Towards the integration of computational systems biology and high-throughput data: supporting differential analysis of microarray gene expression data.

Nicola Segata[1], Enrico Blanzieri[1], and Corrado Priami[1 2]

[1]Dipartimento di Ingegneria e Scienza dell'Informazione, University of Trento, Italy
http://www.disi.unitn.it

[2]The Microsoft Research-University of Trento Centre for Computational and Systems Biology
http://www.cosbi.eu

Summary

The paradigmatic shift occurred in biology that led first to high-throughput experimental techniques and later to computational systems biology must be applied also to the analysis paradigm of the relation between local models and data to obtain an effective prediction tool. In this work we introduce a unifying notational framework for systems biology models and high-throughput data in order to allow new integrations on the systemic scale like the use of *in silico* predictions to support the mining of gene expression datasets. Using the framework, we propose two applications concerning the use of system level models to support the differential analysis of microarray expression data. We tested the potentialities of the approach with a specific microarray experiment on the phosphate system in *Saccharomyces cerevisiae* and a computational model of the PHO pathway that supports the systems biology concepts.

1 Introduction

The systems biology concept [12] that is now becoming crucial in the computational modelling field, received the definitive incentive with the success of the "globalists" over the "localists" [9]. A major motivation for the shift from the local study of biochemical pathway to the genomic scale analysis of interaction networks was the developing of massively-parallel and high-throughput techniques [4] which made available a huge amount of unstructured gene-specific or protein-specific data. A consequence of this paradigmatic shift is the need of integrative tools for combining the quantitative and structural information with the functional one. In particular, there is the need to integrate high-throughput expression data and computational systems biology modelling.

One of the most important and widely used high-throughput techniques is the microarray technology, applied for example in [5] to detect the genes related to the phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae*. The identification of genes that show significant changes in expression associated with experimental variables of interest on a microarray is called differential analysis of gene expression data [6, 16] and it is not a trivial

This work was partially supported by FIRB - *Computational tools for Systems Biology* and by BISCA - *Sistemi e calcoli di ispirazione biologica e loro applicazioni*.

task because the high-throughput datasets are large, extremely noisy, possibly with missing or corrupted expression values and subject to biological bias. As a consequence the differences in gene expression caused by the experimental variables can be at least partially hidden by other undesirable sources of variations. Therefore the set of differentially expressed genes is usually quite large since it contains a lot of non-regulated genes that need to be removed. So the interpretation of this set is a hot topic in research mainly because its typical high cardinality prevents the functional profiling and the validation of each gene singularly. Some authors addressed the problem through literature profiling [3] or Gene Ontology-based tools [10, 14].

Computational modelling of biological phenomena [1] combines representations of complex biochemical systems with the possibility to quantitatively reproduce and predict their behaviour. The sophistication reached by computers and programming languages permits one to effectively simulate the models allowing *in silico* predictions at the level of intricate metabolic and signalling pathways like in [7, 8]. Even though the complexity of these models can be very high, they still describe *local* aspects of biochemical systems. The paradigmatic shift in the dimensions and complexity of the models relies on modelling formalisms that natively handle high parallelism and incremental model construction [17, 15]. An example of application of these formalisms is the PHO pathway model presented in [19] in which we model and simulate the phosphate systems in *Saccharomyces cerevisiae* with the stochastic π-calculus, a language that support compositionality (the ability of building models incrementally). The high expressivity power of the π-calculus process algebra and the level of abstraction of the model, that is different from basic chemical reactions in order to handle partial specification of the pathway, make an automatic translation of the model to an analytical model with elementary reactions not straightforward. However this drawback is amply compensated by the compositional operator of the algebra that can be applied on *local models* to obtain a *compositional model* that reaches the systems biology level. Although the system level modelling of simple microorganisms is still more theoretical than effective, the potentialities of the *in silico* approach in this context are extremely promising for the biological community [11].

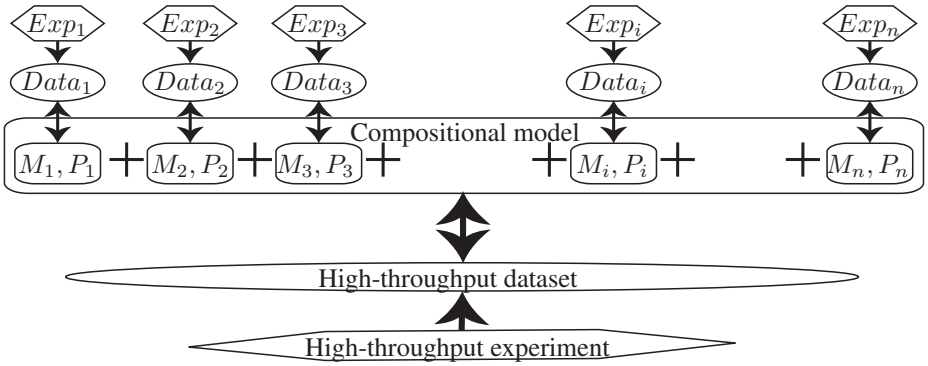

Figure 1: Relation between local models and data and between systemic model and high-throughput data. Each local model M_x with the corresponding parameters P_x can be tuned and validated with data ($Data_x$) produced by specific experiments (Exp_x). The systems biology model, built composing the local sub-models, have new possibilities of integration with high-throughput data. The $+$ symbol stands for the abstract compositional operator of the models.

