Gene Expression Array exploration using $\mathcal{K}$-Formal Concept Analysis

J.M. González Calabozo, Carmen Peláez-Moreno, and Francisco J. Valverde-Albacete [*]

Dpto. de Teoría de la Señal y de las Comunicaciones.
Universidad Carlos III de Madrid
Avda. de la Universidad, 30. Leganés 28911. Spain
fva,carmen,jmgc@tsc.uc3m.es

Abstract. DNA micro-arrays are a mechanism for eliciting gene expression values, the concentration of the transcription products of a set of genes, under different chemical conditions. The phenomena of interest—up-regulation, down-regulation and co-regulation—are hypothesized to stem from the functional relationships among transcription products. In [1,2,3] a generalisation of Formal Concept Analysis was developed with data mining applications in mind, $\mathcal{K}$-Formal Concept Analysis, where incidences take values in certain kinds of semirings, instead of the usual Boolean carrier set. In this paper, we use $(\overline{\mathbb{R}}_{\min,+})$- and $(\overline{\mathbb{R}}_{\max,+})$-Formal Concept Analysis to analyse gene expression data for *Arabidopsis thaliana*. We introduce the mechanism to render the data in the appropriate algebra and profit by the wealth of different Galois Connections available in Generalized Formal Concept Analysis to carry different analysis for up- and down-regulated genes.

1 Introduction

The *transcriptome* of a species is the set of gene expression products, be they proteins or messenger RNA (mRNA) chains. DNA micro-arrays are a mechanism to take measures of such data in the form of an *expression profile*, a record of the concentration of different mRNA associated to a subset of the species genome with respect to a *condition*, a particular state or sequence of states undergone by the cells under study. Roughly, each of these mRNA sequences comes from the expression of a particular gene and is translated into a protein inside ribosomes.

Transcriptomics studies these expression profiles for multiple purposes: *body maps*—creating records of baseline abundance of mRNA in different tissues—, *case vs. control studies*—studying particular states vs a control profile—, *parsing pathways*—elucidating the signalling networks associated to sets of genes— and

[*] This work has been partially supported by the Spanish Government-Comisión Interministerial de Ciencia y Tecnología projects 2008-06382/TEC and 2008-02473/TEC and the regional projects S-505/TIC/0223 (DGUI-CM) and CCG08-UC3M/TIC-4457 (Comunidad Autónoma de Madrid - UC3M).

studying *functional response patterns*, the exploration of a systematically varied set of conditions in the expectation that *co-regulation* of genes across a set of biological conditions reveals functional gene groups [4].

In this context, the concentration of the transcribed product (usually mRNA) is the *(gene) expression value*, and the expression values of a set of genes under the same condition an *expression profile.* Therefore, given a *genome* —a set of *genes*—$G = \{g_i\}_{i=1}^{n}$ the *gene expression data* taken to analyse their functional influence consists of the expression value of every gene C_{ij}—an expression profile—under one condition m_j in a non-explicitly given set of conditions $M = \{m_j\}_{j=1}^{p}$, which grows as we take more measurements.

Under these premises, *co-regulation* refers to the increment or decrement of the expression value in a set of genes brought about by the change in expression value of other genes. At each condition and for each gene, co-regulation results either in *up-regulation*, an increment in expression value, or *down-regulation*, a decrement in expression value, and these changes are expected to reveal functional relations between genes.

This emphasis on up-regulation and down-regulation make gene profile exploration an ideal candidate to be explored by means of $\mathcal{K}$-Formal Concept Analysis, a flavour of Formal Concept Analysis where incidences take value in a multi-valued algebra $\mathcal{K}$ which is an idempotent semifield—an analogue of a field replacing addition with an idempotent law [1,2,3].

In this paper we will undertake the exploration of expression profiles with $(\overline{\mathbb{R}}_{\min,+})$- $(\overline{\mathbb{R}}_{\max,+})$-Formal Concept Analysis with the purpose of researching into functional response patterns. For that purpose, in Section 2 we review data-preprocessing, $\mathcal{K}$-Formal Concept Analysis and lattice-building procedures applied to expression profiles. Next we describe our results in a database of *Arabidopsis thaliana* profiles, and conclude in Section 4 by comparing ours to previous work on using Formal Concept Analysis on gene expression data.

2 Methods and tools

2.1 Data preparation

The main problem with expression data is noise: mRNA concentrations profiles are irreproducible from experiment to experiment due to conditions difficult or impossible to control—such as the thermodynamic environment of reactions or individual specimen ontogenesis, respectively. Besides, measurement techniques also introduce their own kind of noise, since they are also based in chemical reactions—hybridization of mRNA with fluorescent markers. For this reason most measurements are repeated a number of times for each condition. Sometimes these measurements are used to obtain variance- and mean-normalized profiles for each condition. Finally, an actual profile for condition m_j is obtained which we gather in a single matrix C_{ij} of *positive numbers* where i runs over genes and j over conditions.

For each experiment, a special kind of profile, called a *control*, may be measured as a reference for other measurements. Controls are adapted to the kind of experiment and might be the profile of a mix of cells of a whole specimen—to obtain a *body map*—or a particular mix of specific cells under study—for instance, healthy cells to be compared against cancerous cells. Since controls may be extracted from population of specimens grown in controlled conditions, they are expected to be less noisy. In our experiments, we designate a set of measurements for the same condition as controls and coalesce them into their geometrical mean $\bar{c}_i$. This produces a single control at the expense of reducing the set of measurements.

