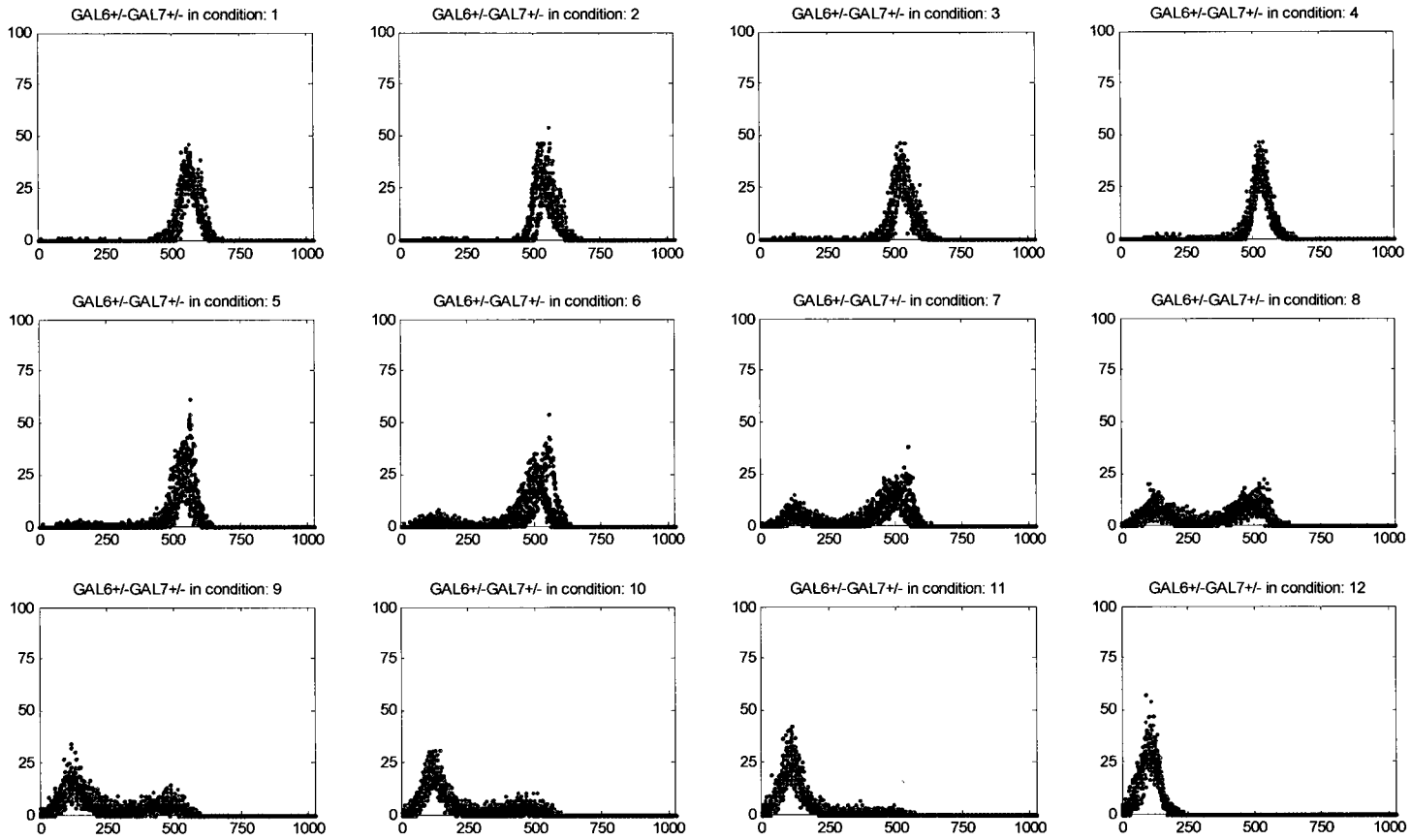

GAL3+/-GAL7+/- in condition: 1
GAL3+/-GAL7+/- in condition: 2
GAL3+/-GAL7+/- in condition: 3
GAL3+/-GAL7+/- in condition: 4
GAL3+/-GAL7+/- in condition: 5
GAL3+/-GAL7+/- in condition: 6
GAL3+/-GAL7+/- in condition: 7
GAL3+/-GAL7+/- in condition: 8
GAL3+/-GAL7+/- in condition: 9
GAL3+/-GAL7+/- in condition: 10
GAL3+/-GAL7+/- in condition: 11
GAL3+/-GAL7+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000

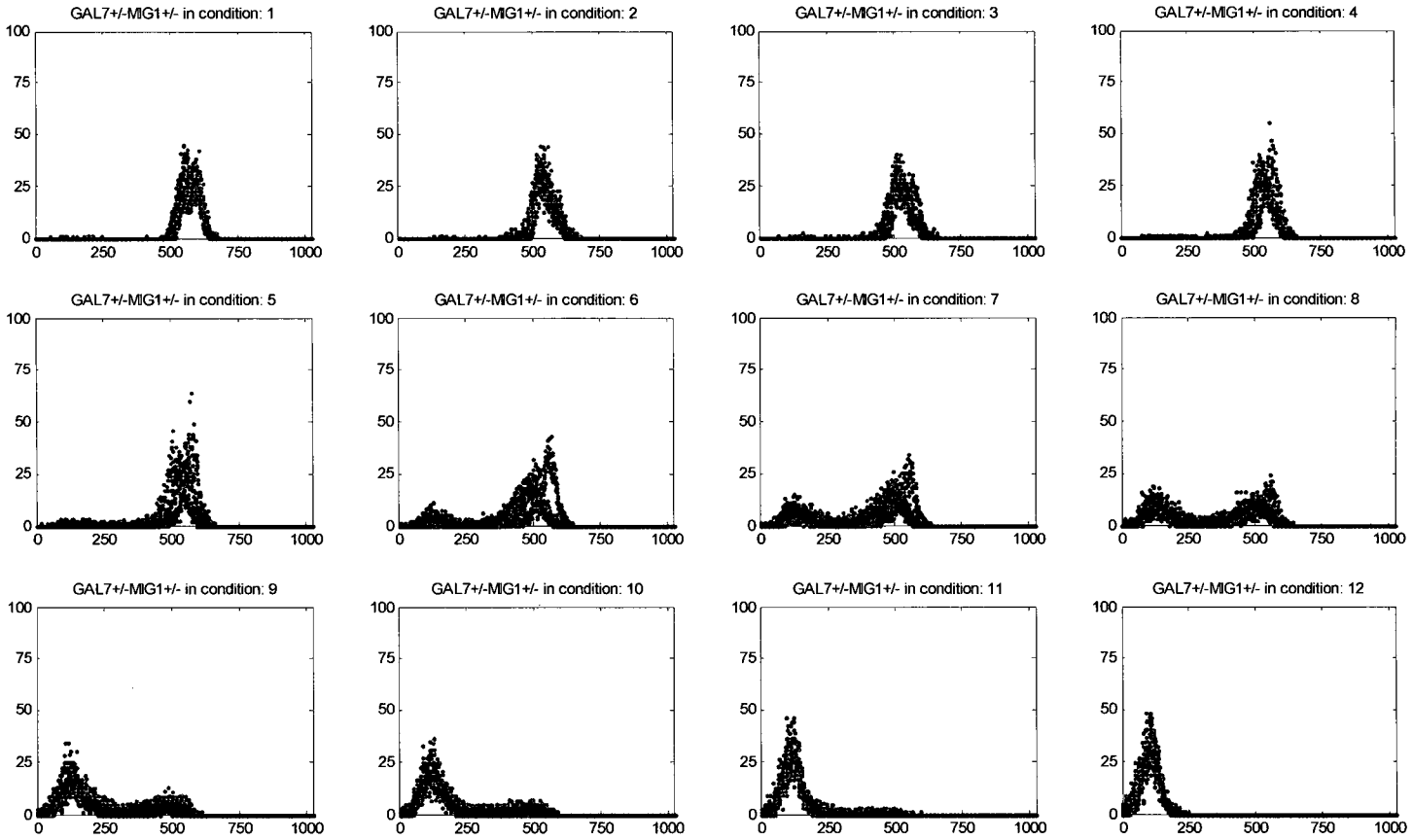

GAL4+/-MIG1+/- in condition: 1
GAL4+/-MIG1+/- in condition: 2
GAL4+/-MIG1+/- in condition: 3
GAL4+/-MIG1+/- in condition: 4
GAL4+/-MIG1+/- in condition: 5
GAL4+/-MIG1+/- in condition: 6
GAL4+/-MIG1+/- in condition: 7
GAL4+/-MIG1+/- in condition: 8
GAL4+/-MIG1+/- in condition: 9
GAL4+/-MIG1+/- in condition: 10
GAL4+/-MIG1+/- in condition: 11
GAL4+/-MIG1+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000


GAL4+/-GAL6+/- in condition: 1
GAL4+/-GAL6+/- in condition: 2
GAL4+/-GAL6+/- in condition: 3
GAL4+/-GAL6+/- in condition: 4
GAL4+/-GAL6+/- in condition: 5
GAL4+/-GAL6+/- in condition: 6
GAL4+/-GAL6+/- in condition: 7
GAL4+/-GAL6+/- in condition: 8
GAL4+/-GAL6+/- in condition: 9
GAL4+/-GAL6+/- in condition: 10
GAL4+/-GAL6+/- in condition: 11
GAL4+/-GAL6+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000


GAL4+/-GAL7+/- in condition: 1
GAL4+/-GAL7+/- in condition: 2
GAL4+/-GAL7+/- in condition: 3
GAL4+/-GAL7+/- in condition: 4
GAL4+/-GAL7+/- in condition: 5
GAL4+/-GAL7+/- in condition: 6
GAL4+/-GAL7+/- in condition: 7
GAL4+/-GAL7+/- in condition: 8
GAL4+/-GAL7+/- in condition: 9
GAL4+/-GAL7+/- in condition: 10
GAL4+/-GAL7+/- in condition: 11
GAL4+/-GAL7+/- in condition: 12
100
75
50
25
0
0
250
500
750
1000


GAL6+/-MIG1+/- in condition: 1
GAL6+/-MIG1+/- in condition: 2
GAL6+/-MIG1+/- in condition: 3
GAL6+/-MIG1+/- in condition: 4
GAL6+/-MIG1+/- in condition: 5
GAL6+/-MIG1+/- in condition: 6
GAL6+/-MIG1+/- in condition: 7
GAL6+/-MIG1+/- in condition: 8
GAL6+/-MIG1+/- in condition: 9
GAL6+/-MIG1+/- in condition: 10
GAL6+/-MIG1+/- in condition: 11
GAL6+/-MIG1+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000


GAL6+/-GAL7+/- in condition: 1
GAL6+/-GAL7+/- in condition: 2
GAL6+/-GAL7+/- in condition: 3
GAL6+/-GAL7+/- in condition: 4
GAL6+/-GAL7+/- in condition: 5
GAL6+/-GAL7+/- in condition: 6
GAL6+/-GAL7+/- in condition: 7
GAL6+/-GAL7+/- in condition: 8
GAL6+/-GAL7+/- in condition: 9
GAL6+/-GAL7+/- in condition: 10
GAL6+/-GAL7+/- in condition: 11
GAL6+/-GAL7+/- in condition: 12
0
25
50
75
100
0
250
500
750
1000