The relation between data, local models, high-throughput data and compositional models is complex, and Figure 1 tries to schematize its main characteristics. In general, a model has a biological valence only if there are experimental data for the same biological aspect. Typical relations between data and models are parameter estimation, model validation and output data

interpretation. When the biological description is not on a system level, there is a one-to-one relation between a local model and the possibly available data that permits *in silico* simulations and thus biological predictions (schematized in Figure 1 with the double arrows between each local model and each dataset derived from specific experiments). However, high-throughput data are not suitable for tuning and validating local pathway models because data at the genomic scale are not specific for each local pathway and are affected by a large number of sources of uncertainty. With the introduction of computational systems biology, the relation between compositional models and high-throughput data need to be explored (schematized in Figure 1 by the big double arrow between the compositional model and the high-throughput dataset derived from high-throughput experiment). For example, such analysis highlights the relation between the PHO pathway model [19] and the genic expression data on the phosphate metabolism in *Saccharomyces cerevisiae* [5]. Therefore, as the shift from biology to systems biology is a paradigmatic change, also this new relation needs new analysis paradigms and tools to understand how to handle some critical aspects such as: (i) the uncertainty and dimensionality of the high-throughput data, (ii) the computational weight of the models that makes too many replicas of *in silico* predictions not realistic, (iii) the management and comparison of dependent and independent variables in both techniques, (iv) the mutual assessment of the reliability of the techniques in case of partial disagreement between measures and simulations and (v) the combination of large datasets produced by completely different techniques and namely with different statistical properties. For reaching these purposes a necessary step is the development of a general framework able to provide a formal means to represent this relation.

In this paper, we investigate the possibility that the expression information obtained with *in silico* simulations of system level models could support the differential microarray analysis. This approach is possible assuming to have a reasonable large model built in a compositional way, thus permitting the independent validation of each pathway with biological data and it is complementary to other knowledge-based mining methods. We define a unifying notational framework for the experimental high-throughput data and *in silico* high-throughput expression values obtained with a computational model for allowing integrations of the two approaches. The model indirectly reflects the state of the art of the biological knowledge, so the framework allows the biological information to be included in the process of microarray expression mining with the prediction potentialities of the systems biology models. The first application we propose in the framework aims to remove from the set of regulated genes those genes that the model predicts as normal to be regulated in the specific conditions. In this way the experimenter can focus only on the genes with potentially more relevant biological information. We effectively test the introduced approach and the first application to the microarray experiment of [5] using the PHO pathway model [19]. Even if this model is not on the genomic scale, it highlights the utility of the methods and confirms that, as the compositional models scale up to the system level, our approach can give a systematic support to the the analysis of regulated genes in microarray experiments. The second application tackles the problem of the genes that are regulated by dependent and non directly controllable conditions that crowd the set of differentially expressed genes, possibly hiding genes regulated only by the direct and desired experimental conditions. This application is based on the *in silico* prediction of the effects on gene expression of the dependent variables only. Moreover, the framework and the two proposed applications can suggest which are the aspects that need new software tools to be handled properly in the integrative perspective.

The paper is organized as follows: Section 2 introduces the unifying notational framework for

microarray and *in silico* gene expression experiments. Section 3 focuses on the specialization of the framework for microarray differential analysis proposing some measures for quantifying the level and the precision of the integration. Section 4 describes the two specific application based on the framework and in Section 5 we discuss the possibility of developing general software tools in this context. The running example based on the microarray work of [5] and the PHO pathway model [19] is discussed step by step in the sections.

2 A unifying notational framework

We define a framework to formally specify the microarray experiments and the model-based estimation of gene expression in order to allow integrations of the two approaches. We focus on *Affymetrix* oligonucleotide chips [13] as far as the microarray technology is concerned, even though the framework can be adapted to cDNA chips[1]. We apply the framework to the running example which has a simple experimental design; however, in general, there are no particular limitations on the application to more complex designs.

2.1 Microarray experiments and microarray dataset

An Affymetrix microarray experiment consists in the absolute quantification of the expression profiles of a set of genes. The parameters of a microarray experiment are essentially the controlled independent variables with their levels, and the measured dependent variables. Formally, the parameters are $q_{\mu A} = (G_{\mu A}, C, M, P)$ where:

$G_{\mu A}$ is the set of genes that are spotted on the microarray[2].

$C = \{(V_1, l_1), \ldots, (V_k, l_k)\}$ is the set of k different conditions applied on the chip. A condition C_i with $1 \leq i \leq k$ is a pair (V_i, l_i) where V_i is an independent variable controlled by the experimenter and $l_i \in \mathbb{R}$ is a real value assigned to the variable. Hereafter, we assume that the set of k independent variables $V = \{V_1, \ldots, V_k\}$ can be retrieved from the experiment with the function $IV(q_{\mu A}) = V$.

$M = \{(V_{k+1}, m_{k+1}), \ldots, (V_n, m_n)\}$ is the set of $n - k$ different measured dependent variables of the chip. A dependent measured parameters M_i with $k + 1 \leq i \leq n$ is a pair (V_i, m_i) where V_i is a dependent measured variable and $m_i \in \mathbb{R}$ is the value assigned to the variable. Hereafter we assume that the set of dependent variables $V_D = \{V_{k+1}, \ldots, V_n\}$ can be retrieved with the function $DV(q_{\mu A}) = V_D$ and that $V \cap V_D = \emptyset$.

P represents the information regarding the experiment. It should contains the parameters that are not variable and that are sufficient to reproduce the experiment in a rigorous way. P may contain the sample used, the extract preparation and labeling, the procedure and parameters for the hybridization and instruments information. In general it can contains all the information of the MIAME standard [2] not handled by the other defined parameters.

[1]The cDNA chips, in fact, can be modelled as the combinations of two *Affymetrix* chips reflecting the expression of Cy3- and Cy5-labeled probes, like in the case of the microarray experiment of the running example.

[2]Note that it is not always trivial to associate a gene to each spot on the microarray if an oligonucleotide microarray chip is used.

The cDNA chip used in [5] (FODB are the initials of the authors) for the expression profiling between low and high phosphate conditions can be seen as two Affymetrix experiments

$$FODB_{\mu A,Cy3} = (G_{\mu A}^{FODB}, \{(P_i, 0.2mM)\}, \emptyset, P^{FODB})$$
$$FODB_{\mu A,Cy5} = (G_{\mu A}^{FODB}, \{(P_i, 10mM)\}, \emptyset, P^{FODB})$$

Where $G_{\mu A}^{FODB}$ is the set of genes considered in the experiment (approximatively 6400 distinct DNA sequences, available in the additional materials of [5]), P_i is the phosphate concentration, and P^{FODB} contains an exhaustive description of the experimental parameters that allows experimental reproduction. The only experimental independent variable is the phosphate concentration $IV(FODB_{\mu A,Cy3}) = IV(FODB_{\mu A,Cy5}) = P_i$ which assumes two different values in the two chips. The work does not measure any dependent variables so $M = \emptyset$ and consequently $DV(q_{\mu A}) = \emptyset$.