Since both up-regulation and down-regulation of genes occur in gene expression we would like to cater to exploring both. All profiles excepting controls are entry-wise normalized by the control profile and their logarithm[1] taken to make the resulting number range in $[-\infty, \infty]$ $R_{ij} = \log \frac{C_{ij}}{\bar{c}_i}$. Log-quotients of gene expression values are

- $R_{ij} \leq 0$ if g_i is down-regulated by m_j ,
- $R_{ij} \geq 0$ if g_i is up-regulated by m_j , and
- $R_{ij} = 0$ if the control and the condition expression value are equal.

Call the doubly completed set of reals $\overline{\overline{\mathbb{R}}} = \mathbb{R} \cup \{\pm\infty\}$. The reasoning above would suggest using as carrier set for log-quotient values $\overline{\overline{\mathbb{R}}}$ where further:

- $R_{ij} = \log \frac{0}{k} = -\infty, k \neq 0$ when g_i is not expressed at all in m_j ,
- $R_{ij} = \log \frac{k}{0} = \infty$ when g_i is not expressed in the control condition.

With $|G| = n$, $|M| = p$, we collect all expression profiles into a $(\overline{\overline{\mathbb{R}}})$-valued matrix $R \in \overline{\overline{\mathbb{R}}}^{n \times p}$, and call the triple (G, M, R) a *multi-valued formal context*, where $R_{ij} = \lambda$ reads as "the expression value of gene g_i in condition m_j is λ" . The procedure to obtain specific concept lattices from this context is roughly sketched in the next subsection.

2.2 $\mathcal{K}$-Formal Concept Analysis of expression data

A generalisation of Formal Concept Analysis called $\mathcal{K}$-Formal Concept Analysis (kFCA) was introduced in [1,2,3] to cater for the notion of a *degree of incidence*, where $\overline{\mathcal{K}}$ is a complete idempotent semifield $\mathcal{K} = \langle K, \oplus, \otimes, \cdot^{-1}, \bot, e, \top \rangle$. This allows the analysis of real-valued incidences by embedding them into a convenient algebra, to be investigated next.

$\mathcal{K}$-Formal Concept Analysis. Complete idempotent semifields are already lattices with $a \wedge b = a \oplus b, a \vee b = a \otimes (a \oplus b)^{-1} \otimes b$. For a complete idempotent semifield a *semimodule* or vector space $\overline{\mathcal{K}}^n = \langle \overline{K}^n, \oplus, \bot_n \rangle$ is an additive monoid with a scalar multiplication inherited from the multiplication in the semifield. A unitary vector e_i in this vector space is $e_i(i) = e$ and $e_i(k) = \bot_{\mathcal{K}}, i \neq k$. Notice that semimodules have an order induced by that of the underlying semiring. In

[1] All logarithms are base 2 in this paper.

the case of idempotent semifields, this order is compatible with the $\oplus$ operation $x \leq y \Leftrightarrow x \oplus y = y$ turning them into join-semilattices.

Matrices over completed idempotent semifields $R \in \overline{\mathcal{K}}^{n \times p}$are linear forms between vector spaces. For the analysis of expression values we call:

- a row vector in $\mathcal{Y} = \overline{\mathcal{K}}^{1 \times n}$ a *$\mathcal{K}$-set of genes*,
- a column vector in $\mathcal{X} = \overline{\mathcal{K}}^{p \times 1}$ a *$\mathcal{K}$-set of conditions*,
- a column vector in the range of R, $\mathrm{Im}(R) \subseteq \overline{\mathcal{K}}^{n \times 1}$ a *(gene) expression profile*,
- a row vector in the range of R^{T}, $\mathrm{Im}(R^{\mathrm{T}}) \subseteq \overline{\mathcal{K}}^{1 \times p}$ a *condition profile*.

Note that DNA micro-arrays actually obtain a set expression values for a particular condition m_j later transformed into an expression profile $p(m_j) = R \otimes e_j$ (§2.1). However, the *condition profile* for g_i, the vector of its expression values for different conditions $q(g_i) = e_i^{\mathrm{T}} \otimes R$ is seldom considered of interest in analyses.

Consider the context $(G, M, R)_{\overline{\mathcal{K}}}$ and row- and column-vector spaces $\mathcal{Y} = \overline{\mathcal{K}}^n$ and $\mathcal{X} \equiv \overline{\mathcal{K}}^p$. The bracket $\langle y \mid R \mid x \rangle = y \otimes R \otimes x$ between left and right vector spaces over $\overline{\mathcal{K}}$ is proven in [3] to induce a Galois connection $[(\cdot)^{+}_{R,\varphi}, {}^{+}_{R,\varphi}(\cdot)] : \overline{\mathcal{K}}^n \leftrightharpoons \overline{\mathcal{K}}^p$. Given an invertible $\varphi \in K$, the *φ-polars* are the dually adjoint maps

$$(y)^{+}_{R,\varphi} = \bigvee \{\, x \in X \mid \langle y \mid R \mid x \rangle \leq \varphi \,\} \quad {}^{+}_{R,\varphi}(x) = \bigvee \{\, y \in Y \mid \langle y \mid R \mid x \rangle \leq \varphi \,\}$$

For row- and column-vectors a and b , the *φ-formal concept* $(a, b)_\varphi$ is a pair such that $(a)^{+}_{R,\varphi} = b$ and ${}^{+}_{R,\varphi}(b) = a$ with a the *φ-extent* and b the *φ-intent*. The parameter $\varphi \in K$ is called the *threshold of existence* and it can be proven to describe a *maximum expression value* allowed for pairs $(a, b) \in \overline{\mathcal{K}}^n \times \overline{\mathcal{K}}^p$ to be considered as members of the φ-formal concept set $\mathfrak{B}^\varphi(G, M, R)_{\overline{\mathcal{K}}}$ [3]. As usual, φ-concepts can be ordered by extents or dually by intents

$$(a_1, b_1) \leq (a_2, b_2) \Leftrightarrow a_1 \leq a_2 \Leftrightarrow b_1 \leq^{\mathrm{d}} b_2 \tag{1}$$

and the set of φ-concepts with this order is the φ-concept lattice $\underline{\mathfrak{B}}^\varphi(G, M, R)_{\overline{\mathcal{K}}}$.