GAL7+/-MIG1+/- in condition: 1
GAL7+/-MIG1+/- in condition: 2
GAL7+/-MIG1+/- in condition: 3
GAL7+/-MIG1+/- in condition: 4
GAL7+/-MIG1+/- in condition: 5
GAL7+/-MIG1+/- in condition: 6
GAL7+/-MIG1+/- in condition: 7
GAL7+/-MIG1+/- in condition: 8
GAL7+/-MIG1+/- in condition: 9
GAL7+/-MIG1+/- in condition: 10
GAL7+/-MIG1+/- in condition: 11
GAL7+/-MIG1+/- in condition: 12
100
75
50
25
0
0
250
500
750
1000

Fitness Scores

Structure ID	Fitness Score wrt knock-out strain (GAL3-/-, GAL4-/-, GAL80-/-)	Fitness Score wrt knock-down strains: (GAL3, GAL4, GAL80 combinations)
1	1.84	7.01
2	2.11	5.57
3	2.49	4.69
4	2.69	6.33
5	4.49	6.41
6	7.26	6.38
6	4.74	7.41
7	7.60	6.99
8	6.41	5.83
9	4.67	7.23
10	4.92	7.28
	1.99	5.77
11	3.40	5.87
12	3.49	9.62
13	3.51	6.33
14	1.49	6.32
	6.66	5.81
15	3.79	5.76
16	3.80	6.06
17	3.88	5.62
18	3.91	7.39
19	3.92	5.58
20	4.07	6.94
21	4.14	10.18
22	4.23	8.46
23	4.35	5.91
24	5.67	7.99
	3.41	5.52
25	4.49	7.18
26	4.49	5.40
27	4.77	5.70
28	6.44	5.93
	4.67	6.96
	4.49	5.24
29	5.46	5.74
30	5.59	7.62
	3.47	6.31
	6.74	6.28
31	6.81	9.62
	3.96	9.20
32	5.67	6.58
33	5.77	5.73
34	4.49	7.49

	4.43 7.43	6.19 7.83
35	5.85	6.10
36	5.90	6.66
37	6.15	6.85
38	6.74	6.03
39	7.37	6.29
40	8.47	5.98
41	4.08 8.58 5.47 9.08	8.40 8.15 5.72 6.30

Table S.10: Network structures and their respective fitness score. The ranking is based on fitness score from the knock-out strains. In some cases, two or more parameter sets generate the same structure.

Parameter Value

ID	Tcoefficients: T21,	T23,	T24,	T34,	T32,	T42,	T43,	T53	Basal Level: B1,	B2,	B3,	B4,	B5
1	5.8E+00	4.9E+02	0.0E+00	-1.4E+03	6.0E+01	-1.2E+01	-4.6E+02	1.7E+05	4.5E-01	1.2E+01	3.2E+01	8.5E-01	7.3E-01
2	-2.0E+01	3.8E+03	3.3E+01	1.2E+01	-8.2E+01	2.3E+03	-2.2E+04	2.0E+03	-5.4E-01	-5.0E+00	1.9E+01	2.8E+02	7.6E-01
3	4.5E+03	-1.7E+01	-7.4E+02	-1.5E+02	4.1E+01	-8.1E+02	8.9E+01	1.5E+00	1.7E+01	-8.8E+00	-6.0E-01	4.1E+00	7.8E-01
4	2.5E+02	-1.9E+03	-2.1E+01	-2.7E+03	-1.0E-03	3.7E+04	0.0E+00	-3.0E-01	-6.5E+00	3.4E+00	-2.0E+03	2.1E+02	-4.2E-01
5	2.3E+02	-6.6E+02	-9.4E+01	4.9E+01	-1.2E+02	1.4E+05	-4.1E+03	-1.3E+03	-8.4E+00	3.8E+01	-3.8E+01	3.1E+02	-8.3E-01
6	3.5E+02	-7.2E+02	-5.4E+03	1.9E+03	-8.2E+01	1.3E+01	7.8E+01	-6.8E+03	-1.9E+01	1.3E+02	-4.6E+01	9.1E+01	-2.1E+00
6	3.6E+01	-9.8E+01	-7.6E+01	6.7E+01	-9.6E+01	3.6E+05	1.9E+04	-7.1E+00	-1.8E+00	3.3E+01	-1.6E+01	-6.0E+00	-7.1E+00
7	1.9E+01	-1.5E+02	3.8E+01	-1.1E+04	-7.5E+03	-1.5E+03	-2.4E+01	-1.8E+00	-4.4E+04	1.5E+01	-9.3E+00	5.3E+00	-1.7E+00
8	3.2E+04	-2.1E+03	2.9E+04	-3.4E+03	1.7E-01	5.8E+01	-5.8E+03	-1.6E+03	3.7E+01	2.4E+01	1.2E+01	7.9E+01	-2.3E+00
9	1.5E+03	-9.3E+01	-7.5E+01	1.9E+04	-2.0E+02	3.9E-03	-2.6E+03	-2.3E+00	-5.4E+00	7.6E+00	-1.7E+02	-1.0E+01	-2.2E+00
10	2.8E+03	-1.9E+02	2.7E-01	2.5E+03	-1.6E+02	-1.3E+02	-2.7E+02	-4.6E+01	2.2E+01	8.2E+00	1.9E+00	1.1E+02	-2.2E+00
	-5.1E+01	-2.7E+02	-3.4E+03	1.9E+02	-1.7E-01	-2.1E+03	-6.5E+00	-9.4E-01	-5.1E+02	-1.5E+01	-1.6E+00	-9.1E+00	-5.6E-01
11	2.8E+01	-6.5E+01	-1.2E+01	-2.5E-01	-1.3E+03	-5.3E+02	-7.4E-01	-1.3E+00	-1.1E+06	6.8E+00	-1.2E+01	-7.4E+00	-7.1E-01
12	1.8E+01	-3.2E+01	-2.4E+00	1.7E+02	-2.3E+02	-3.5E+01	-5.6E+02	-3.7E+04	2.1E-01	6.2E+00	-1.8E+01	-4.1E+01	-1.0E+01
13	1.2E+02	1.0E+02	-1.3E+03	8.1E+01	-1.9E+02	-2.8E+01	-1.9E+05	-1.4E+00	-5.5E-01	1.7E+02	-5.0E+01	-2.5E+01	-5.7E-01
14	2.5E+01	-1.5E-03	-2.7E+03	-1.7E+02	-5.5E+01	-3.0E+02	-3.3E+02	-1.1E+02	-5.3E-01	-1.1E+01	-1.7E+01	-4.4E+00	-1.1E+00
	-8.5E+01	-2.9E+03	-2.2E+02	-1.5E+01	-4.1E+01	-2.4E+02	-1.6E+02	-3.9E+02	-6.8E+00	-4.7E+00	-6.9E+00	-1.3E+01	-7.7E-01
15	1.2E+02	1.6E+02	-2.6E+02	2.5E+01	5.1E+01	2.0E+01	-2.2E+05	1.4E+00	-2.8E-01	1.8E+02	5.0E+01	-3.0E+01	5.7E-01

16	-1.9E+01	1.1E+02	4.8E+02	2.7E-03	5.7E+02	-4.1E+02	-2.0E+02	-5.3E-01	-7.5E+01	-1.8E+00	3.3E+02	7.9E+01	-4.2E-01
17	-7.3E+02	-2.3E+02	2.4E+03	1.9E+04	5.9E+07	4.4E+06	-6.9E+01	-1.4E+00	-1.4E+03	-3.6E+01	6.3E+01	3.3E+00	-7.8E-01
18	-6.7E+01	1.7E+02	7.9E+00	0.0E+00	6.3E+03	-2.2E+02	3.4E+01	-8.2E-01	-1.2E+00	-5.5E+00	9.2E+01	4.4E+00	-9.5E-01
19	9.3E+00	2.5E+01	1.1E+02	-7.1E+02	1.2E+03	-2.4E+01	-6.4E+02	9.6E-01	-1.4E+04	3.9E+01	2.5E+00	-3.0E+00	3.5E-01
20	8.5E+01	3.6E-01	5.7E+02	7.1E+02	-2.6E+02	-1.8E+02	-5.8E+01	-8.3E+01	1.9E+00	2.2E+01	8.6E+00	-1.1E+01	-6.2E-01
21	1.4E+02	-9.2E+01	-5.3E+00	2.9E+01	2.0E+01	2.0E+01	-1.4E+02	1.0E+01	-9.5E+00	2.9E+00	2.0E+01	-1.1E+02	8.9E+00
22	2.7E+00	1.6E+03	0.0E+00	-9.3E+02	0.0E+00	-1.0E+01	1.3E+01	2.7E+02	-4.2E-01	1.4E+01	3.1E+00	-1.2E+00	7.0E-01
23	7.0E+00	1.5E+03	0.0E+00	-9.3E+02	0.0E+00	-1.2E+01	0.0E+00	1.3E+02	-1.6E-01	1.4E+01	3.0E+00	-1.4E+00	6.0E-01
24	4.6E+01	-1.4E-03	-1.0E-03	-1.0E+03	-3.5E-03	3.8E+02	-4.2E+02	-2.2E+01	2.5E-03	3.1E+01	-8.7E+00	3.3E+00	-2.2E+01
	4.3E+01	-5.7E+01	-1.4E+03	-7.2E+01	-6.2E+02	-2.4E+03	-3.7E+02	-9.2E-01	-1.2E+03	9.6E+00	-1.5E+01	4.1E+00	-5.3E-01
25	2.3E+01	-1.2E+01	4.9E-01	1.1E+02	1.3E+01	-2.2E+01	-2.1E+01	-7.5E+00	2.4E+00	1.5E+00	1.4E+01	-9.6E+00	-3.8E+00
26	1.3E+02	-4.5E+02	-5.9E+00	7.7E+01	-1.3E+02	-3.6E+01	-7.3E+01	-2.1E+04	-1.9E+01	6.9E+01	4.7E-01	-4.9E+01	-2.6E+00
27	-4.1E+01	-1.5E+03	-2.4E+03	1.0E+03	-1.2E+02	-2.2E+02	-7.2E+02	-1.6E+02	2.2E+00	-7.5E+00	8.1E+00	5.9E+00	4.1E-01
28	2.1E+01	-1.0E+05	4.2E+01	2.3E+01	-1.8E+04	4.1E+04	9.1E+03	-5.0E+02	-1.9E+07	5.8E+01	1.3E+01	5.4E+01	-5.8E-01
	5.0E+01	5.0E+01	-7.1E+01	-1.2E+00	-5.0E+01	-2.7E+00	-4.0E+01	-7.1E+00	-2.1E+00	-1.3E+01	-2.7E+01	-4.2E+01	-7.0E+00
	-7.5E+01	2.8E-01	-3.0E+02	-5.5E+01	-9.5E+04	-1.1E+03	-3.0E+02	-1.4E+03	-2.6E+04	-3.6E+01	-2.9E+01	-1.1E+00	-4.5E-01
29	-1.8E+01	4.5E+01	2.3E+03	0.0E+00	1.1E+05	-2.6E+00	-6.5E+01	-1.2E+00	-1.1E+00	-1.8E+00	-1.2E+03	-8.7E+00	-6.3E-01
30	-7.0E+01	2.2E+03	-2.0E+03	5.3E+02	3.4E+03	6.0E+03	-4.8E+02	-1.3E+02	-6.8E+00	-5.8E+00	-6.2E+00	-2.7E+01	-1.6E+00
	-1.6E+02	3.8E+02	-2.8E+04	1.3E+03	2.9E+03	5.5E+03	-1.2E+04	-7.5E+02	-4.3E+00	-1.3E+02	-1.7E+01	-4.3E+00	-1.4E+00
	-3.3E+03	1.3E+02	-6.2E+02	1.0E+02	8.0E+03	5.3E+02	-7.0E+02	-8.6E+01	-4.0E+00	-1.3E+00	-6.4E+00	-1.8E+00	-1.4E+00
31	2.7E+02	1.2E+02	-4.1E+03	5.4E+02	-1.4E+03	-1.3E+01	-1.7E+00	-9.8E+02	-4.5E-02	-2.1E+00	-3.8E+00	1.6E-02	-9.3E-01