For a specific microarray experiment we can define a function reflecting the experimental procedure that promotes the biochemical reactions on the chip and results in values of absolute expression detected by the instruments. This function for a microarray experiment $q_{\mu A} = (G_{\mu A}, C, M, P)$ has the form $Expr_{q_{\mu A}} : G \mapsto \mathbb{R}$ and associates to a gene $g \in G_{\mu A}$ a value reflecting its expression. All genes spotted on the microarray and the corresponding expression values are included in a dataset called $E_{q_{\mu A}}$, defined as $E_{q_{\mu A}} = \{(g, Expr_{q_{\mu A}}(g)) \mid g \in G_{\mu A}\}$.

The dataset of the microarray experiment of our example is $E_{FODB_{\mu A}}$ and is available in the additional materials of [5] with the relative quantification between $FODB_{\mu A,Cy3}$ and $FODB_{\mu A,Cy5}$ since a cDNA technology is used.

2.2 The *in silico* model-based simulation of expression experiments

Here we propose how to simulate *in silico* an experiment to obtain a dataset of expression profiles. The model can be viewed as a set of metabolic and signalling pathways interacting with each others. In particular, the prerequisites for a model to be suitable in this context are: (i) The model must consider the gene transcription and allow the quantification of gene expression during the simulations. (ii) The model must have a genomic scale; it is not necessary to have a comprehensive model of all the genes of the cell, but the number of considered genes must be comparable to number of genes spotted on a microarray chip. (iii) The model must allow *in silico* experiments that accepts as inputs the environment conditions (like concentrations of the nutrients, temperature, pH, etc.) as independent variables.

The PHO model we use to test the framework, matches the first and the last conditions. The genomic scale, instead, is not respected, and so the model is not suitable for real large-scale microarray mining, but it can still test the usefulness and quality of the approach. Moreover, the used modelling language support the incremental development, and so the model can be extended to other pathways influencing more genes. Obviously, the choice of the values of the model parameters (e.g. reaction rates and species concentrations) should be independent from the data of the microarray experiments whose analysis we want to support with the model.

In the definition of the microarray experiment we have the controllable conditions, the dependent variables and the parameters; the intuition is that they match the input requirements of an *in silico* simulation of a sufficient large subset of the biological network of a cell. So, similarly to the microarray experiment we can give the definition of the parameters of a model-based *in*

gene	expr. (molecules/sec)
PHO2	0.05
PHO4	0.06
PHO81	1.31
PHO5	2.08

(a)

gene	expr. (molecules/sec)
PHO2	0.05
PHO4	0.06
PHO81	0.63
PHO5	0.28

(b)

Figure 2: The datasets of the *in silico* PHO pathway experiment. $E_{q^{SBP}_{m,lp}}$ (a) reflects the low phosphate conditions, while $E_{q^{SBP}_{m,hp}}$ (b) reflects the high phosphate conditions. The expression levels are in terms of molecules per second.

silico expression experiment as $q_m = (G_m, C, M, P)$, where C, M and P are the conditions, the measured parameters and the experimental information as defined for $q_{\mu A}$, while G_m is the set of genes for which the model is able to estimate the expression profile.

The simulated expression experiments described in [19] (SBP are the initials of the authors, *lp* and *hp* denote low and high phosphate conditions) of the reference microarray work [5], are

$$SBP_{m,lp} = (G_m^{SBP}, \{(P_i, 0.2mM)\}, \emptyset, P^{SBP}) \;\; SBP_{m,hp} = (G_m^{SBP}, \{(P_i, 10mM)\}, \emptyset, P^{SBP})$$

with the same definition given for $FODB_{\mu A,Cy3}$ and $FODB_{\mu A,Cy5}$ except for P^{SBP} which is the *in silico* correspondent of P^{FODB} and G_m^{SBP} which contains very few genes with respect to $G_{\mu A}^{FODB}$ since the used model has not a genomic scale. In particular we have that $G_m^{SBP} = \{$PHO2,PHO4,PHO81,PHO5$\}$. The expression of a gene can be seen as the result of a particular instance of the model that is simulated with the particular inputs. So, with a conceptual analogous of the microarray expression function, for every q_m there exists an intensionally defined function that associates a real value to each gene as follows: $Expr_{q_m} : G \mapsto \mathbb{R}$. $Expr_{q_m}$ reflects the simulated biochemical reactions occurring in a living cell, whereas $Expr_{q_{\mu A}}$ reflected the biochemical reactions occurring in the microarray experiment preparation. $Expr_{q_m}$ and $Expr_{q_{\mu A}}$ can be seen as *in silico* estimation and high-throughput measure of the real gene expression in a living cell. The corresponding dataset for $Expr_{q_m}$ is $E_{q_m} = \{(g, Expr_{q_m}(g)) \mid g \in G_m\}$.

Figure 2 shows the datasets of the *in silico* PHO pathway experiment ($E_{q^{SBP}_{m,lp}}$ and $E_{q^{SBP}_{m,hp}}$). Notice that the *in silico* expression datasets and the microarray ones adopt different expression quantification: an absolute value reflecting the intensity level of the spots the firsts, the rate of new molecules the seconds.

3 Differential analysis of microarray gene expression data

The main objective in gene expression analysis is the detection of the genes that are differentially expressed (or regulated) between two biological samples denoted as sample 1 and sample 2 (a concept relative to one of the sample will be denoted with the corresponding apex) with some experimental differences. In differential analysis the classification of the genes is made with complex statistical techniques that analyse the entire distribution of gene expressions [6, 20].