A drawback for data mining purposes is that the φ-concept lattice, has a huge number of concepts—infinite, in the typical case—and is hard to visualize. Therefore, we define the *structural (gene expression) lattice* $\underline{\mathfrak{B}}(G, M, I_R^\varphi)$ of the φ-concept lattice as the concept lattice of a binary incidence, I_R^φ , *related to R and intended to focus on those concepts below a threshold of existence φ.*

The following is a procedure to *build and explore* a structural lattice:

Step 1 Fix a threshold φ. Compute the closures of the n unitary row vectors of dimension $1 \times n$, $\gamma(e_i) = \left({}^{+}_{R,\varphi}((e_i)^{+}_{R,\varphi}), (e_i)^{+}_{R,\varphi}\right)$ and p unitary column vectors of dimension $p \times 1$, $\mu(e_j) = \left({}^{+}_{R,\varphi}(e_j), ({}^{+}_{R,\varphi}(e_j))^{+}_{R,\varphi}\right)$.

Step 2 Define a binary incidence I_R^φ between genes and conditions associated to those concepts by $g_i I_R^\varphi m_j \Leftrightarrow \gamma(e_i) \leq \mu(e_j)$.

Step 3 Use a standard tool for Formal Concept Analysis—ConExp [**?**]—to build and visualize the standard lattice $\underline{\mathfrak{B}}\,(G, M, I_R^\varphi)$.

Because the procedure that selects the formal concepts depends on the threshold φ, typically the algorithm above must be carried out a number of times—one for each choice of φ that is deemed interesting—a process we call *lattice exploration*. This allows us to analyse non-boolean expression matrices using several thresholds of existence.

The choice of idempotent semiring. For the case at hand, therefore, a proper choice for $\overline{\mathcal{K}}$ is $\overline{\mathbb{R}}_{\max,+}$ (read "completed max-plus semiring"), actually an *idempotent semifield*:

$$\overline{\mathbb{R}}_{\max,+} = \langle \overline{\underline{\mathbb{R}}}, \max, \underset{\cdot}{+}, -\cdot, -\infty, 0, \infty \rangle$$

This is the completed set of reals with the "max" operation used as addition and normal addition as multiplication, and subtraction as the multiplicative inverse. As noted elsewhere, completed idempotent semifields come in dually ordered pairs[3, §2.2.2]. The order dual of $\overline{\mathbb{R}}_{\max,+}$ is $\overline{\mathbb{R}}_{\min,+}$, the completed min-plus semiring

$$\overline{\mathbb{R}}_{\min,+} = \langle \overline{\underline{\mathbb{R}}}, \min, \overset{\cdot}{+}, -\cdot, \infty, 0, -\infty \rangle$$

Notice that then $\top_{\overline{\mathbb{R}}_{\min,+}} = -\infty, \bot_{\overline{\mathbb{R}}_{\min,+}} = \infty$ and $-\cdot$ is actually a dual order isomorphism between both lattice structures. In this notation we have $-\infty \underset{\cdot}{+} \infty = -\infty$ and $-\infty \overset{\cdot}{+} \infty = \infty$, which solves several issues in dealing with the separately completed dioids. This structure actually carries a complete lattice structure

$$\langle L, \vee, \wedge, \bot, \top \rangle := \langle \overline{\underline{\mathbb{R}}}, \max, \min, -\infty, \infty \rangle \ .$$

.

Therefore we posit this structure as an appropriate means for modelling increments with respect to an average value.

Exploring down-regulation with $(\overline{\mathbb{R}}_{\max,+})$-Formal Concept Analysis. By taking $\overline{\mathcal{K}} := \overline{\mathbb{R}}_{\max,+}$ and the bracket $\langle y \mid R \mid x \rangle = y \underset{\cdot}{\otimes} R \underset{\cdot}{\otimes} x$ the *polars* are the dually adjoint maps [2]

$$\begin{aligned}(y)^{+}_{R,\varphi} &= (y \underset{\cdot}{\otimes} R) \underset{\cdot}{\backslash} \varphi & {}^{+}_{R,\varphi}(x) &= \varphi \underset{\cdot}{/} (R \underset{\cdot}{\otimes} x) \\ &= R^{\circledast} \overset{\cdot}{\otimes} y^{\circledast} \overset{\cdot}{\otimes} \varphi & &= \varphi \overset{\cdot}{\otimes} x^{\circledast} \overset{\cdot}{\otimes} R^{\circledast}\end{aligned} \qquad (2)$$

Recall that $e = 0$ is the *unit for multiplication in* $\overline{\mathbb{R}}_{\min,+}$. Since $\langle y \mid R \mid x \rangle = \max_{i,j}\{y_i \underset{\cdot}{+} R_{ij} \underset{\cdot}{+} x_j\}$ selects the highest expression value(s) in R_{ij} subject to the weights in y_i and x_j which act as focusing mechanisms, by keeping $y_i = 0 = x_j$ and $\varphi \leq 0$ we concentrate on negative expression values $R_{ij} \leq 0$, that is down-regulated genes in the concepts defined by (2). Hence to find down-regulated genes of (G, M, R) we have to explore $\underline{\mathfrak{B}}^{\varphi}(G, M, R)_{\overline{\mathbb{R}}_{\max,+}}$ with $\varphi \in (-\infty, 0]$.