	2.0E+01	7.1E+01	-6.1E+01	2.7E+03	-5.2E+01	-1.7E+02	-2.6E+02	-2.0E+01	1.0E+00	1.0E+01	4.5E+00	-5.2E+01	-4.0E-01
32	-1.8E+01	7.6E+01	2.3E+03	-3.6E+01	1.4E+04	7.5E+01	-1.2E+02	-1.5E+00	-1.1E+00	1.8E+00	1.2E+03	8.7E+00	-8.0E-01
33	-4.1E+03	-5.6E+00	8.5E+03	6.1E+02	-2.4E+01	1.5E+01	-4.5E+01	8.3E+00	2.4E+01	3.5E+00	1.1E+00	2.6E+01	3.1E-01
34	-3.9E+02	1.6E+03	1.6E+02	-1.2E+02	5.6E+01	-5.6E+03	-5.5E+02	-1.3E+05	-1.3E+00	-5.0E+01	-4.1E+00	-7.5E-01	-4.0E+00
	-1.2E+01	1.2E+04	2.1E+02	-9.8E+01	6.8E+00	-1.3E+03	-1.4E+04	-2.0E+02	-4.9E+03	-1.1E+01	-1.3E+01	-1.8E+01	-7.2E-01
	-1.7E+01	1.0E+02	6.1E+03	-4.3E+02	5.4E+01	-5.4E+04	-1.5E+02	-3.0E+01	-1.5E+01	2.0E+01	9.1E-01	-1.0E+01	-5.5E+00
35	1.0E+02	-2.9E+02	-2.3E+02	1.2E+02	-3.0E+01	-1.5E+01	7.7E+02	-2.7E+02	2.7E+00	1.5E+00	2.2E+01	3.9E+00	-7.8E-01
36	-3.3E+02	9.5E+02	2.9E+01	6.0E+03	1.4E+03	0.0E+00	-2.8E+01	-2.3E+00	-3.4E+01	-3.2E+01	9.0E+00	-1.1E+01	-1.3E+00
37	-1.2E+01	1.2E+04	2.1E+02	-9.8E+01	6.8E+00	-1.3E+03	0.0E+00	-2.0E+02	-4.9E+03	-1.1E+01	-1.3E+01	-1.8E+01	-7.2E-01
38	-1.3E+04	-2.4E+05	2.6E+03	-1.9E+01	3.9E+02	-1.1E+02	-4.1E+04	-3.1E+05	-2.1E+00	-5.4E+00	-6.1E+02	-8.4E+01	-9.9E-01
39	-9.2E+01	1.9E+02	-1.4E+03	-1.8E+02	5.3E+03	9.6E+02	-2.3E+00	-2.0E+02	-1.1E+00	-1.1E+01	-1.9E+00	-4.3E+00	-3.3E+00
40	-1.7E+01	1.2E+03	3.6E+01	-1.0E+02	3.1E+01	-3.5E+01	2.6E+02	-1.3E+01	-1.8E+01	-4.8E-02	-7.3E+00	-1.6E+01	-8.0E-01
41	-1.1E+02	2.5E+02	3.0E+02	1.1E+02	-1.3E+03	-2.2E+03	-7.8E+01	-4.2E+01	-5.3E+04	-5.2E+01	-5.5E+00	-1.8E+00	-1.8E+00
	-1.7E+02	7.0E+02	1.0E+02	-1.3E+01	3.4E+02	1.6E+07	-1.1E+05	-1.4E+03	-2.4E+04	-2.7E+01	-3.5E+01	-1.5E+01	-1.3E+00
	-2.5E+01	7.7E+01	2.3E+03	9.4E+02	2.0E+04	2.5E+01	-3.5E+01	-1.2E+00	-1.1E+00	-1.1E+01	8.2E+01	4.7E+00	-6.3E-01
	-6.4E+01	6.4E+01	4.8E+02	7.9E+01	1.3E+02	1.0E-03	-8.7E+02	-6.5E+03	2.9E-01	-4.1E+00	-1.7E+01	-5.8E+00	-6.7E+00

Table S.11: Network structures with their respective parameter values.

Detailed Mathematical Model of GAL System

I have implemented the model of the Galactose Utilization Pathway developed by Ramsey S. *et al.,* (2006); the original implementation was done in Dizzy language (a chemical kinetics modeling program). The equations in the publication were not a correct translation of the simulated system. I have therefore done a manual conversion of the chemical equations to ODEs and implemented them in Matlab in order to compare my experimental data to the simulations. The metabolic reactions, and the ODEs are given here. Some simplifications based on steady-state assumptions are provided and briefly discussed as well.

To create the knock-outs in the model, the rate of production of the specific mRNA was reduced by half. In the model, GAL6, MIG1 and GAL7 are not included.

Specie's name	Description	Initial data (molecules)
R1	*GAL1 mRNA*	*0.2647*
R2	*GAL2 mRNA*	*0.3305*
R3	*GAL3 mRNA*	*0.9044*
R4	*GAL4 mRNA*	*0.4*
Rrep	*reporter mRNA*	*0.2647*
R80	*GAL80 mRNA*	*1.1871*
G1	*Gal1p*	*132.3267*
G2	*Gal2p*	*1156.7*
G3 & G3i	*Gal3p / Gal3p (activated by galactose)*	*4341.2 / 0*
G4 & G4d	*Gal4p: (monomer) / (homodimer)*	*0.1563 / 308.92*
G80 & G80c & G80d & G80cd	*Gal80p: monomer, nucleus / monomer, cytoplasm / homodimer, nucleus /homodimer, cytoplasm*	*0.1138 / 0.1095 / 157.229 / 157.229*
Grep	*GFP reporter*	*132.327*
G3i80	*complex of Gal3p (activated) and Gal80cd*	*0*
Gic	*intracellular galactose*	*0*

Table S.12: Description of species used in the GAL model

Chemical reactions included in the model

kinetic parameters are defined as:

alpha_TR = 1; // dimensionless

Km_TR = 1.0; // molecules

k_TR = 4350.0; // min^-1

kcat_GK = 3350.0; // min^-1

Km_GK = 0.6 / molecTomM; // molecules

Other parameters are defined as:

q = 30; // cooperativity for multiple repressor (G80d) binding

kfp = 6.5/max_G4d;

kfr = 5*kfp;

krp = 1.0;

krr = 1.0;

Reactions:

Metabolic reaction velocities are defined as:

vTR = [k_TR * G2 * (GAE_mM - GAI_mM) / (Km_TR + GAE_mM + GAI_mM + (alpha_TR*GAE_mM*GAI_mM/Km_TR))];

vGK = [kcat_GK * G1 * GAI / (Km_GK + GAI)];

where: GAE_mM is concentration of external galactose in mM and,
GAI_mM is concentration of internal galactose in mM

Galactose import and export are defined as:

Import of galactose: → GAI , [vTR];

Export of galactose: GAI → , [vGK];

Chemical reaction are defined as:

Dimer formation of Gal80 in cytosol with rate constant of 100:

G80C + G80C → G80Cd, 100;

Monomer formation of Gal80 in cytosol with rate constant of 0.001:

G80Cd → G80C + G80C, 0.001;

Dimer formation of Gal80 in nucleus with rate constant of 100:

G80 + G80 → G80d , 100;

Monomer formation of Gal80 in nucleus with rate constant of 0.001:

G80d → G80 + G80, 0.001;

Shutteling of dimer form of Gal80 (nucleus ↔ cytosol) with rate constant of 50:

G80d → G80Cd, 50; and G80Cd → G80d, 50;

Shutteling of monomer form of Gal80 (nucleus ↔ cytosol) with rate constant of 50:

G80 → G80C, 50; and G80C → G80, 50;

Dimer formation of Gal4 in nucleus with rate constant of 100:

G4 + G4 → G4d , 100;

Monomer formation of Gal4 in cytosol with rate constant of 0.001:

G4d → G4 + G4, 0.001;

Formation of active form of Gal3 (Gal3 gets activated by binding to galactose):

G3 + GAI → G3i , 0.000000745;

Formation of inactive form of Gal3:

G3i → G3 + GAI, 890;

Complex formation (active Gal3 binding to dimer form of cytosolic Gal80):

G80Cd + G3i → G80G3i , 0.025716;

Complex deformation:

G80G3i → G80Cd + G3i, 0.0159616;

dimer decay reactions are defined as follow; the decay rates of the dimer form are approximated to be the same as the monomer form

G4d →, kdp_gal4;

G80d → , kdp_gal80;

G80C → , kdp_gal80;

G80Cd → , kdp_gal80;

Other molecular decay processes

G3i → , kdp_gal3;

G80G3i → , 0.5*kdp_gal3;

Production rates and decay rates are parameterized for all the species:

make_r1, → R1, [SaturatedFraction_FourBidingSites * kir_gal1];

make_r2, → R2, [SaturatedFraction_FiveBidingSites * kir_gal2];

make_r3, → R3, [SaturatedFraction_OneBidingSite * kir_gal3];

make_r4, → R4, [1 * kir_gal4];

make_r80, → R80,[SaturatedFraction_OneBidingSite * kir_gal80];

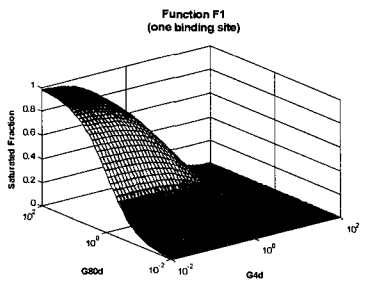
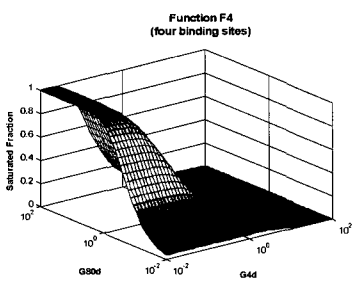
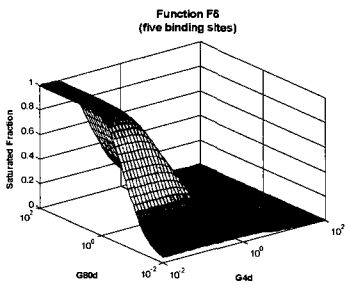
make_rep_rna, → reporter_rna, [SaturatedFraction_FourBidingSites * kir_rep];

degrade_r1, R1 → , kdr_gal1;
degrade_r2, R2 → , kdr_gal2;
degrade_r3, R3 → , kdr_gal3;
degrade_r4, R4 → , kdr_gal4;
degrade_r80, R80 → , kdr_gal80;
degrade_rep_rna, reporter_rna → , kdr_rep;

make_g1, $R1 → G1, kip_gal1;
make_g2, $R2 → G2, kip_gal2;
make_g3, $R3 → G3, kip_gal3;
make_g4, $R4 → G4, kip_gal4;
make_g80, $R80 → G80, kip_gal80;
make_reporter, $reporter_rna → reporter, kip_rep;

degrade_g1, G1 → kdp_gal1;
degrade_g2, G2 → kdp_gal2;
degrade_g3, G3 → kdp_gal3;
degrade_g4, G4 → kdp_gal4;
degrade_g80, G80 → kdp_gal80;
degrade_reporter, reporter → kdp_repoter;

Saturated fraction: Saturated fraction depends on the concentration of Gal80 nuclear and Gal4 nuclear both in the dimer form. To understand the concept of saturated fraction, consider a simple model where binding of a gene (GAL4) leads to promotion of transcription and binding of a second gene, Q, to G (GAL80-GAL4) represses the transcription. If a third gene, R, is able to bind to Q, retaining it from repressing the transcription, then increasing R will increase the transcription. To model the GAL network, the concept presented here was developed and utilized. In addition to that, we know that biding of a protein does affect the binding of subsequent protein, therefore it is important to model the system taking into account the cooperativity of the system. (for detailed mathematical formulation see publication by Atauri *et al.,* 2004). The exact formulation of the saturation fraction is given in the second row of the next table and it is translated to a set of three equations (see ODEs); the three plots are graphical representations of these complex equations. These complex equations are incorporated in the mathematical model of the GAL network (Ramsey S. *et al.,* 2006).