We define the following microarray experiments:

$$\begin{array}{l} q^1_{\mu A} = (G_{\mu A}, C^1, M^1, P) \\ q^2_{\mu A} = (G_{\mu A}, C^2, M^2, P) \end{array} \quad \text{with} \quad \begin{array}{l} IV(q^1_{\mu A}) = IV(q^2_{\mu A}) \text{ and} \\ DV(q^1_{\mu A}) = DV(q^2_{\mu A}) \end{array} \tag{1}$$

$q^1_{\mu A}$ and $q^2_{\mu A}$ can differ only in the level of the dependent and independent variables.

The set $R_{\mu A} \subseteq G_{\mu A}$ represents the genes that are regulated between the two microarray experiments $q^1_{\mu A}$ and $q^2_{\mu A}$ which reflect the differences between conditions C^1 and C^2 and between the measured dependent variables M^1 and M^2. In the formalism defined for the microarray experiments this set can be detected by a class of functions called $\delta_{\mu A}$ that, in general, takes two datasets and returns the genes that are regulated[3]: $R_{\mu A} = \delta_{\mu A}(E_{q^1_{\mu A}}, E_{q^2_{\mu A}})$. In the PHO microarray experiment [5], the $\delta_{\mu A}$ is based on the two-fold derepression ratio, and the set of regulated genes $R^{FODB}_{\mu A}$, calculated as $\delta_{\mu A}(E_{FODB_{\mu A,Cy3}}, E_{FODB_{\mu A,Cy5}})$, is

$$R^{FODB}_{\mu A} = \{\text{PHO5,PHO11,PHO12,PHO8,PHO84,PHO89,PHO86,PHO81,SPL2,PHM1,} \\ \text{PHM2,PHM3,PHM4,PHM5,PHM6,PHM7,PHM8,HOR2,CTF19,HIS1}\} \tag{2}$$

The same procedure can be applied to the expressions retrieved from the *in silico* simulations. The experiments are defined maintaining the same conditions of the microarray experiment:

$$\begin{array}{l} q^1_m = (G_m, C^1, M^1, P) \\ q^2_m = (G_m, C^2, M^2, P) \end{array} \quad \text{with} \quad \begin{array}{l} IV(q^1_m) = IV(q^2_m) \text{ and} \\ DV(q^1_m) = DV(q^2_m) \end{array} \tag{3}$$

On q^1_m and q^2_m we can now apply the classification function δ_m which is, in general, different from $\delta_{\mu A}$ because the assumptions on the distribution of the expression dataset of a microarray and of *in silico* experiment can differ. The set of regulated genes with the *in silico* approach is denoted with $R_m \subseteq G_m$ and computed with a δ_m function of the form $R_m = \delta_m(E_{q^1_m}, E_{q^2_m})$. This formula, in the case of the PHO pathway model, is $R^{SBP}_m = \delta_m(E_{q^{SBP}_{m,lp}}, E_{q^{SBP}_{m,hp}})$. Applying a simple δ_m function that detects the genes that show some expression difference in the two datasets shown in Figure 2, the set of regulated genes for the *in silico* prediction is $R^{SBP}_m = \{\text{PHO81,PHO5}\}$.

In ideal conditions we should have $R_{\mu A} = R_m$ (and $G_m = G_{\mu A}$), but in a scenario where all kind of systematic and random errors can occur the $G_{\mu A}$, G_m, $R_{\mu A}$ and R_m sets are all potentially different. We further discuss the subsets of $R_{\mu A}$ which is the set of genes we want to "clean", distinguishing the cases $G_m \subseteq G_{\mu A}$, $G_m \equiv G_{\mu A}$ and $G_{\mu A} \subseteq G_m$. Since the microarray chips can handle almost the whole genome while the computational models are still far from it (as in our case), we focus on the first case represented in Figure 3:

$R_{\mu A} \setminus G_m$ contains the microarray regulated genes which are not considered in the computational model. The genes in this set are those whose investigation can potentially increase the biological knowledge and improve the model. For the running example we have $R^{FODB}_{\mu A} \setminus G^{SBP}_m = \{$PHO11,PHO12,PHO8,PHO84,PHO89,PHO86,PHO81,SPL2,PHM1,PHM2,PHM3, PHM4,PHM5,PHM6,PHM7,PHM8,HOR2,CTF19,HIS1$\}$

$R = R_{\mu A} \cap R_m$ represents the set of genes that are regulated in both approaches. In the example $R^{PHO} = \{$PHO81,PHO5$\}$

[3]The $\delta_{\mu A}$ as used here accepts two expression datasets, but it can be generalised to consider two sets of expression data in presence of experimental designs with replicas.

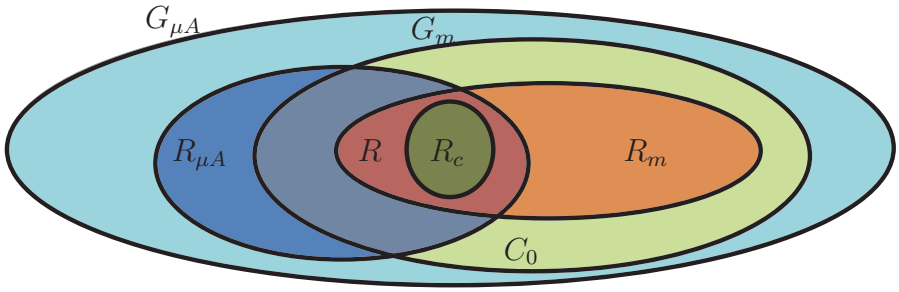

Figure 3: Representation of the possible intersections between $G_{\mu A}$ (the set of genes with a spot on the microarray), G_m (the set of genes considered in the model), $R_{\mu A}$ (the set of microarray regulated genes) and R_m (the set of *in silico* regulated genes) assuming $G_m \subseteq G_{\mu A}$.