[2] Notice how the polars are given a closed expression in the *dual idempotent semifield* $\overline{\mathbb{R}}_{\min,+}$.

Exploring up-regulation with ($\mathbb{R}_{\min,+}$)-Formal Concept Analysis. To cater to up-regulated genes we simply consider matrix R to be part of a the context $\overline{\mathbb{R}}_{\min,+}$-*valued formal context* $(G, M, R)_{\overline{\mathbb{R}}_{\min,+}}$. By taking the bracket $[y \mid R \mid x] = y \overset{\centerdot}{\otimes} R \overset{\centerdot}{\otimes} x$ the dually adjoint maps over the *dual order* now define a *minimum degree of existence* required for pairs of vectors to be considered ϕ-concepts.

$$
\begin{aligned}
(y)^{+}_{R,\phi} &= \bigwedge \{ x \in X \mid [x \mid R \mid y] \geq \phi \} \quad & {}^{+}_{R,\phi}(x) &= \bigwedge \{ y \in Y \mid [x \mid R \mid y] \geq \phi \} \\
&= R^{\circledast} \underset{\centerdot}{\otimes} y^{\circledast} \underset{\centerdot}{\otimes} \phi & &= \phi \underset{\centerdot}{\otimes} x^{\circledast} \underset{\centerdot}{\otimes} R^{\circledast}
\end{aligned}
\tag{3}
$$

Since $[y \mid R \mid x] = \min_{i,j}\{y_i \overset{\centerdot}{+} R_{ij} \overset{\centerdot}{+} x_j\}$ selects the lowest expression value(s) in R_{ij} subject to the weights in y_i and x_j—which act as a masking mechanisms—by keeping $y_i = 0 = x_j$ and $\phi \geq 0$ we concentrate on *positive* expression values $R_{ij} \geq 0$, that is up-regulated genes in the concepts defined by (3). Hence to find up-regulated genes of (G, M, R) we have to explore $\underline{\mathfrak{B}}^{\phi}(G, M, R)_{\overline{\mathbb{R}}_{\min,+}}$ with $\phi \in (0, \infty)$.

Note that since the unitary vectors in $\overline{\mathbb{R}}^{n}_{\min,+}$ are $(e_i)^{-1}$, another way of exploring $\underline{\mathfrak{B}}^{\phi}(G, M, R)_{\overline{\mathbb{R}}_{\min,+}}$ with $\phi \in (0, \infty)$ is to explore $\underline{\mathfrak{B}}^{-\phi}(G, M, -R)_{\overline{\mathbb{R}}_{\max,+}}$.

3 Results

3.1 Data conditioning

We selected transcriptomic data for *A. thaliana* to analyse the behaviour of the root and the shoots in a Selenium-rich environment. The data used for this simulation was downloaded from the NCBI database [3], the same data has been analysed in [**?**]. The data come from an Affymetrix Arabidopsis ATH1 Genome Array [**?**] which measures concentration of predefined mRNA sequences in a given biological sample.

We perform this preprocessing with the *Bioconductor* R-package as in [**?**] which also allows MAS preprocessing. A different comparison among different preprocessing [**?**] types suggests that RMA—also supported by Bioconductor—can provide better results, but MAS preprocessing seems to be more widely accepted.

The data has 8 different gene expression profiles:

- root tissues, two control samples: **root1** and **root2**
- root tissues, two samples with Selenium: **rootSe1** and **rootSe2**
- shoot tissues, two control samples: **shoot1** and **shoot2**
- shoot tissues, two samples with Selenium: **shootSe1** and **shootSe2**

Each of these profiles provides the expression value of $|G| = 22\,810$ genes.

The data were preprocessed as described in Section 2.1 to obtain two different contexts:

[3] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9311

- *normalised in the mean of the normal root profiles* $\mathbb{K}_r = (G, M_r, R_r)_{\overline{\mathbb{R}}}$. Thus all the gene expression values, for the gene i, will be normalized by:

$$c_i^r = \sqrt{C_{i\mathbf{root1}} \cdot C_{i\mathbf{root2}}} \tag{4}$$

  The final gene expression will be:

$$R_{ij} = \log \frac{C_{ij}}{c_i^r} \tag{5}$$

  Where $j \in \{\mathbf{shoot1}, \mathbf{shoot2}, \mathbf{rootSe1}, \mathbf{rootSe2}, \mathbf{shootSe1}, \mathbf{shootSe2}\}$ is one of the remaining 6 different profiles after removing **root1** and **root2**.
- *normalised in the mean of the normal shoot profiles* $\mathbb{K}_s = (G, M_s, R_s)_{\overline{\mathbb{R}}}$. As before the gene expression values, for the gene i, are normalized by:

$$c_i^s = \sqrt{C_{i\mathbf{shoot1}} \cdot C_{i\mathbf{shoot2}}} \tag{6}$$

  The final gene expression will be:

$$R_{ij} = \log \frac{C_{ij}}{c_i^s} \tag{7}$$

  Where $j \in \{\mathbf{root1}, \mathbf{root2}, \mathbf{rootSe1}, \mathbf{rootSe2}, \mathbf{shootSe1}, \mathbf{shootSe2}\}$ is one of the 6 different profiles remaining after removing **shoot1** and **shoot2**.