9	kp = kfp/krp;
	kr = kfr/krr;
	kpf0 = [kp*G4d];
	kpf0_2 = [kpf0*kpf0];
	kpf0_3 = [kpf0_2*kpf0];
	kpf0_4 = [kpf0_3*kpf0];
	krf1 = [kr*G80d];
	krf1_2 = [q*krf1*krf1];
	krf1_3 = [q*krf1_2*krf1];
	krf1_4 = [q*krf1_3*krf1];
	numerator = [(4.0 * kpf0) +
	(12.0 * kpf0_2 * krf1) +
	(6.0 * kpf0_2) +
	(4.0 * kpf0_3) +
	(12.0 * kpf0_3 * krf1) +
	(12.0 * kpf0_3 * krf1_2) +
	(kpf0_4) +
	(4.0 * kpf0_4 * krf1) +
	(4.0 * kpf0_4 * krf1_3) +
	(6.0 * kpf0_4 * krf1_2)];
	fracsat_foursites= [numerator /
	(numerator +
	1.0 +
	(4.0 * kpf0 * krf1) +
	(6.0 * kpf0_2 * krf1_2) +
	(4.0 * kpf0_3 * krf1_3) +
	(kpf0_4 * krf1_4))];
		kp = kfp/krp;
		kr = kfr/krr;
		kpf0 = [kp*G4d];
		kpf0_2 = [kpf0*kpf0];
		kpf0_3 = [kpf0_2*kpf0];
		kpf0_4 = [kpf0_3*kpf0];
		kpf0_5 = [kpf0_4*kpf0];
		krf1 = [kr*G80d];
		krf1_2 = [q*krf1*krf1];
		krf1_3 = [q*krf1_2*krf1];
		krf1_4 = [q*krf1_3*krf1];
		krf1_5 = [q*krf1_4*krf1];
		numerator = [(5.0 * kpf0) +
		(10.0 * kpf0_2) +
		(20.0 * kpf0_2 * krf1) +
		(30.0 * kpf0_3 * krf1_2) +
		(30.0 * kpf0_3 * krf1) +
		(10.0 * kpf0_3) +
		(5.0 * kpf0_4) +
		(30.0 * kpf0_4 * krf1_2) +
		(20.0 * kpf0_4 * krf1) +
		(20.0 * kpf0_4 * krf1_3) +
		(kpf0_5) +
		(5.0 * kpf0_5 * krf1) +
		(5.0 * kpf0_5 * krf1_4) +
		(10.0 * kpf0_5 * krf1_3) +
		(10.0 * kpf0_5 * krf1_2)];
		fracsat_fivesites = [numerator /
		(numerator +
		1.0 +
		(5.0 * kpf0 * krf1) +
		(10.0 * kpf0_2 * krf1_2) +
		(10.0 * kpf0_3 * krf1_3) +
		(5.0 * kpf0_4 * krf1_4) +
		(kpf0_5 * krf1_5))];

Table S.13: Parameterized fractional saturation function for genes with multiple binding sites.

The ODEs are obtained following a conversion from the above metabolic equations:

$$dR1/dt = Kirgal\ 1(F4(KpG\ 4d, KqG\ 80d, qr)) - Kdrgal\ 1(R1)$$
$$dR2/dt = Kirgal\ 2(F5(KpG\ 4d, KqG\ 80d, qr)) - Kdrgal\ 2(R2)$$
$$dR3/dt = Kirgal\ 3(q3)(F1(KpG\ 4d, KqG\ 80d)) - Kdrgal\ 3(R3)$$
$$dR4/dt = R4 - R4(Kdrgal\ 4)$$
$$dRrep/dt = Kirrep\ (F4(KpG\ 4d, KqG\ 80d, qr)) - Kdrrep\ (Rrep)$$
$$dR80/dt = Kirgal\ 80(F1(KpG\ 4d, KqG\ 80d)) - Kdrgal\ 80(R80)$$
$$dG1/dt = Kipgal\ 1(R1) - Kdpgal\ 1(G1)$$
$$dG2/dt = Kipgal\ 2(R2) - Kdpgal\ 2(G2)$$
$$dG3/dt = Kipgal\ 3(R3) - Kdpgal\ 3(G3) - Kfi(G3)(Gic) + Kri(G3i)$$
$$dG3i/dt = Kfi(G3)(Gic) - Kri(G3i) - Kdpgal\ 3(G3i) - Kfd\ 3i80(G80cd)(Krd\ 3i80)(C3i80)$$
$$dG4/dt = Kipgal\ 4(R4) - Kdpgal\ 4(G4) - 2(Kfd)(G4)(G4) + 2(Krd)(G4d)$$
$$dG4d/dt = Kfd(G4)(G4) - Krd(G4d) - Kdpgal\ 4(G4d)$$
$$dGrep/dt = Kiprep\ (Rrep) - Kdprep\ (Grep)$$
$$dG80/dt = Kipgal\ 80(R80) - Kdpgal\ 80(G80) - Kf\ 80(G80) + Kr\ 80(G80c) - 2(Kfd)(G80)(G80) + 2(Krd)(G80d)$$
$$dG80c/dt = Kf\ 80(G80) - Kr\ 80(G80c) - 2(Kfd)(G80c)(G80c) + 2(Krd)(G80cd) - Kdpgal\ 80(G80c)$$
$$dG80d/dt = Kfd(G80)(G80) - Krd(G80d) - Kdpgal\ 80(G80d) - Kf\ 80(G80d) + Kr\ 80(G80cd)$$
$$dG80cd/dt = Kfd(G80c)(G80c) - KrdG\ 80cd - Kdpgal\ 80(G80cd) + Kf\ 80(G80d) - Kr\ 80(G80cd) - Kfd\ 3i80(G80cd)(G3i) + Krd\ 3i80(C3i80)$$
$$dC3i80/dt = Kfd\ 3i80(G80cd)(G3i) - Krd\ 3i80(C3i80) - 0.5(Kdpgal\ 3)(C3i80)$$
$$dGic/dt = Vtr - Vgk - Kfi(G3)(Gic) + Kri(G3i)$$

$$Vtr = \frac{Ktr(G2)(Gex - Gic)}{(Kmtr + Gex + Gic + \frac{Atr}{Kmtr}(Gex)(Gic)}$$

$$Vgk = \frac{Kcatgk\ (G1)(Gic)}{Kmgk + Gic}$$

$$F1(P,Q) = \frac{P}{1 + P + PQ}$$

$$F4(P,Q,q) = \frac{4P + 12P^2Q + 6P^2 + 4P^3 + 12P^3Q + 12qP^3Q^2 + P^4 + 4P^4Q + 4q^2P^4Q^3 + 6qP^4Q^2}{Numarator + 1 + 4PQ + 6qP^2Q^2 + 4q^2P^3Q^3 + q^3P^4Q^4}$$

$$F5(P,Q,q) = \frac{5P + 10P^2 + 20P^2Q + 30qP^3Q^2 + 30P^3Q + 10P^3 + 5P^4 + 30qP^4Q^2 + 20P^4Q + 20q^2P^4Q^3 + P^5 + 5P^5Q + 5q^3P^5Q^4 + 10q^2P^5Q^3 + 10qP^5Q^2}{Numarator + 1 + 5PQ + 10qP^2Q^2 + 10q^2P^3Q^3 + 5q^3P^4Q^4 + q^4P^5Q^5}$$

Galactose utilization pathway has been modeled in detail as previously described in Chapter 1. I have used a recent model of this network (Ramsey S. *et al.,* 2006) to investigate its potential to reproduce the experimental data presented here. The model is based on chemical reactions and was implemented in Dizzy language (a chemical kinetic modeling program). After a conversion of the reaction equations to a set of ODEs, I have produced simulated results for some of the knockdown strains. To generate the gene knockdown strains, the RNA synthesis rate was divided by two assuming that decreasing gene copy number from two copies to one copy results in reducing the protein concentration by 50%. The simulated data had the same trends as the experimental datasets for the average expression level (Figure S.34). For instance, reducing GAL80's gene copy number from two to one, increased the expression level; this was also observed experimentally. Opposite was also found for GAL3 knockdown strain. Following this validation, we hoped to reproduce the bimodality observed when single cell data was analyzed. To achieve that, we have performed an exhaustive parameter optimization to search for a set of parameters that would capture the bimodality of the system. We have not found any parameter set that would satisfy the bistability condition; the latter does not mean that the system in not capable of such behavior. However, with 19 ODEs, which were found to be stiff due to significant difference in parameter values, and 44 parameters, it was computationally expensive to further search the parameter space. A reduction of set of equations with assumption of steady-state measurement was also attempted with no success. Figure S.35 shows an example of simulated results following parameter optimization process; convergence was reached and bimodality was never detected.

This model is based on binding of Gal3p-galactose (active form of Gal3p) to dimer form of Gal80p; however the experimental observations indicates that the active form of Gal3p can bind to Gal80p monomer and this complex formation is in competition with Gal80p dimer formation (Pilauri, V. *et al.,* 2005). The model also assumes same decay rate for monomer and dimer forms of both Gal80p and Gal4p; this assumption does not take into account the fact that dimer formation has stabilizing property and therefore, the decay rate of a dimer is usually much lower specially if the dimer formation is strongly favored, which was the case for both Gal4p and Gal80p (Keq was estimated to 100/0.001). Finally we were not able to find a parameter set to fit our single cell experimental data using this model.