$R_c \subseteq R$ contains, for definition, the genes of the model that are regulated in a consistent way with respect to the regulations detected by the microarray:

$$R_c = \left\{ g \in R_{\mu A} \cap R_m \;\middle|\; \frac{Expr_{q^1_{\mu A}}(g)}{Expr_{q^2_{\mu A}}(g)} \simeq \frac{Expr_{q^1_m}(g)}{Expr_{q^2_m}(g)} \right\} \tag{4}$$

The expression comparison operator ($\simeq$) is not precisely defined since different levels of consistency can be considered[4]. Also the comparison of the expression ratios is not the only comparison technique [20]. The set of the genes that are regulated consistently (with a over- or under-expressed comparison) between $R^{FODB}_{\mu A}$ and R^{SBP}_m of our example is

$$R^{PHO}_c = \{\text{PHO81,PHO5}\} \tag{5}$$

$(R_{\mu A} \cap G_m) \setminus R_m$ contains the genes regulated in the microarray but not in the *in silico* simulations. The mismatch can be due to errors, or to the lack of biological information on which the model is constructed. In the running example $(R^{FODB}_{\mu A} \cap G^{SBP}_m) \setminus R^{SBP}_m = \emptyset$

$C_0 = (G_{\mu A} \cap G_m) \setminus (R_{\mu A} \cup R_m)$ the set of constitutive (i.e. not regulated) genes in both microarray and *in silico* experiments. In the running example $C^{PHO}_0 = \{\text{PHO2,PHO4}\}$

$R_m \setminus R$ the set of genes that are regulated only in the *in silico* experiments. In our example $R^{SBP}_m \setminus R^{PHO} = \emptyset$.

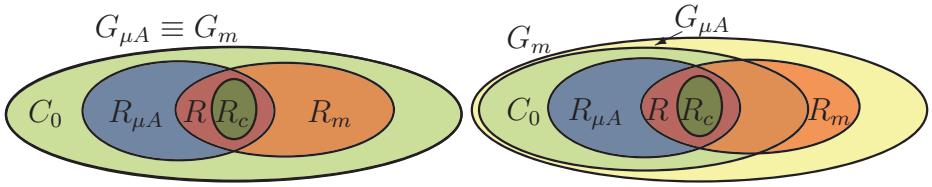

Figure 4: Representation of $G_{\mu A}$, G_m, $R_{\mu A}$ and R_m assuming $G_m \equiv G_{\mu A}$ and $G_{\mu A} \subseteq G_m$.

The cases $G_m \equiv G_{\mu A}$ and $G_{\mu A} \subseteq G_m$ are represented in Figure 4. Both cases are nowadays not realistic and they foresee the scenarios of genome wide modelling and synthetic biology modelling respectively. The main differences with the $G_m \subseteq G_{\mu A}$ case are that $R_{\mu A} \setminus G_m = \emptyset$ and that in the $G_{\mu A} \subseteq G_m$ case there is the $R_m \setminus G_{\mu A}$ set of the genes considered in the computational model but not on the microarray chip.

[4]The most restrictive is the $=$, while the less restrictive is an operator that check if the gene is over or under expressed (expression ratio different from 1) in both approaches, but other intermediate operators can be defined.

3.1 Accuracy and coverage measures

From a methodological point of view, for the consistently regulated genes R_c and the consistently costitutive genes C_0, the modelling approach agrees with the microarray one, while the genes in $R \setminus R_c$, $(R_{\mu A} \cap G_m) \setminus R$ and $(R_m \cap G_{\mu A}) \setminus R$ represent an incongruence between the approaches. With these sets we define the following measures of accuracy and coverage:

$$A = \frac{|R_c| + |C_0|}{|G_{\mu A} \cap G_m|} \times 100 \quad Cov = \frac{|G_{\mu A} \cap G_m|}{|G_{\mu A}|} \times 100 \quad Cov_e = \frac{|R_{\mu A} \cap G_m|}{|R_{\mu A}|} \times 100 \quad (6)$$

The accuracy A measure the level of mutual reliability of computational modelling (and indirectly of the biological knowledge) and of microarray expression data. Cov is the chip coverage and it reflects the coverage level of the model with respect to the microarray chip. It is an *a priori* quantification of the percentage of the genes in the microarray chip that are also considered in the model. However, a low chip coverage does not implies that the approach is useless because the microarray regulated genes can be highly covered by the model even if the chip is poorly covered. So another coverage measure, that we call experimental coverage (Cov_e), is introduced. The Cov_e can be calculated only when both the microarray and the *in silico* experiment are performed and it is referred only to a single experiment, but it is a more accurate evaluation of the experimental quality. Moreover, minimizing the chip coverage after maximizing the experimental coverage, gives the same results of a complete chip coverage, but it requires a much lower computational effort for the *in silico* experiments. In any case, the coverage and the accuracy must be considered together for evaluating the overall quality of the approach for a performed experiment.

Observing on the running example that $|R_c^{PHO}| = 2$, $|C_0^{PHO}| = 2$, $|G_{\mu A}^{FODB} \cap G_m^{SBP}| = |G_m^{SBP}| = 4$, $|G_{\mu A}^{FODB}| = 6400$, $|R_{\mu A}^{FODB}| = 20$ and $|R_{\mu A}^{FODB} \cap G_m^{SBP}| = 2$, the introduced measure are: $A^{PHO} = 100\%$, $Cov^{PHO} = 0,0625\%$ and $Cov_e^{PHO} = 10\%$.

This means that the model predicts the expression profiles consistently with the microarray experiment, but the coverage is very low as we expected since the model is local. The relatively high experimental coverage means that the model fits at least partially the pathways whose genes are interested in the regulation.

4 Prospective applications in microarray differential analysis

The set of genes detected with the microarray differential analysis whose regulation is in relation with a specific condition, can be very large and so hard to analyse. The applications we propose here are based on the described framework and are intended to be applied on the set of regulated genes in order to detect the most informative genes before further analysis.