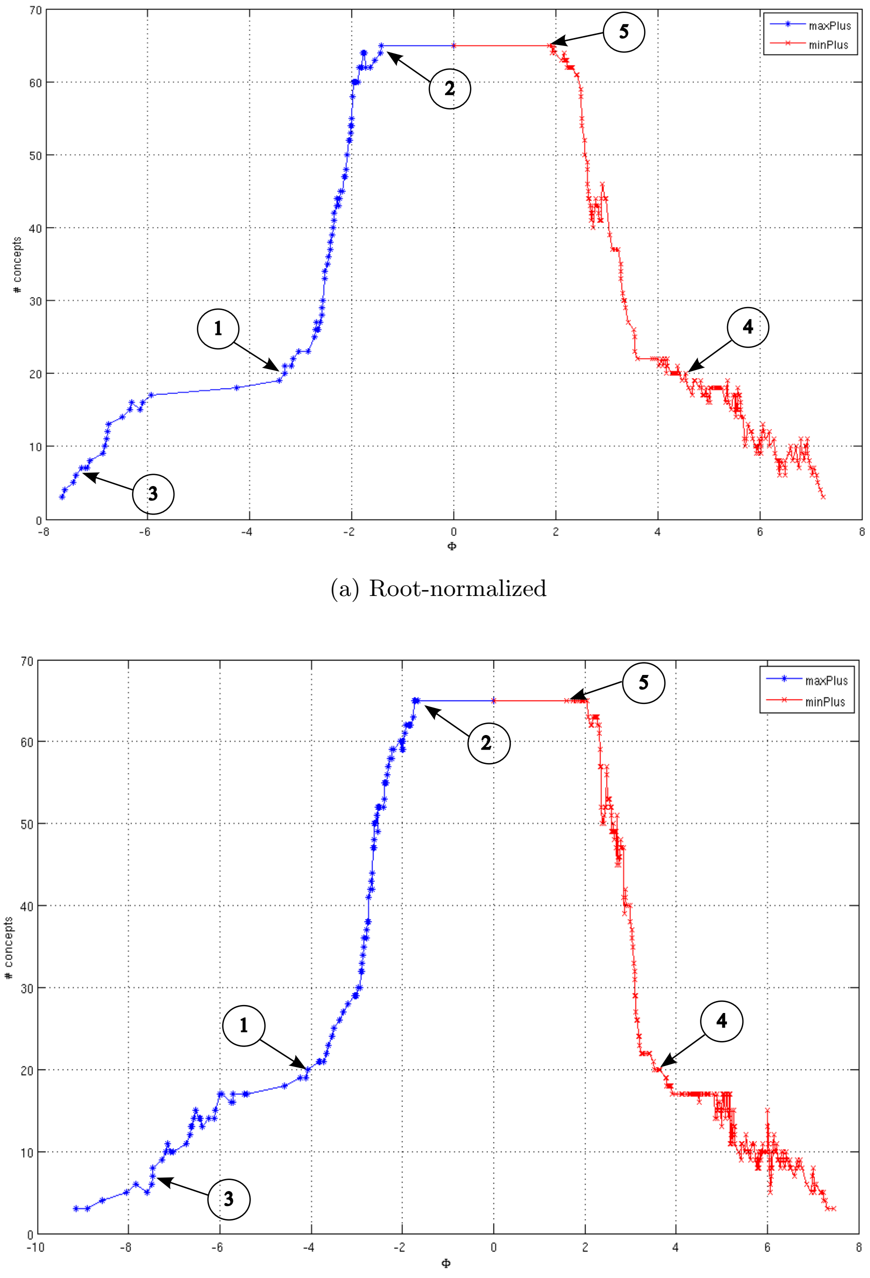
The idea is that each of the lattices explored for each of these contexts will shed light on the Selenium-modified analogue of the control, but the other conditions will further identify expression behaviour. As previously said the number of conditions for, say $\mathbb{K}_r$ is reduced to 6: the conditions used to find the control no longer appear, and the other six profiles are normalized by it. Therefore M_r and M_s are different albeit related.

The contexts were processed with our in-house $\mathcal{K}$-Formal Concept Analysis toolbox, running in MatLab.

3.2 Lattice exploration

Lattice exploration was carried out on each context using $\left(\overline{\mathbb{R}}_{\max,+}\right)$- and $\left(\overline{\mathbb{R}}_{\min,+}\right)$—to investigate under-expressed and over-expressed genes, respectively—for different values of the thresholds, with φ ranging in $(-\infty, 0)$ and ϕ in $(0, \infty)$, as described in Section 2.2. The resulting number of concepts are shown in Figure 1 for either context.

The overall shape of both curves is very similar. The left halves with $\varphi \in (-\infty, 0)$ start from two concepts when the threshold of existence is below the minimum entry in R, attaining the maximum 2^p in a neighbourhood of 0. On the other hand, the right halves with $\phi \in (0, \infty)$, are roughly symmetric collapsing again into a two-concept lattice when ϕ is above the maximum entry in R . It is worth mentioning, that it is possible to detect a change in the slope of the curves around $\varphi = -6$ and $\phi = 4$. This will be further looked into in the next subsections.



(a) Root-normalized

(b) Shoot-normalized

Fig. 1: (Colour on-line) Number of concepts as a function of the threshold level $\varphi \in (-\infty, 0)$ for down-regulated (blue dots) and $\phi \in (0, \infty)$ for up-regulated (red crosses) genes in root-normalized (a) and shoot-normalized (b) data. Points of interest to draw structural lattices from are the leftmost (an example labeled with arrow #3) and rightmost extremes, those points close to the plateaus, coming from either side (examples labeled with arrows #2 and #5), but specially the shoulders are each side of the "mesas" (examples labeled with arrows #1 and #4). The structural lattices for these examples are depicted in the subsequent figures.