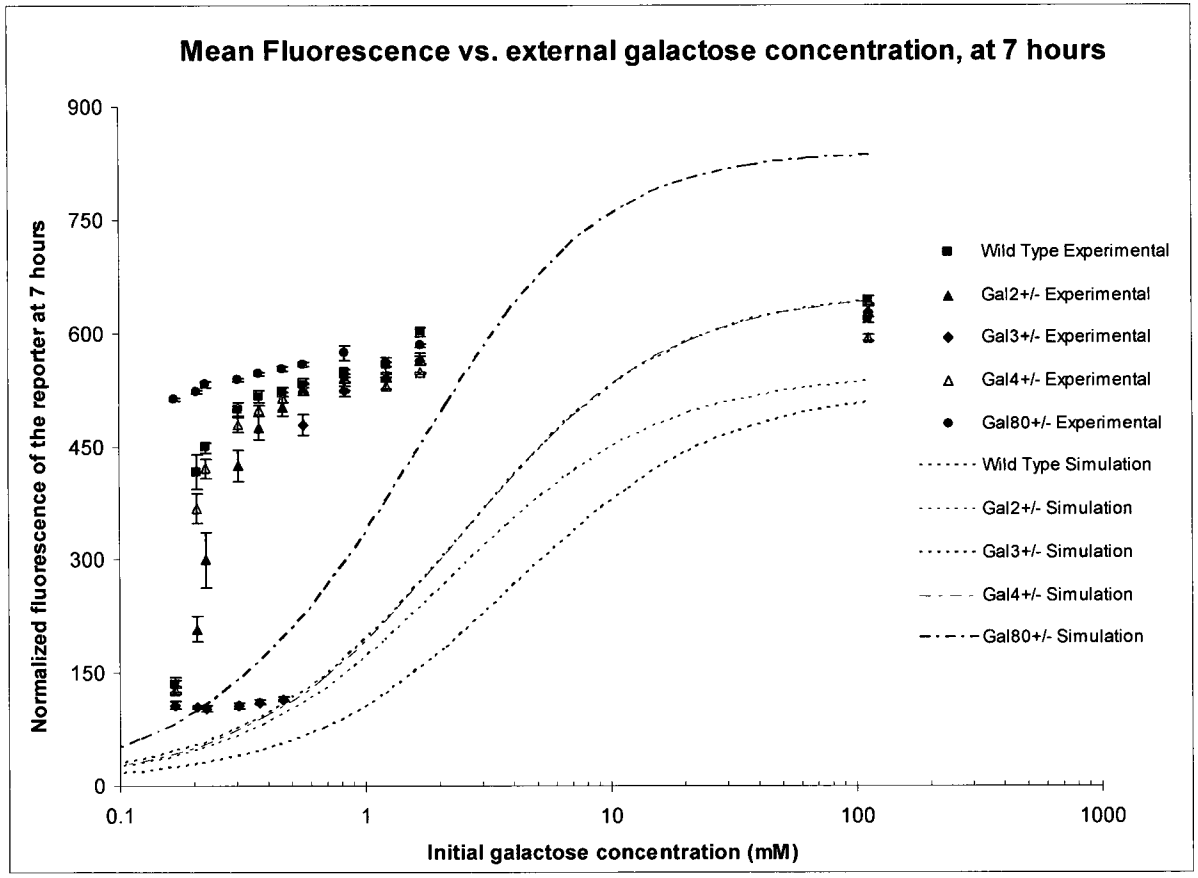


Figure S.34: Average fluorescence from the reporter gene as a function of external galactose concentration. Simulation results have been normalized relative to the experimental expression level. The simulation is obtained using the model developed by Ramsey S. *et al.*, (2006). Error bars represent standard error (n=3 replicates) on experimental data.

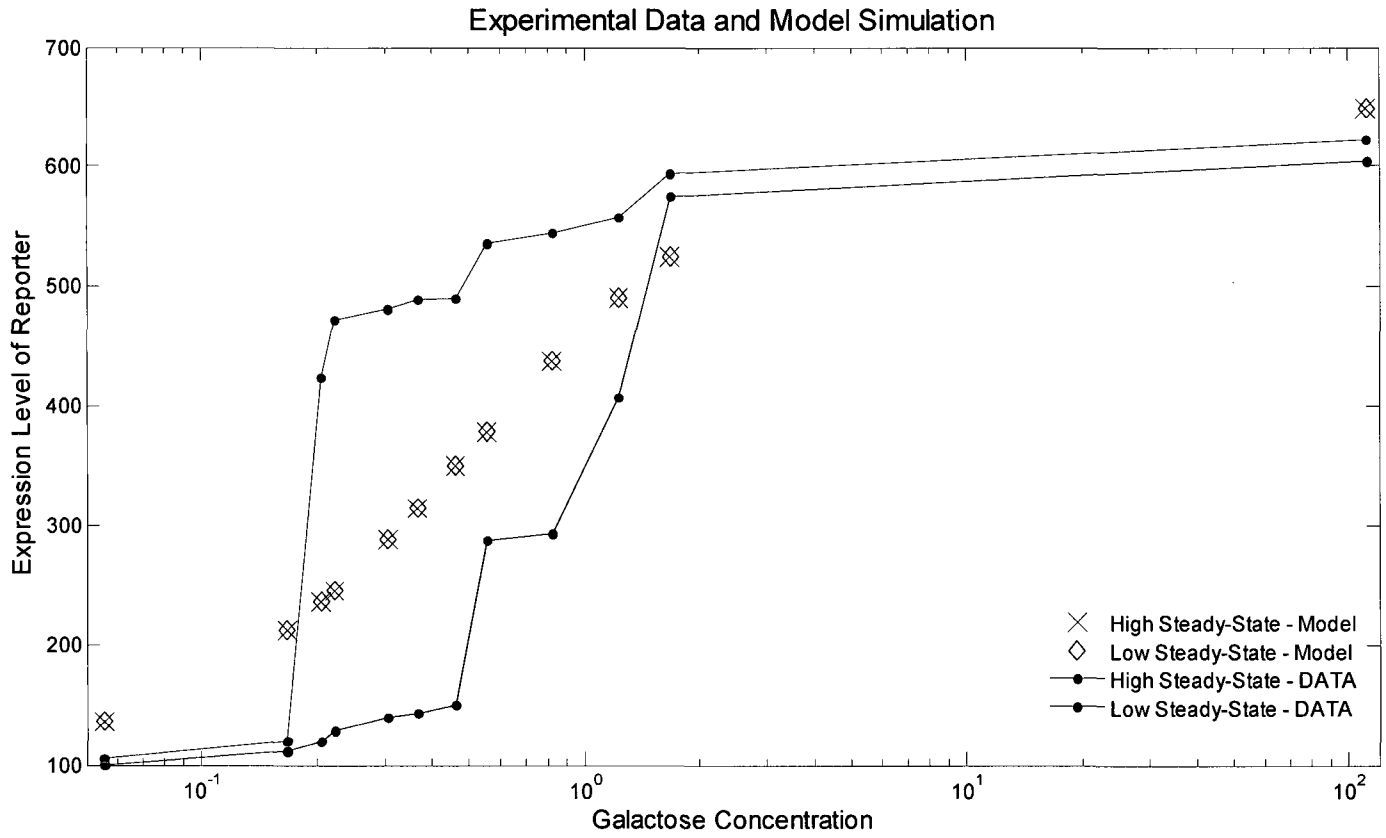


Figure S.35: Dose response curves for experimental data and model simulation using wild-type strain. An exhaustive parameter search fits the data to an average value between the expressing and non-expressing cells. The model does not detect bistability that was continuously observed in the experimental data. The search of parameter space was halted after many unsuccessful trials.

Simplification of the ODEs from Aitchison's Model of the GAL network in order to reduce computational time when solving the ODEs in Matlab.

The reduced form will have seven equations instead of 19. The following two assumptions are taken into account for model simplification: 1) reactions are at steady-state and 2) concentration of GAL80 dimer and GAL4 dimer are significantly higher than the concentration of the monomer form.

Step-by-step simplification

reaction number 1: dR1/dt
reaction number 2: dR2/dt
reaction number 3: dR3/dt
reaction number 4: dRrep/dt
reaction number 5: dR80/dt
reaction number 6: dG1/dt
reaction number 7: dG2/dt
reaction number 8: dG3/dt
reaction number 9: dG3i/dt
reaction number 10: dG4/dt
reaction number 11: dG4d/dt
reaction number 12: dGrep/dt
reaction number 13: dG80/dt
reaction number 14: dG80C/dt
reaction number 15: dG80d/dt
reaction number 16: dG80Cd/dt
reaction number 17: dC3i80/dt
reaction number 18: dGic/dt
reaction number 19: dR4/dt

Step 1: at steady state the rate of change of RNA is zero, therefore we have the following steady-state relationships:

uss(1) = Kirgal1*F(1)/Kdrgal1;
uss(2) = Kirgal2*F(2)/Kdrgal2;
uss(3) = Kirgal3*q3*F(3)/Kdrgal3;
uss(4) = Kirrep*F(4)/Kdrrep;
uss(5) = Kirgal80*F(5)/Kdrgal80;
uss(19) = Kirgal4*F(6)/Kdrgal4;

Therefore the 6 and 7th equations can be rewritten as:
f(6) = Kipgal1*uss(1) -Kdpgal1*u(6) ;
f(7) = Kipgal2*uss(2) -Kdpgal2*u(7) ;

Step 2 : GAL3 containing complexes are: GAL3, GAL3i and, C3i80
Therefore: GAL3T = u(8) + u(9) + u(17)
→ fGAL3T = Kipgal3*uss(3) [-Kfi*u(8) *u(18) +Kri*u(9)] -Kdpgal3*u(8) + [Kfi*u(8) *u(18) -Kri*u(9)] [-Kfd3i80*u(16) *u(9) +Kdr3i80*u(17)] -Kdpgal3*u(9) + [Kfd3i80*u(16) *u(9) -Kdr3i80*u(17)] -0.5*Kdpgal3*u(17)
= Kipgal3*uss(3) -Kdpgal3*u(8) -Kdpgal3*u(9) -0.5*Kdpgal3*u(17)
= Kipgal3*uss(3)-Kdpgal3*(u(8)+u(9)+0.5*u(17))

ASSUME ALL GAL3 COMPLEXES DECAY AT SAME RATE:
fGAL3T = Kipgal3*uss(3)-Kdpgal3*(u(8)+u(9)+u(17)) = Kipgal3*uss(3)-Kdpgal3*GAL3T

Step 3: GAL4 containing complexes are: G4 and G4d
Therefore: GAL4T = u(10) + 2*u(11)
fGAL4T = Kipgal4*uss(19) - Kdpgal4*u(10) - 2*-Kdpgal4*u(11)
fGAL4T = Kipgal4*uss(19) - Kdpgal4*GAL4T
u(11) = 0.5*GAL4T

Step 4: Total reporter

f(12) = Kiprep*uss(4) -Kdprep*u(12) ;

Step 5: Need to find U(11) and U(15) as functions of GAL3T, GAL80T and GAL4T

U(11)

ASSUME u(10) << u(11) → GAL4 is largely in the dimer form (the rate constant is 100 vs. 0.001)

SET **U(11) = 0.5*GAL4T** → from Step 3

U(15)

From GAL80T = u(13) + u(14) + 2*u(15) + 2*u(16) + 2*u(17)

Need to express u(13), u(14), u(16), u(17) in terms of u(16) from equilibrium reactions,

then isolate u(15) in terms of GAL80T

ASSUME NEGLIGIBLE AMOUNTS OF MONOMERS → u(13) + u(14) << u(15) + u(16) + u(17) *(see figure S.36)*