4.1 Removing the genes regulated in the *in silico* predictions

The set of regulated genes can include many genes that are already indirectly known to be related with the experimental conditions. The removing of these genes allows the biologist to focus only on the really unknown genes. The idea is to use computational models, built on the

current biological knowledge, to predict the genes that will be regulated in a specific microarray experiment and remove them from the microarray experimental results.

The aim of this application is to filter out from the regulated genes of the microarray experiment ($R_{\mu A}$), those genes that the model suggests to be regulated (R_m). We would remove directly the genes in R_m from the genes in $R_{\mu A}$. In other words the set of really interesting genes R_{int} would be $R_{int} = R_{\mu A} \setminus R_m$. In presence of errors and approximations, this definition is no more acceptable mainly because the uncertainty on R_m. We need a more robust definition of the set of genes that can be safely removed from the microarray regulated ones. So instead of subtracting from $R_{\mu A}$ the R_m set, we subtract the genes that are regulated in a consistent way in both approaches, i.e. the R_c set as defined in (4). So $R_{int} = R_{\mu A} \setminus R_c$.

Returning to the example, this definition is $R^{PHO}_{int} = R^{FODB}_{\mu A} \setminus R^{PHO}_c$ where the $R^{FODB}_{\mu A}$ and R^{PHO}_c are computed as shown in (2) and (5). The resulting set of really interesting genes is

$$R^{PHO}_{int} = \{\text{PHO11,PHO12,PHO8,PHO84,PHO89,PHO86,SPL2,PHM1,PHM2,PHM3,} \\ \text{PHM4,PHM5,PHM6,PHM7,PHM8,HOR2,CTF19,HIS1}\} \tag{7}$$

Our application has filtered out from the set of microarray regulated genes of [5] two genes (PHO5 and PHO81) predicted by the computational model [19]. So we have reduced the number of genes that represent new biological information suggesting that this application is useful.

The genes considered in the model are only 4, and the two that have been removed from R^{PHO}_{int} are those that the model predicts sensitive to different phosphate metabolism. So, as the accuracy measure ($A^{PHO} = 100\%$) suggests, the model predictions are in this case the more desirable ones with respect to the very low coverage. We can conclude that this application can be a helpful tool for a biologists, especially in the cases where the microarray experiment detects an high number of regulated genes and the computational model has a reasonable good experimental coverage in addition to the accuracy.

4.2 Removing the genes regulated by the non controlled variables

The definition of the microarray experiment includes the notion of controlled conditions and dependent not-directly controlled variables. Obviously, the effects of the dependent variables in terms of regulated genes cannot be detected in isolation or separated from the independent variable effects within the same microarray experiment. However, if biological knowledge on the effects of the dependent variables is available, it is possible to incorporate this in a model. The idea is to estimate *in silico* the genes that are regulated because of the dependent variables in order to remove them from the set of microarray regulated genes. Note that the values of the dependent variables included in the specification of the *in silico* gene expression experiments are those measured contextually with the microarray experiment. Since, as seen, the microarray experiment of [5] does not provide any dependent variables, we cannot test this conceptual application with our running example.

Consider the following three microarray experiments[5]:

$$q_{\mu A} = (G, C, M, P) \qquad q^{cntr}_{\mu A} = (G, C^{cntr}, M^{cntr}, P) \qquad q^{h}_{\mu A} = (G, C, M^{cntr}, P)$$

[5] In this subsection we assume that $G = G_{\mu A} \equiv G_m$, without losing generality.

with $IV(q_{\mu A}) = IV(q^{cntr}_{\mu A}) = IV(q^{h}_{\mu A})$ and $DV(q_{\mu A}) = DV(q^{cntr}_{\mu A}) = DV(q^{h}_{\mu A})$. The first two chips are a microarray experiment ($q_{\mu A}$) and the relative control chip ($q^{cntr}_{\mu A}$), while the third ($q^{h}_{\mu A}$) is a hypothetical variation of the microarray chips in which the actual values of dependent variable are replaced with the control ones. Notice that the experiment $q^{h}_{\mu A}$ cannot be really performed since the values of the independent variables C force the values of the dependent controlled variables to be M and not M^{cntr}. Suppose one applies a $\delta_{\mu A}$ function in the following way:

$$R_{\mu A} = \delta_{\mu A}(E_{q_{\mu A}}, E_{q^{cntr}_{\mu A}}) \qquad R^{V}_{\mu A} = \delta_{\mu A}(E_{q^{h}_{\mu A}}, E_{q^{cntr}_{\mu A}}) \qquad R^{V_D}_{\mu A} = \delta_{\mu A}(E_{q_{\mu A}}, E_{q^{h}_{\mu A}})$$

$R_{\mu A}$ is the standard set of genes that are regulated because of the differences between levels of the independent and dependent variables between the two chips, $R^{V}_{\mu A}$ contains the genes that are regulated only by the independent variables (since $q^{h}_{\mu A}$ and $q^{cntr}_{\mu A}$ have the same values of dependent variables M^{cntr}), and $R^{V_D}_{\mu A}$ represents the genes that are regulated only by the dependent variables (since $q_{\mu A}$ and $q^{h}_{\mu A}$ have the same values of the independent variables C). $R^{V}_{\mu A}$ is the set of genes that the experimenter would have because it is not influenced by the dependent variables, but it is not possible to obtain because the $q^{h}_{\mu A}$ chip is only hypothetical.

Under the assumption that the sets of genes regulated by the dependent and independent variables are disjoint, the set of genes regulated by the independent variables in the microarray experiment $R^{V}_{\mu A}$ could be estimated subtracting from the regulated genes $R_{\mu A}$ those genes that are regulated because of the dependent variable $R^{V_D}_{\mu A}$. So

$$R^{V}_{\mu A} = R_{\mu A} \setminus R^{V_D}_{\mu A} \tag{8}$$

but also $R^{V_D}_{\mu A}$ is not possible to obtain, since it needs the hypothetical $q^{h}_{\mu A}$ chip.