Down-regulation analysis. To obtain an interpretation of the structure of the genes that are down-regulated in the presence of Selenium, structural lattices for negative φ should be explored. Figure 2 depicts two structural lattices at a middle value of the left part of the curve where the slope has been found to be lower—a *shoulder*. A clear separation between root-related and shoot-related conditions is appreciated in the form of adjoint sublattices in Figure 2a and almost adjoint sublattices in Figure 2b.

In Figure 2a, the four shoot-related conditions make up a boolean lattice with sets of genes labelled in all possible combinations of the four mentioned conditions. This implies that these conditions cannot be separated at this level in the root normalization. Interestingly, the **RootSe** conditions join at a node with a singleton extent, gene 259161*_at* related to carbon and nitrogen metabolism.

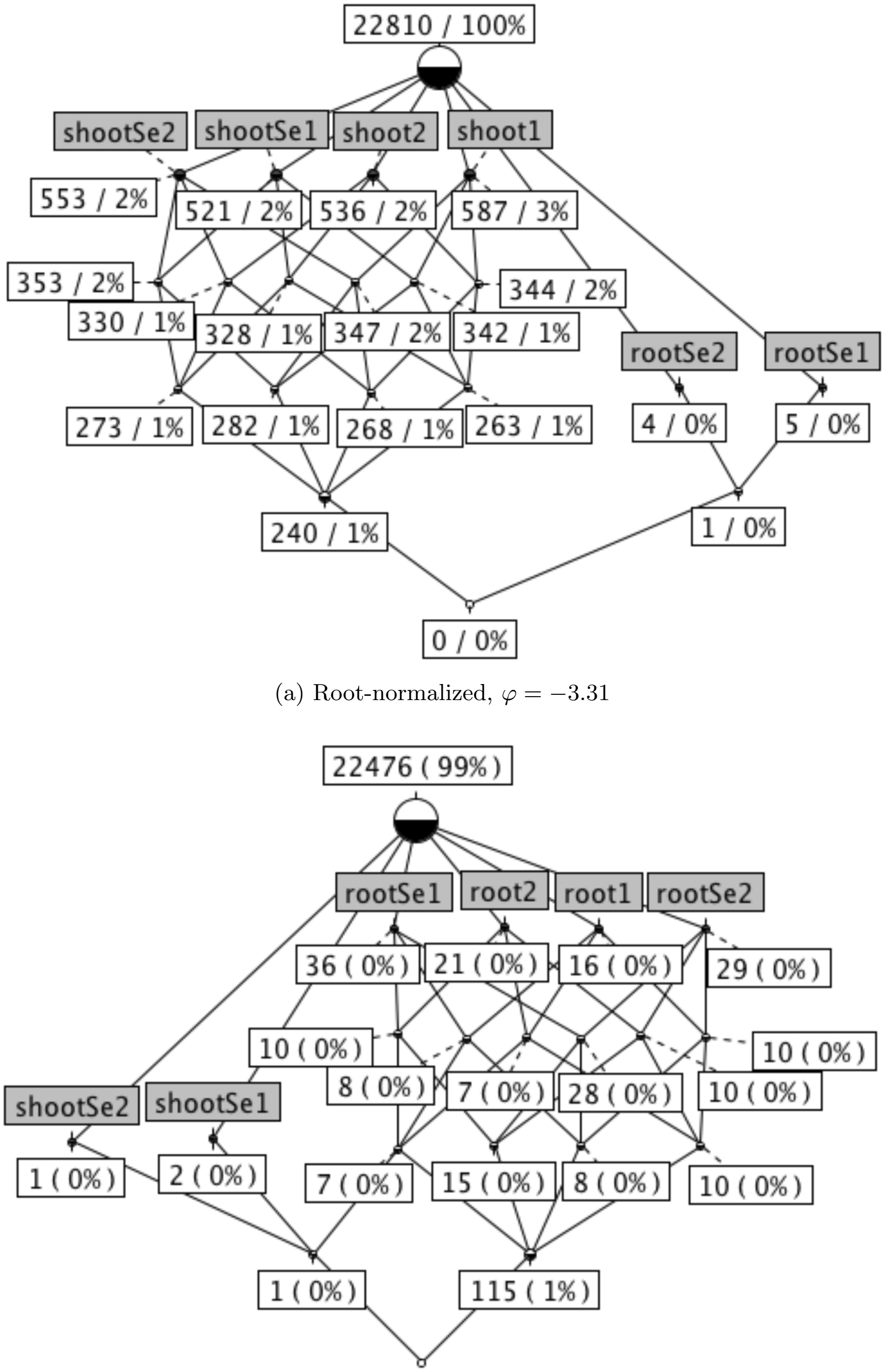
On the other hand, in Figure 2b a different situation can be noticed where the boolean sublattice is now generated by the four root-related conditions while the **ShootSe** conditions are apart. However, they are not so clearly differentiated due to the connection that exists with a lower node of the boolean sublattice. Interestingly, these conditions join at a node with a singleton extent, gene 251196*_at* or *glutaredoxin*, an enzyme normally related to stress signalling which is here inhibited.

Figure 3 depicts the projection of **rootSe** labels in Figure 3a and **shootSe** labels in Figure 3b from the full boolean lattice of 2^p concepts that appears close to $\varphi = 0$. The bottom nodes represent the 89 (118) genes that are down-regulated by Selenium in the root (the shoots), in which an agreement between both realizations of condition **rootSe** exists. Nonetheless, it is important to acknowledge that at this level of the observation the measurements are not very reliable due to the empirical limitations explained in 2.1, and we will concentrate on the findings for the previous case in Subsection 3.3.

Figure 4 presents, finally, the most salient down-regulated genes in a lattice for a low φ. As can be noted, for the root normalization (resp. shoot) the threshold of existence for **rootSe** (resp. **shootSe**) is too low to allow any gene down-regulated by that condition to appear. However, an incipient structure concerning shoot-related (resp. root-related) conditions is beginning to be discernible which we refuse to analyse in this first attempt.