Therefore: GAL80T ~ 2*u(15) + 2*u(16) + 2*u(17)

→ in the model GAL80 cytosolic dimer (u(17)) binds to GAL3i; however it has been shown that it is the monomer form of GAL80 that binds to the GAL3i (see introduction). That change to the actual model has a major effect and cannot be made without having to adjust the parameter values and other kinetics of the equation.

u(16) = kf80/kr80*u(15)

u(17) = Kfd3i80*kf80/kr80*u(15)*u(9)/Kdr3i80

NEED TO FIND U(9)

GAL3T = u(8) + u(9)*(1 + Kfd3i80*kf80/kr80*u(15)/Kdr3i80)

NEED TO FIND U(8)

Kfi*u(8) *u(18) -Kri*u(9) = 0 → **u(8) = Kri*u(9)/Kfi/u(18)**

Therefore: GAL3T = u(9)*(1+Kri/Kfi/u(18)+Kfd3i80*kf80/kr80*u(15)/Kdr3i80))

Therefore: **u(9) = GAL3T/(1+Kri/Kfi/u(18)+Kfd3i80*kf80/kr80*u(15)/Kdr3i80))**

INSERT INTO GAL80T ~ 2*u(15) + 2*u(16) + 2*u(17)

GAL80T =2*u(15) +
2* kf80/kr80*u(15) +
2*Kfd3i80*kf80/kr80*u(15)*GAL3T/(1+Kri/Kfi/u(18)+Kfd3i80*kf80/kr80*u(15)/Kdr3i80))/Kdr3i80

ASSUMPTIONS:

Since Kri/Kfi = 1.2e+09

Therefore: IF u(18) < 10^6 THEN GAL3T = u(9)*(Kri/Kfi/u(18)+Kfd3i80*kf80/kr80*u(15)/Kdr3i80))

Since:

Kfd3i80/Kdr3i80 = 0.02572/0.01596 = 1.6, and

kf80/kr80 = 50/50 = 1,

Therefore: IF u(18) < 10^6 AND u(15) < 10^3 THEN GAL3T =~ u(9)*(Kri/Kfi/u(18))

→ u(9) = GAL3T/(Kri/Kfi/u(18)) → **u(9) = GAL3T*Kfi*u(18)/Kri**

GAL80T ~ 2*u(15) + 2*u(16) + 2*u(17)

u(16) = kf80/kr80*u(15)

u(17) = Kfd3i80*kf80/kr80*u(15)*GAL3T*Kfi*u(18)/Kri/Kdr3i80

GAL80T ~ 2*u(15) + 2*kf80/kr80*u(15) + 2*Kfd3i80*kf80/kr80*u(15)*GAL3T*Kfi*u(18)/Kri/Kdr3i80 →

GAL80T ~ 2*u(15)*(1+kf80/kr80+Kfd3i80*kf80/kr80*GAL3T*Kfi*u(18)/Kri/Kdr3i80) →

u(15) = GAL80T/2/(1+kf80/kr80+Kfd3i80*kf80/kr80*GAL3T*Kfi*u(18)/Kri/Kdr3i80)

Step 6: Find total galactose, u(18)

GALACT = u(18) + u(9) +u(16)

fGATACT = (Ktr*u(7) *(Gex-(u(18)*4.65*10^-8))/(Kmtr+Gex+(u(18)*4.65*10^-8) +(Atr/Kmtr)*Gex*(u(18)*4.65*10^-8)))
-(Kcatgk*u(6)*u(18) /(Kmgk+u(18)))-Kdpgal3*u(9)-Kdpgal80*u(16)+Kfd*u(14)*u(14)-Krd*u(16)+Kf80*u(15)-Kr80*u(16)

ASSUME GALACTOSE BINDING TO GAL3 IS IN STEADY STATE:

Kfi*u(8)*u(18) - Kri*u(9) = 0 →

f(18) = (Ktr*u(7)*(Gex-(u(18)*4.65*10^-8))/(Kmtr+Gex+(u(18)*4.65*10^-8)+(Atr/Kmtr)*Gex*(u(18)*4.65*10^-8)))
-(Kcatgk*u(6) *u(18) /(Kmgk+u(18))) ;

Step 7: Reduced set of equations are:

u(11) = 0.5*GAL4T

u(15) = GAL80T/2/(1+kf80/kr80+Kfd3i80*kf80/kr80*GAL3T*Kfi*u(18)/Kri/Kdr3i80)

```
u(11) = 0.5*u(10);
u(15) = u(13)/2/(1+Kf80/Kr80+Kfd3i80*Kf80/Kr80*u(8)*Kfi*u(18)/Kri/Kdr3i80);
```

Saturated Fractions: F(1) to F(6) correspond to saturated fractions for R1, R2, R3, Rrep, R80 and R4 respectively

```
F(1) = ((4*(Kp*u(11) )+12*(Kp*u(11) )^2*(Kq*u(15) )+6*(Kp*u(11) )^2+4*(Kp*u(11) )^3+12*(Kp*u(11) )^3*(Kq*u(15)
)+12*qr*(Kp*u(11) )^3*(Kq*u(15) )^2+(Kp*u(11) )^4+4*(Kp*u(11) )^4*(Kq*u(15) )+4*qr^2*(Kp*u(11) )^4*(Kq*u(15)
)^3+6*qr*(Kp*u(11) )^4*(Kq*u(15) )^2)/((4*(Kp*u(11) )+12*(Kp*u(11) )^2*(Kq*u(15) )+6*(Kp*u(11) )^2+4*(Kp*u(11)
)^3+12*(Kp*u(11) )^3*(Kq*u(15) )+12*qr*(Kp*u(11) )^3*(Kq*u(15) )^2+(Kp*u(11) )^4+4*(Kp*u(11) )^4*(Kq*u(15)
)+4*qr^2*(Kp*u(11) )^4*(Kq*u(15) )^3+6*qr*(Kp*u(11) )^4*(Kq*u(15) )^2)+1+4*(Kp*u(11) )*(Kq*u(15) )+6*qr*(Kp*u(11)
)^2*(Kq*u(15) )^2+4*qr^2*(Kp*u(11) )^3*(Kq*u(15) )^3+qr^3*(Kp*u(11) )^4*(Kq*u(15) )^4));

F(2) = ((5*(Kp*u(11) )+10*(Kp*u(11) )^2+20*(Kp*u(11) )^2*(Kq*u(15) )+30*qr*(Kp*u(11) )^3*(Kq*u(15) )^2+30*(Kp*u(11)
)^3*(Kq*u(15) )+10*(Kp*u(11) )^3+5*(Kp*u(11) )^4+30*qr*(Kp*u(11) )^4*(Kq*u(15) )^2+20*(Kp*u(11) )^4*(Kq*u(15)
)+20*qr^2*(Kp*u(11) )^4*(Kq*u(15) )^3+(Kp*u(11) )^5+5*(Kp*u(11) )^5*(Kq*u(15) )+5*qr^3*(Kp*u(11) )^5*(Kq*u(15)
)^4+10*qr^2*(Kp*u(11) )^5*(Kq*u(15) )^3+10*qr*(Kp*u(11) )^5*(Kq*u(15) )^2)/((5*(Kp*u(11) )+10*(Kp*u(11) )^2+20*(Kp*u(11)
)^2*(Kq*u(15) )+30*qr*(Kp*u(11) )^3*(Kq*u(15) )^2+30*(Kp*u(11) )^3*(Kq*u(15) )+10*(Kp*u(11) )^3+5*(Kp*u(11)
)^4+30*qr*(Kp*u(11) )^4*(Kq*u(15) )^2+20*(Kp*u(11) )^4*(Kq*u(15) )+20*qr^2*(Kp*u(11) )^4*(Kq*u(15) )^3+(Kp*u(11)
)^5+5*(Kp*u(11) )^5*(Kq*u(15) )+5*qr^3*(Kp*u(11) )^5*(Kq*u(15) )^4+10*qr^2*(Kp*u(11) )^5*(Kq*u(15) )^3+10*qr*(Kp*u(11)
)^5*(Kq*u(15) )^2)+1+5*(Kp*u(11) )*(Kq*u(15) )+10*qr*(Kp*u(11) )^2*(Kq*u(15) )^2+10*qr^2*(Kp*u(11) )^3*(Kq*u(15)
)^3+5*qr^3*(Kp*u(11) )^4*(Kq*u(15) )^4+qr^4*(Kp*u(11) )^5*(Kq*u(15) )^5));

F(3) = ((Kp*u(11) )/(1+(Kp*u(11) )+(Kp*u(11) )*(Kq*u(15) )));

F(4) = ((4*(Kp*u(11) )+12*(Kp*u(11) )^2*(Kq*u(15) )+6*(Kp*u(11) )^2+4*(Kp*u(11) )^3+12*(Kp*u(11) )^3*(Kq*u(15)
)+12*qr*(Kp*u(11) )^3*(Kq*u(15) )^2+(Kp*u(11) )^4+4*(Kp*u(11) )^4*(Kq*u(15) )+4*qr^2*(Kp*u(11) )^4*(Kq*u(15)
)^3+6*qr*(Kp*u(11) )^4*(Kq*u(15) )^2)/((4*(Kp*u(11) )+12*(Kp*u(11) )^2*(Kq*u(15) )+6*(Kp*u(11) )^2+4*(Kp*u(11)
)^3+12*(Kp*u(11) )^3*(Kq*u(15) )+12*qr*(Kp*u(11) )^3*(Kq*u(15) )^2+(Kp*u(11) )^4+4*(Kp*u(11) )^4*(Kq*u(15)
)+4*qr^2*(Kp*u(11) )^4*(Kq*u(15) )^3+6*qr*(Kp*u(11) )^4*(Kq*u(15) )^2)+1+4*(Kp*u(11) )*(Kq*u(15) )+6*qr*(Kp*u(11)
)^2*(Kq*u(15) )^2+4*qr^2*(Kp*u(11) )^3*(Kq*u(15) )^3+qr^3*(Kp*u(11) )^4*(Kq*u(15) )^4));

F(5) = GRF(3);
F(6) = 1;
```

Steady-state reactions:

```
uss(1) = Kirgal1*GRF(1)/Kdrgal1;
uss(2) = Kirgal2*GRF(2)/Kdrgal2;
uss(3) = Kirgal3*q3*GRF(3)/Kdrgal3;
uss(4) = Kirrep*GRF(4)/Kdrrep;
uss(5) = Kirgal80*GRF(5)/Kdrgal80;
uss(19) = Kirgal4*GRF(6)/Kdrgal4;
```

Simplified ODEs:

```
f(1) =0;
f(2) =0;
f(3) =0;
f(4) =0;
f(5) =0;
```

EQ1: f(6) = Kipgal1*uss(1) -Kdpgal1*u(6) ;
EQ2: f(7) = Kipgal2*uss(2) -Kdpgal2*u(7) ;
fGAL3T = Kipgal3*uss(3)-Kdpgal3*GAL3T:
EQ3: f(8) =Kipgal3*uss(3)-Kdpgal3*u(8);
f(9) =0;
fGAL4T = Kipgal4*uss(19) – Kdpgal4*GAL4T:
EQ4: f(10) = Kipgal4*uss(19) - Kdpgal4*u(10);
f(11) = 0;
REPORTER:
EQ5: f(12) = Kiprep*uss(4) -Kdprep*u(12);
fGAL80T = Kipgal80*uss(5) – Kdpgal3*GAL80T:
EQ6: f(13) =Kipgal80*uss(5) - Kdpgal3*u(13);

```
f(14) = 0;
f(15) = 0;
f(16) = 0;
f(17) = 0;
```

EQ7: f(18) = (Ktr*u(7) *(Gex-(u(18)*4.65*10^-8))/(Kmtr+Gex+(u(18)*4.65*10^-8) +
(Atr/Kmtr)*Gex*(u(18)*4.65*10^-8))) -(Kcatgk*u(6) *u(18) /(Kmgk+u(18))) ;

```
f(19) = 0;
```

Even with only seven equations and the indicated assumptions, the reduced form perform better in speed by about 18% (2.87 seconds versus 2.36 seconds for the reduced equations). The expression level of the reporter gene showed up to 55% variation (Figure S.38).