However, in this framework we have the possibility to estimate expression experiments with the computational model. In particular, all the following experiments can be performed *in silico*:

$$q_m = (G, C, M, P) \qquad q^{cntr}_m = (G, C^{cntr}, M^{cntr}, P) \qquad q^{h}_m = (G, C, M^{cntr}, P)$$

With these *in silico* experiments it is possible to detect the gene regulated only by the independent (R^{V}_{m}) and only by the dependent variables ($R^{V_D}_{m}$): $R^{V}_{m} = \delta_m(E_{q^{h}_m}, E_{q^{cntr}_m})$ and $R^{V_D}_{m} = \delta_m(E_{q_m}, E_{q^{h}_m})$. The direct approximation of $R^{V}_{\mu A}$ with R^{V}_{m} is useless because in this way we rely only on the model, losing the information of the microarray experiment.

The idea for integrate the microarray and the *in silico* data consists in filtering out from the set of microarray regulated genes those genes that are regulated because of the dependent variables, substituting $R^{V_D}_{\mu A}$ with $R^{V_D}_{m}$ in (8): $R^{V}_{\mu A} \simeq R_{\mu A} \setminus R^{V_D}_{m}$. If a gene is regulated both because of the independent and because of the dependent variables, the estimation of $R^{V}_{\mu A}$ will not include that gene. For this reason is necessary to assure that the independent variables V and the dependent variables V_D regulate two different set of genes meaning that the influence on gene regulation of the two set of variables is disjoint.

Definition 4.1 *V and V_D are disjoint with respect to the gene regulation (or ge-disjoint) if and only if for every possible values associated to the variables the following holds:* $R^{V}_{\mu A} \cap R^{V_D}_{\mu A} = \emptyset \ \wedge \ R^{V}_{m} \cap R^{V_D}_{m} = \emptyset$

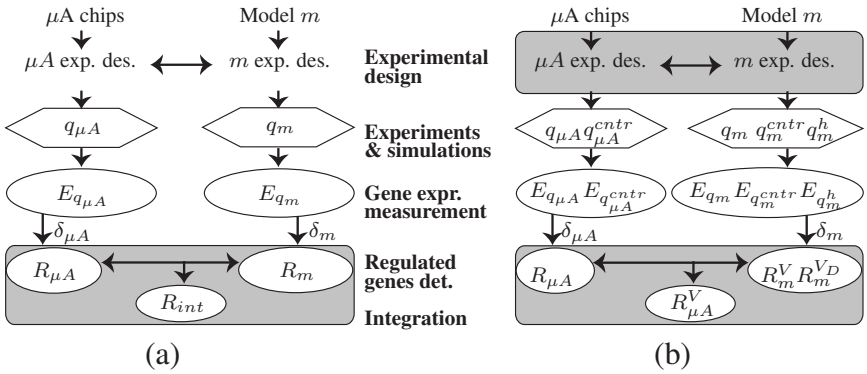

Figure 5: The modules interested by the application for removing the *in silico* regulated genes (a) and by the application for removing the genes regulated by non controlled variables (b).

This definition is too restrictive because it is impossible to directly obtain $R^V_{\mu A}$ and $R^{V_D}_{\mu A}$. However, the definition can be made less strict for an effective application if we have an high accuracy and a good confidence in the quantitative estimation of gene expression of the model.

Definition 4.2 *V and V_D are ge-disjoint, if and only if for every possible values associated to the variables the following holds:* $R^V_m \cap R^{V_D}_m = \emptyset$

The complete definition for removing from the set of regulated genes of a microarray, the genes that are regulated only by the dependent variables, is $R^V_{\mu A} \simeq R_{\mu A} \setminus R^{V_D}_m$ if $R^V_m \cap R^{V_D}_m = \emptyset$.

If V and V_D are not ge-disjoint (or if it is not possible to show it), we can partition V_D in two subsets $V'_D, V''_D \subset V_D$ with $V'_D \cap V''_D = \emptyset$ and $V'_D \cup V''_D = V_D$ such that V'_D and V are ge-disjoint. We can still remove some genes from $R_{\mu A}$, and precisely the genes regulated because of V'_D.

So in the cases where it is possible to show that the set of independent variables (or a subset) are ge-disjoint from the set (or from a subset) of the dependent measured variables, we can remove all (or some of) the genes that are not directly regulated by the experimental conditions.

5 From a conceptual to a software framework

Our notational framework can be a guide for the development of software tools for supporting analyses that combine the two fields of the microarray technology and the system level modelling and simulation of biological networks. While for the first a lot of bioinformatics tools have been developed for every aspect of the technology, the second still needs research in order to make the development and the simulation of systems biology models effective.

Our conceptual framework assumes to have a computational model and microarray chips and can actively act on different phases: the experimental design, the experiments and simulations, the gene expression measurement, the regulated genes detection and the integration of regulated genes. The first application (Figure 5.a) concerns only the manipulation of the regulated genes belonging to microarray and *in silico* experiments designed with the same conditions, while the second (Figure 5.b) interests also the design, requiring some experiments with particular settings of the dependent and independent conditions. Other applications can regard also

other phases, and for this reason we discuss the availability of implemented software or formal specification for each phase.

Experimental design In-depth studies have been done to make microarray experiments maximally informative, given the effort and the resources [21, 22]. In this work we assume that an *in silico* simulated microarray experiment has the same abstract behaviour of the real one; for this reason all the designs reported in literature can be applied also to the simulated case. However, some improvements can be done relying on the fact that once the model is developed the simulation cost is much lower than the microarray experimental one. Moreover some hybrid designs are possible integrating in the same design simulated and real microarray experiments.

Experiments and simulations The microarray experimental procedure has reached a good level of standardization [2]. Instead, the possible sources of variability in the computational simulation rely on model errors and approximations in the simulation algorithm; however, both problems regard in general the computational modelling of biological systems, while the simulation procedure is intrinsically standard.