Up-regulation analysis. Changing the choice of semiring from $\overline{\mathbb{R}}_{\max,+}$ to $\overline{\mathbb{R}}_{\min,+}$ allows us to analyse up-regulation. For this case, structural lattices for positive ϕ should be explored.

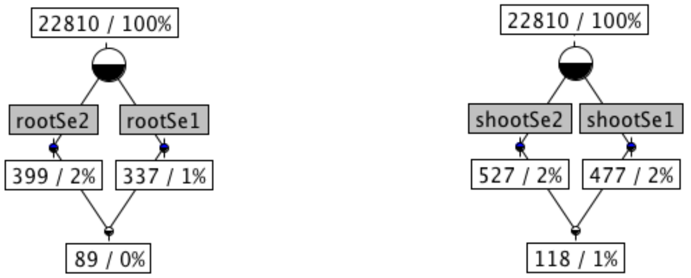
Figure 5 depicts two structural lattices at both middle values to the right of the curves in Figure 1 where the slopes have been found to be less decreasing. A clear separation between root-related conditions is again evident in the form of adjoint sublattices in Figure 5a. The same cannot be asserted for the shoot-related conditions in Figure 5b as it is not possible to find any structural lattice in which **shootSe1** and **shootSe2** are joined in an independent (not labelled with any other condition) concept different than bottom.



(a) Root-normalized, $\varphi = -3.31$

(b) Shoot-normalized, $\varphi = -4.06$

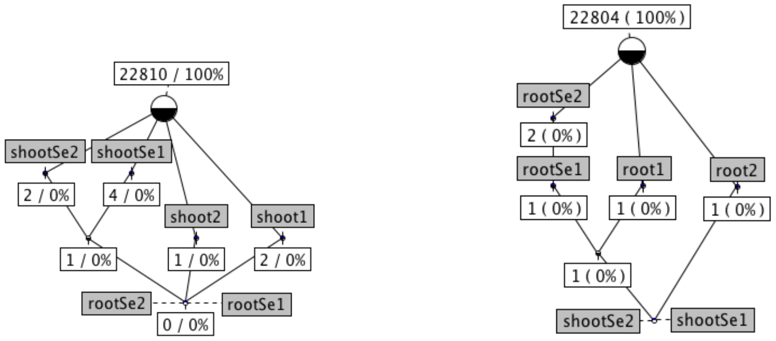
Fig. 2: Structural lattices for down-regulation analysis at a low φ where the **RootSe** (a) and **ShootSe** (b) conditions are apart from the rest. These lattices correspond to the points signaled with arrows #1 in both plots of figure 1.



(a) Root-normalized, $\varphi = -1.42$

(b) Shoot-normalized, $\varphi = -1.66$

Fig. 3: Structural lattices for down-regulation analysis at a high φ. Only rootSe (a) and shootSe (b) related conditions are retained. A fully connected boolean lattice is obtained at this level involving all conditions what indicates that the absolute value of φ is too low to consider down-regulation precise. These lattices correspond to the points signaled with arrows #2 in both plots of figure 1.

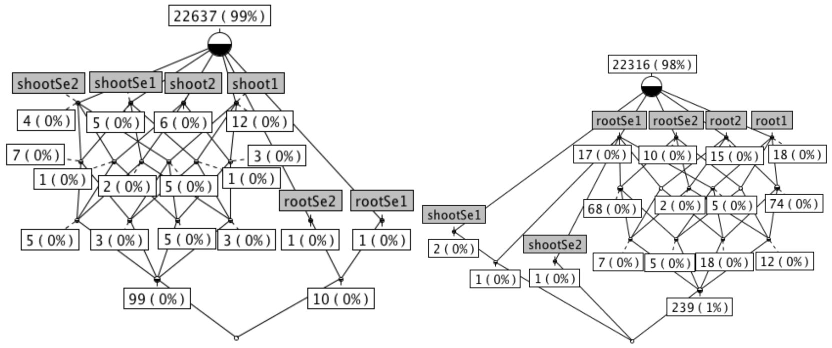


(a) Root-normalized, $\varphi = -7.29$

(b) Shoot-normalized, $\varphi = -7.47$

Fig. 4: Structural lattices for down-regulation analysis at a low φ. These lattices correspond to the points signaled with arrows #3 in both plots of figure 1.

The structure encountered in Figure 5a is analogue to the one in Figure 2a with the four shoot-related conditions conforming a boolean lattice (to the left) and an adjoint sublattice condensing root-related conditions (to the right). The



(a) Root-normalized, $\phi = 4.41$

(b) Shoot-normalized, $\phi = 3.65$

Fig. 5: Structural lattices for up-regulation analysis at a ϕ where the **RootSe** (a) and **ShootSe** (b) conditions separate from the rest of the conditions. These lattices correspond to the points signaled with arrows #4 in both plots of figure 1.

object counts of the concepts are different, however, involving considerably fewer genes in the boolean lattice and many more in the **root** sublattice. As in down-regulation both **rootSe** conditions join at a node that in this case contains 10 exclusive genes whose analysis can be found in Section 3.3.

Unfortunately, and thought almost the inverse situation can be noticed in Figure 5b, where the boolean sublattice is now generated by the four root-related conditions, the **shootSe** conditions do not appear totally apart not even joining at a common concept different than bottom. This divergence between the two realizations of the experiments prevents us from providing findings in this situation.