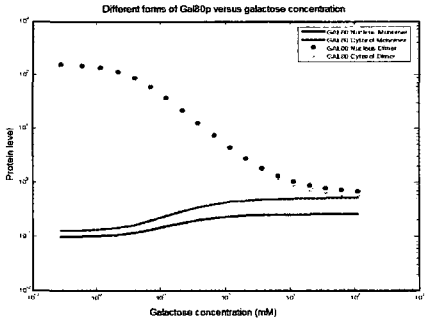


Figure S.36: Different forms of GAL80 vs external Galactose concentrations (mM). The amount of the dimer is 10 to 600 fold higher than the amount of the monomer. GAL80 dimer in both nucleus and cytosol is significantly more abundant. The protein level versus galactose concentration of the dimer and monomer form of GAL80 both in cytosol and nucleus demonstrate this effect. Only at very high galactose concentrations this effect is less pronounced (~10 fold difference). The assumption of GAL80Cd >> GAL80C and GAL80d >> GAL80 is therefore valid and the simplification is acceptable.

Reporter Protein Level vs time

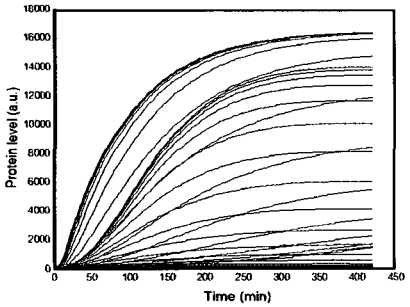


Figure S.37: Protein level versus time at different galactose concentrations. Integration time was taken to be 420min (equivalent to 6hour). Blue indicates the solution to ODEs prior to simplification, red represents the solution to the ODEs following simplification. The graph demonstrates that steady-state was reached following at that time. The assumption was based on the fact that the dimer formation is fast and therefore most of GAL80 is in dimer form both in cytosol and in nucleus (Keq=100/0.001=100000)

Reporter Protein Level vs galactose (mM)

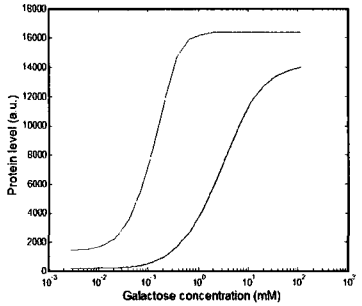


Figure S.38: Reporter protein level versus galactose concentration (in mM). Blue indicates the solution to ODEs prior to simplification, green represents the solution to the ODEs following simplification. The difference between the two dose response curves is significant (up to 55% change in expression level). Gain in speed was only about 18%. The assumption was based on the fact that the dimer formation is fast and therefore most of GAL80 is in dimer form both in cytosol and in nucleus (Keq=100/0.001=100000)

References

Acar, M., Becskei, A., & van Oudenaarden, A. Enhancement of cellular memory by reducing stochastic transitions. *Nature* **435**, 228-232 (2005).

Ahuatzi, D., Riera, A., Pelaez, R., Herrero, P., & Moreno, F. Hxk2 regulates the phosphorylation state of Mig1 and therefore its nucleocytoplasmic distribution. *Journal of Biological Chemistry* **282**, 4485-4493 (2007).

Ahuja R.K. & Orlin J.B. Developing fitter GAs. *INFORMS J on Computing* **9**, 251-253 (1997).

Aloy, P. & Russell, R. B. Taking the mystery out of biological networks. *Embo Reports* **5**, 349-350 (2004).

Bajwa, W., Torchia, T. E., & Hopper, J. E. Yeast Regulatory Gene Gal3 - Carbon Regulation - Uasgal Elements in Common with Gal1, Gal2, Gal7, Gal10, Gal80, and Mel1 - Encoded Protein Strikingly Similar to Yeast and Escherichia-Coli Galactokinases. *Mol.Cell.Biol.* **8**, 3439-3447 (1988).

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., & di Bernardo, D. How to infer gene networks from expression profiles. *Molecular Systems Biology* **3**, (2007).

Barabasi, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509-512 (1999).

Barnett, J. A. A history of research on yeasts 10: foundations of yeast genetics. *Yeast* **24**, 799-845 (2007).

Becskei, A., Seraphin, B., & Serrano, L. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *Embo Journal* **20**, 2528-2535 (2001).

Bhalla, U. S. & Iyengar, R. Robustness of the bistable behavior of a biological signaling feedback loop. *Chaos* **11**, 221-226 (2001).

Boone, C., Bussey, H., & Andrews, B. J. Exploring genetic interactions and networks with yeast. *Nat Rev Genet* **8**, 437-449 (2007).

Brandman, O., Ferrett, J. E., Li, R., & Meyer, T. Interlinked fast and slow positive feedback loops drive reliable cell decisions. *Science* **310**, 496-498 (2005).

Brown, N. E. *et al.* Mechanism reduction via principal component analysis. *Int.J.Chem.Kinet.* **29**, 393-414 (1997).

Bussey, H. *et al.* The nucleotide sequence of Saccharomyces cerevisiae chromosome XVI. *Nature* **387**, 103-105 (1997).

Churcher, C. *et al.* The nucleotide sequence of Saccharomyces cerevisiae chromosome IX. *Nature* **387**, 84-87 (1997).

Collins, S.R., *et al.* A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. Genome Biology **7**, R63 (2006)

Csete, M. E. & Doyle, J. C. Reverse Engineering of Biological Complexity. *Science* **295**, 1664-1669 (2002).

Danø, S. *et al.* Reduction of a biochemical model with preservation of its basic dynamic properties. *FEBS J* **273**, 4862-4877 (2006).

Davidson, E. H. Genomic Regulatory Systems: Development and Evolution. 2001. San Diego, CA, Academic San Diego.

Davis, L. Epistasis variance: a view-point on GA-hardness in *Foundations of Genetic Algorithms* 23-35. Morgan Kaufmann, San Mateo, CA (1991).

De Atauri, P., Orrell, D., Ramsey, S., & Bolouri, H. Evolution of 'design' principles in biochemical networks. *Systems Biology* **1**, 28-40 (2004).

De Jong, K.A. **An Analysis of the Behavior of a Class of Genetic Adaptive Systems**. PhD Dissertation. University of Michigan, (1975).

Deans, T. L., Cantor, C. R., & Collins, J. J. A Tunable Genetic Switch Based on RNAi and Repressor Proteins for Regulating Gene Expression in Mammalian Cells. *Cell* **130**, 363-372 (2007).

Deutscher, D., Meilijson, I., Kupiec, M., & Ruppin, E. Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nature Genetics* **38**, 993-998 (2006).

Devit, M. J., Waddle, J. A., & Johnston, M. Regulated nuclear translocation of the Mig1 glucose repressor. *Molecular Biology of the Cell* **8**, 1603-1618 (1997).

Di Ventura, B., Lemerle, C., Michalodimitrakis, K., & Serrano, L. From in vivo to in silico biology and back. *Nature* **443**, 527-533 (2006).

Dietrich, F. S. *et al.* The nucleotide sequence of Saccharomyces cerevisiae chromosome V. *Nature* **387**, 78-81 (1997).

Douglas, H. C. & Hawthorne, D. C. Enzymatic Expression + Genetic Linkage of Genes Controlling Galactose Utilization in Saccharomyces. *Genetics* **49**, 837-& (1964).

Dujon, B. *et al.* The nucleotide sequence of Saccharomyces cerevisiae chromosome XV. *Nature* **387**, 98-102 (1997).

Edelson, D. Computer simulation in chemical kinetics. *Science* **214**, 981-986 (1981).

Fell, D. A. Undrestanding the Control of Metabolism. 1997. London, Portland Press.

Feng, X. J. *et al.* Optimizing genetic circuits by global sensitivity analysis. *Biophysical Journal* **87**, 2195-2202 (2004).

Ferrell, J. E. Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Current Opinion in Cell Biology* **14**, 140-148 (2002).

Fogarty, T.C. Varying the probability of mutation in the genetic algorithm. ed. J.D.Schaffer. 3rd International Conference on Genetic Algorithms , 104-109. 1989. Los Altos, CA, Morgan Kaufmann.

Foury, F. Human genetic diseases: A cross-talk between man and yeast. *Gene* **195**, 1-10 (1997).

Frolova, E., Johnston, M., & Majors, J. Binding of the glucose-dependent Mig1p repressor to the GAL1 and GAL4 promoters in vivo: regulation by glucose and chromatin structure. *Nucleic Acids Research* **27**, 1350-1358 (1999).

Gancedo, J.M. Yeast carbon catabolite repression. *Microbiology and Molecular Biology Reviews* **62**, 334-+ (1998).

Gardner, T.S. & Faith, J. J. Reverse-engineering transcription control networks. *Physics of Life Reviews* **2**, 65-88 (2005).

Gardner, T.S., et al., Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**, 102-105 (2003).

Gautier, O., Carr R.W., Seigneur, C. Variational sensitivity analysis of a photochemical smog mechanism. *Int.J.Chem.Kinet.* **17**, 1347-1364 (1985).

Gavin, A. C., *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147 (2002).

Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737-741 (2003).

Giaever, G. *et al.* Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387-391 (2002).

Goffrini, P. Ferrero, I., Donnini C. Respiration-Dependent Utilization of Sugars in Yeasts: a Determinant Role for Sugar Transporters. *J Bacteriol* **184** (2), 427-432 (2002).

Goldberg D.E. & Deb K., A comparative analysis of selection schemes used in genetic algorithms in *Foundations of genetic algorithms* 69-93. Morgan Kauufmann, San Mateo, CA (1991).

Grandison, S. & Morris, R. J. Biological pathway kinetic rate constants are scale-invariant. *Bioinformatics* **24**, 741-743 (2008).

Grefenstette, J.J. Optimization of control parameters for genetic algorithms. *IEEE-SMC* **16**, 122-128 (1986).

Griggs, D. W. & Johnston, M. Regulated Expression of the Gal4 Activator Gene in Yeast Provides A Sensitive Genetic Switch for Glucose Repression. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 8597-8601 (1991).

Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99-104 (2004).

Hillenmeyer, M. E. *et al.* The chemical genomic portrait of yeast: Uncovering a phenotype for all genes. *Science* **320**, 362-365 (2008).

Ho, Y. *et al.* Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180-183 (2002).

Holland J.H. Building blocks, cohort genetic algorithms, and hyperplane-defined functions. *Evolutionary computation* **8**, 373-391 (2000).

Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686-691 (2003).

Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929-934 (2001).