Gene expression measurement The output of a microarray experiment is obtained with the optical scan of the array and the analysis of the resulting image. Software tools for this operation are available. Both microarray output after scanning and *in silico* expression results, needs normalization procedures to make meaningful comparisons of expression levels. For microarray experiments, statistical methods are available [18], while for the *in silico* simulations specific normalization procedures must be developed because the distribution of the expression can be different from the microarray ones. Since *in silico* microarray outputs are not available, more precise discussions cannot be done, but it is reasonable to adapt some microarray techniques with specific parameters and thresholds.

Regulated genes detection Statistical techniques for $\delta_{\mu A}$ are available and implemented [20]. δ_m, instead, was never developed, but the tuning of some $\delta_{\mu A}$ functions is reasonable, considering also that the *in silico* simulations are not affected by the random experimental error and by approximation in the optical scanning of the chip and so they are less noisy.

Integration of regulated gene sets The integration of the set of regulated genes are normal operations on sets and can thus be easily implemented.

6 Conclusions

The traditional use of biological models concerns the parameters estimation and the qualitative and quantitative description of not directly observable and high level behaviours. After the paradigmatic shift of systems biology there is the need to integrate high-throughput data and the *in silico* results of genomic-scale compositional models. In this work we propose a framework that allows the use of systems biology models to support analysis of high-throughput data, whose final goal is to compare the current biological knowledge with new genomic experiments. The comparison aims to discover unknown aspects of complex and wide biological networks allowing to focus further investigations only on that very specific subnetworks. The overall procedure is somehow recursive since the new discoveries reached starting from the

model suggestions, permits improvements of the biological knowledge from which it is possible to construct more precise models.

We introduced a unifying notational framework for representing the microarray experiments and the *in silico* simulations of gene expression. The notation was specialized in the context of differential gene expression analysis. Then, we proposed two prospective applications of our conceptual framework with the purpose of supporting the mining of regulated genes detected with microarray expression data. Reference to the real example of the stochastic π-calculus PHO pathway model [19] and of the PHO microarray experiment of [5], was systematically made. The same example, though not with a genomic scale model, was also applied to the first application in which we were able to remove two genes from the set of regulated genes of the microarray experiment, thus obtaining encouraging results and highlighting the utility of the approach. The second application, instead, pointed out how it is possible for a computational model to predict the genes regulated by the dependent measured variables only, in order to focus on the genes regulated by the independent variables. Finally we discuss for each phase, namely the experimental design, the experiments and simulations, the gene expression measurement, the regulated genes detection and the integration of regulated genes, which are the aspects that need further investigation to obtain an effective software framework.

To our best knowledge, this is the first attempt to tackle directly the problem of integration of computational systems biology modelling and high-throughput expression data. Currently, genomic-scale models are not yet available, but as the model size grows reaching the systems biology level the impact of our integration can be very important for bringing together biological knowledge and high-throughput experimental research.

References

[1] J.M. Bower and H. Bolouri. *Computational Modeling of Genetic and Biochemical Networks*. MIT Press, 2001.

[2] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, et al. Minimum information about a microarray experiment (miame)-toward standards for microarray data. *Nat Genet*, 29(4):365–71, 2001.

[3] D. Chaussabel and A. Sher. Mining microarray expression data by literature profiling. *Genome Biol*, 3(10):research0055.1–0055.16, 2002.

[4] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.

[5] G.R. Fink, N. Ogawa, J. DeRisi, and P.O. Brown. New components of a system for phosphate accumulation and polyphosphate metabolism in saccharomyces cerevisiae revealed by genomic expression analysis. *Mol Biol Cell*, 11(12):4309–4321, 2000.

[6] G.W. Hatfield, S. Hung, and P. Baldi. Differential analysis of dna microarray gene expression data. *Mol Microbiol*, 47(4):871–877, 2003.

[7] A. Hoffmann, A. Levchenko, M.L. Scott, and D. Baltimore. The ikappa b-nf-kappa b signaling module: Temporal control and selective gene activation. *Science*, 298(5596):1241, 2002.

[8] C.Y.F. Huang and J.E. Ferrell Jr. Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc Natl Acad Sci USA*, 93:10078–10083, 1996.

[9] S. Huang. Back to the biology in systems biology: What can we learn from biomolecular networks? *Brief Funct Genomic Proteomic*, 2(4):279, 2004.

[10] P. Khatri and S. Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, 2005.

[11] H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

[12] H. Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, 2002.

[13] D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotech*, 14:1675–1680, 1996.

[14] P. Pavlidis, J. Qin, V. Arango, J.J. Mann, and E. Sibille. Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res*, 29(6):1213–1222, 2004.

[15] M. Peleg, I. Yeh, and R.B. Altman. Modelling biological processes using workflow and petri net models. *Bioinformatics*, 18(6):825–837, 2002.

[16] M.S. Pepe, G. Longton, G.L. Anderson, and M. Schummer. Selecting differentially expressed genes from microarray experiments. *Biometrics*, 59(1):133–142, 2003.

[17] C. Priami and P. Quaglia. Modelling the dynamics of biosystems. *Brief Bioinform*, 5(3):259–269, 2004.

[18] J. Quackenbush. Microarray data normalization and transformation. *Nat Genet*, 32:496–501, 2002.

[19] N. Segata, E. Blanzieri, and C. Priami. Stochastic π-calculus modelling of multisite phosphorylation based signaling: in silico analysis of the pho4 transcription factor and the pho pathway in saccharomyces cerevisiae. Technical Report TR-08-2007, Microsoft Research - University of Trento Centre for Computational and Systems Biology, 2007.

[20] V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, 98(9):5116–5121, 2001.

[21] Y.H. Yang and T. Speed. Design issues for cdna microarray experiments. *Nat Rev Genet*, 3(8):579–588, 2002.

[22] S.O. Zakharkin, K. Kim, T. Mehta, L. Chen, S. Barnes, K.E. Scheirer, R.S. Parrish, D.B. Allison, and G.P. Page. Sources of variation in affymetrix microarray experiments. *BMC Bioinformatics*, 6(214):1471–2105, 2005.