Isalan, M. *et al.* Evolvability and hierarchy in rewired bacterial gene networks. *Nature* **452**, 840-845 (2008).

Isidori, A. ***Nonlinear control systems*** Springer, Berlin ; New York (1995).

Ito, T., Chiba, T., & Yoshida, M. Exploring the protein interactome using comprehensive two-hybrid projects. *Trends in Biotechnology* **19**, S23-S27 (2001).

Jacq, C. *et al.* The nucleotide sequence of Saccharomyces cerevisiae chromosome IV. *Nature* **387**, 75-78 (1997).

Jaqaman, K. & Danuser G. Linking data to models: data regression. *Nature Review* **7**, 813-819 (2006).

Johnston, M. *et al.* The nucleotide sequence of Saccharomyces cerevisiae chromosome XII. *Nature* **387**, 87-90 (1997).

Johnston, M. *Molecular Biology of Yeast Saccharomyces cerevisiae*, **2,** Cold Spring Harbor Press, New York (1992).

Kapsalis, A., Rayward-Smith V.J. & Smith G.D. Solving the graphical Steiner tree problem using genetic algorithms. *J.Operational Research Society* **44**, 397-406 (1993).

Kitano, H. BIOLOGICAL ROBUSTNESS. *Nat Rev Genet* **5**, 826-837 (2004).

Kremling, A. *et al.* A benchmark for methods in reverse engineering and model discrimination: Problem formulation and solutions. *Genome Research* **14**, 1773-1785 (2004).

Kumar, P. R., Yu, Y., Sternglanz, R., Johnston, S. A., & Joshua-Tor, L. NADP regulates the yeast GAL induction system. *Science* **319**, 1090-1092 (2008).

Lamphier, M. S. & Ptashne M. Multiple mechanisms mediate glucose repression of the yeast GAL1 gene. *Proceedings of the National Academy of Sciences* **89** , 5922-5926 (1992).

Lee, T. I. *et al.* Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804 (2002).

Levine D. GAs: A practitioner's view. *INFORMS J on Computing* **9**, 256-257 (1997).

Lohr, D., Venkov, P., & Zlatanova, J. Transcriptional Regulation in the Yeast Gal Gene Family - A Complex Genetic Network. *Faseb Journal* **9**, 777-787 (1995).

Lorenz, S., Diederichs E, Telgmann R, Schütte C. Discrimination of Dynamical System Models for Biological and Chemical Processes. *J.Comput.Chem* **28**, 1384-1399 (2007).

Luger, G.F. ***Artificial Intelligence: Structures and Strategies for Complex Problem Solving*** Addison Wesley; Harlow, England (2005).

Lum, P. Y. *et al.* Discovering Modes of Action for Therapeutic Compounds Using a Genome-Wide Screen of Yeast Heterozygotes. *Cell* **116**, 121-137 (2004).

Ma, W. Z., Lai, L. H., Qi, O. Y., & Tang, C. Robustness and modular design of the Drosophila segment polarity network. *Molecular Systems Biology* (2006).

Mani, R., Onge, R. P. S., Hartman, J. L., Giaever, G., & Roth, F. P. Defining genetic interaction. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 3461-3466 (2008).

Marton, M. J. *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Medicine* **4**, 1293-1301 (1998).

Melcher, K. & Xu, H. E. Gal80-Gal80 interaction on adjacent Gal4p binding sites is required for complete GAL gene repression. *Embo Journal* **20**, 841-851 (2001).

Melcher, K. Galactose Metabolism in Saccharomyces cerevisiae: a paradigm for eukaryotic gene regulation. ed. edited by F.K.Zimmermann, K.-D. E. Yeast sugar metabolism : biochemistry, genetics, biotechnology, and applications. 235-269. 1997. Lancaster, PA, Technomic Pub. Co.

Mettetal, J. T., Muzzey, D., Gomez-Uribe, C., & van Oudenaarden, A. The frequency dependence of osmo-adaptation in Saccharomyces cerevisiae. *Science* **319**, 482-484 (2008).

Mitrophanov, A.Y. & Groisman E.A. Positive feedback in cellular control systems. *Bioessays* **30**, 542-555 (2008).

Mizutani, A. & Tanaka, M. Regions of GAL4 critical for binding to a promoter in vivo revealed by a visual DNA-binding analysis. *Embo Journal* **22**, 2178-2187 (2003).

Morohashi, M. *et al.* Robustness as a measure of plausibility in models of biochemical networks. *Journal of Theoretical Biology* **216**, 19-30 (2002).

Okino, M. and Mavrovouniotis, M. Simplification of mathematical models of chemical reaction systems. *Chem.Rev.* (1998).

Oliver, S. G. *et al.* The Complete Dna-Sequence of Yeast Chromosome-Iii. *Nature* **357**, 38-46 (1992).

Papamichos-Chronakis, M., Gligoris, T., & Tzamarias, D. The Snf1 kinase controls glucose repression in yeast by modulating interactions between the Mig1 repressor and the Cyc8-Tupl co-repressor. *Embo Reports* **5**, 368-372 (2004).

Peng, G. & Hopper, J. E. Evidence for Gal3p's cytoplasmic location and Gal80p's dual cytoplasmic-nuclear location implicates new mechanisms for controlling Gal4p activity in Saccharomyces cerevisiae. *Mol.Cell.Biol.* **20**, 5140-5148 (2000).

Peng, G. & Hopper, J. E. Gene activation by interaction of an inhibitor with a cytoplasmic signaling protein. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 8548-8553 (2002).

Perkins, T. J., Jaeger, J., Reinitz, J., & Glass, L. Reverse engineering the gap gene network of Drosophila melanogaster. *Plos Computational Biology* **2**, 417-428 (2006).

Philippsen, P. *et al.* The nucleotide sequence of Saccharomyces cerevisiae chromosome XIV and its evolutionary implications. *Nature* **387**, 93-98 (1997).

Pilauri, V., Bewley, M., Diep, C., & Hopper, J. Gal80 dimeirization and the yeast GAL gene switch. *Genetics* **169**, 1903-1914 (2005).

Platt, A. & Reece, R. J. The yeast galactose genetic switch is mediated by the formation of a Gal4p-Gal80p-Gal3p complex. *Embo Journal* **17**, 4086-4091 (1998).

Ramsey, S. A. *et al.* Dual feedback loops in the GAL regulon suppress cellular heterogeneity in yeast. *Nature Genetics* **38**, 1082-1087 (2006).

Reeves, C.R. & Rowe J.E. ***Genetic algorithms : principles and perspectives : a guide to GA theory*** Kluwer Academic Publishers, Boston (2003).

Reeves, C.R. Tabu search solves GA-hard problems in polynomial time. 15th International Symposium on Mathematical Programming. 1994.

Riehman, K., Crews, C., & Fridovich-Keil, J. L. Relationship between genotype, activity, and galactose sensitivity in yeast expressing patient alleles of human galactose-1-phosphate uridylyltransferase. *Journal of Biological Chemistry* **276**, 10634-10640 (2001).

Ronne, H. Glucose Repression in Fungi. *Trends in Genetics* **11**, 12-17 (1995).

Rowe, J. E., Vose M.D. & Wright A.H. Structural Search Spaces and Genetic Operators. *Evol.Comput* **12**, 461-493 (2004).

Salathe, M., Kouyos, R. D., Regoes, R. R., & Bonhoeffer, S. Rapid parasite adaptation drives selection for high recombination rates. *Evolution* **62** , 295-300 (2008).

Schmidt, H., Madsen, M. F., Dano, S., & Cedersund, G. Complexity reduction of biochemical rate expressions. *Bioinformatics* **24**, 848-854 (2008).

Shimada, H. & Fukasawa, T. Controlled Transcription of the Yeast Regulatory Gene Gal80. *Gene* **39**, 1-9 (1985).

Sil, A. K., Xin, P., & Hopper, J. E. Vectors allowing amplified expression of the Saccharomyces cerevisiae Gal3p-Gal80p-Gal4p transcription switch: Applications to galactose-regulated high-level production of proteins. *Protein Expression and Purification* **18**, 202-212 (2000).

Smidtas, S., Schachter, V., & Kepes, F. The adaptive filter of the yeast galactose pathway. *Journal of Theoretical Biology* **242**, 372-381 (2006).

Sneppen, K., Micheelsen, M. A., & Dodd, I. B. Ultrasensitive gene regulation by positive feedback loops in nucleosome modification. *Molecular Systems Biology* **4**, (2008).

Sontag, E.D. Some new directions in control theory inspired by systems biology. *IET Systems Biology* **1**, 9-18 (2004).

Sopko, R. et al., Mapping pathways and phenotypes by systematic gene overexpression. Mol Cell. 21 (3), 313-330 (2006).

St Onge, R. P. *et al.* Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nature Genetics* **39**, 199-206 (2007).

Stagoj, M. N., Comino, A., & Komel, R. Fluorescence based assay of GAL system in yeast Saccharomyces cerevisiae. *Fems Microbiology Letters* **244**, 105-110 (2005).

Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J., & Doyle, J. Robustness of cellular functions. *Cell* **118**, 675-685 (2004).

Suzuki-Fujimoto, T. *et al.* Analysis of the galactose signal transduction pathway in Saccharomyces cerevisiae: Interaction between Gal3p and Gal80p. *Mol.Cell.Biol.* **16**, 2504-2508 (1996).

Tettelin, H. *et al.* The nucleotide sequence of Saccharomyces cerevisiae chromosome VII. *Nature* **387**, 81-84 (1997).

Thomas, R. Laws for the dynamics of regulatory networks. *International Journal of Developmental Biology* **42**, 479-485 (1998).

Timson, D. J. & Reece, R. J. Kinetic analysis of yeast galactokinase: implications for transcriptional activation of the GAL genes. *Biochimie* **84**, 265-272 (2002).

Tong, A.H.Y., et *al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364-2368 (2001).

Trumbly, R. J. & Bradley, G. Isolation and Characterization of Aminopeptidase Mutants of Saccharomyces-Cerevisiae. *Journal of Bacteriology* **156**, 36-48 (1983).

Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-627 (2000).

Verma, M., Bhat, P. J., & Venkatesh, K. V. Quantitative analysis of GAL genetic switch of Saccharomyces cerevisiae reveals that nucleocytoplasmic shuttling of Gal80p results in a highly sensitive response to galactose. *Journal of Biological Chemistry* **278**, 48764-48769 (2003).

Verma, M., Bhat, P. J., & Venkatesh, K. V. Steady-state analysis of glucose repression reveals hierarchical expression of proteins under Mig1p control in Saccharomyces cerevisiae. *Biochemical Journal* **388**, 843-849 (2005).

Wagner, A. Robustness and Evolvability in Living Systems. 2005. New Jersey, Princeton University Press.

Winge, O. & Roberts, C. Inheritance of enzymatic characters in yeasts, and the phenomenon of long term adaptation. *C.R.Trav.Lab.Carlsberg Ser.Physiol.* **24**, 263-315 (1948).

Wollman, R., Civelekoglu-Scholey, G., Scholey, J. M., & Mogilner, A. Reverse engineering of force integration during mitosis in the Drosophila embryo. *Mol Syst Biol* **4**, (2008).

Zheng, W. J., Xu, H. E., & Johnston, S. A. The cysteine-peptidase bleomycin hydrolase is a member of the galactose regulon in yeast. *Journal of Biological Chemistry* **272**, 30350-30355 (1997).