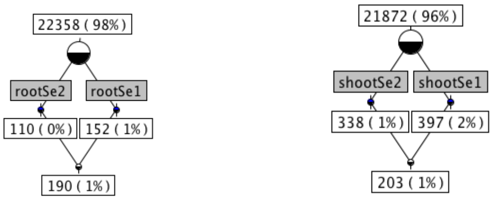
Figure 6 depicts the projection of **rootSe** labels in Figure 6a and **shootSe** labels in Figure 6b from the full boolean lattice of 2^p concepts that appears close to $\phi = 0$. The bottom nodes represent the 190 (203) genes that are up-regulated by Selenium in the root (shoots, respectively) in which an agreement between both realizations of the experiment exists.

Finally, similar lattices as the ones depicted for down-regulation in Figure 4 for low φ can be obtained for high ϕ and up-regulation. However, they are omitted here as they do not add much information for the present analysis.

3.3 Findings

We used the gene identifiers appearing in the more reliable concepts, those with lowest down-regulation and highest up-regulation threshold, to obtain their functional description, when available, from a knowledge database.

Preliminary analyses suggest that for up-regulation in the roots subject to *Se*, our procedure detects over-expressed genes used by *A. thaliana* to sense and signal physiological conditions (Ca^{++} transport), to bind to heavy metals (*Cd*,



(a) Root-normalized, $\phi = 1.86$ (b) Shoot-normalized, $\phi = 1.59$

Fig. 6: Structural lattices for down-regulation analysis at a high ϕ. Only rootSe (a) and shootaSe (b) conditions are retained. A fully connected boolean lattice is obtained at this level involving all conditions. This means that the absolute value of ϕ is too low to consider up-regulation significant. These lattices correspond to the points signaled with arrows #5 in both plots of figure 1.

Zn) and salts (*Se* is introduced as a selenate) and to combat metal-, pathogen- and salt-induced stress. It is encouraging that one of these genes has an unknown function but is suggested by our procedure to engage in some or all of these functions.

The results for down-regulation are less clear. On the one hand, less genes are clearly under-expressed: for the roots the single reliably detected gene engages in the metabolism of carbon by non-photosynthetic means and in that of nitrogen. For the shoots, the clearly inhibited gene, glutaredoxin, is an enzyme involved in signalling stress conditions employing sulphur-redox pairs (cysteine). The overall picture is not clear but might suggest that *Se* is interfering with the sensing of *S* in the plant, pretending that sulphur is over-abundant and thereby affecting the signalling related to it.

Further in-depth analysis should be carried out by plant physiologists.

3.4 Summary

The analysis carried out in the previous subsections allows us to reach the following conclusions:

- **root** and **shoot** conditions appear clearly apart in terms of the genes up- or down-regulated in each case. The disparity in the values of C_{ij} observed in them advise a separate analysis which we have implemented by providing two types of normalizations as described in section 3.1.
- Up- and down-regulation can be analysed with the same procedure by changing the carrier semiring in $\mathcal{K}$-Formal Concept Analysis from $\overline{\mathbb{R}}_{\max,+}$ to $\overline{\mathbb{R}}_{\min,+}$. The evolution of the number of concepts in each case proceeds inversely as can be observed in the overall symmetry of Figure 1.

- A consistency of both realizations of the same condition, e.g. **rootSe1** and **rootSe2**, should be always enforced to provide reliability.
- When our focus of attention is the up- or down-regulation in **rootSe** (resp. **shootSe**) conditions, the presence of shoot-related (resp. root-related) ones obscures the analysis, as they appear for very low values of φ (resp. very high values of ϕ), that is, either extreme of the curves in Figure 1).
- Around the values of $\varphi = 0$ and $\phi = 0$ a full boolean lattice of 2^p concepts appears showing the unreliability of these measurements due to empirical limitations of the measuring technique.
- Finally, a compromise between the two previous situations can be found in the middle of both down and up regulation analysis where figure 1 exhibits a decay of the absolute value of its slope—the *shoulders* of Figure 1. At these positions, **root** and **shoot** conditions separate into two adjoint sublattices for root normalization and a not-so-clear separation for shoot-normalised experiments.
- Pending more thorough analyses, the merely lattice theory-induced findings can be corroborated by gene-function analysis of the extents found for each case.

4 Discussion

In this paper we have introduced a new approach to gene expression data analysis with $\mathcal{K}$-Formal Concept Analysis, a flavour of Formal Concept Analysis where incidences may take values in complete idempotent semifields. Specifically, we directly analyse the $\overline{\underline{\mathbb{R}}}$-valued, non-scaled context of gene expression values by means of $\left(\overline{\mathbb{R}}_{\min,+}\right)$- and $(\mathbb{R}_{\max,+})$-Formal Concept Analysis.

Our analyses show that a combination of these is a promising tool for the interactive exploration of gene co-regulation, since exploring the context with $(\mathbb{R}_{\max,+})$-Formal Concept Analysis captures the phenomenon of gene down-regulation, while using $(\mathbb{R}_{\min,+})$-Formal Concept Analysis for the exploration captures up-regulation, decreases and increases, respectively, of gene concentrations with respect to a normalizing gene expression profile. In this way, we have detected genes that are either up-regulated or down-regulated in specimens subject to a Selenium-induced physiological stress.

Previous work on using Formal Concept Analysis for transcriptomics includes a remarkable proposal for a methodology for gene expression data exploration in [5], which seems to be the schedule adopted by most practitioners. Pensa et al. suggest and iterative process of exploration based in the *inductive databases* paradigm: for each iteration loop against a database of gene expression data, they carry out pre-processing, data discretisation (attribute scaling), Boolean gene expression data enrichment, Constraint-based extraction of Formal Concepts and post-processing.

Note that our methodology shares the first and last steps, but greatly changes the intermediate steps since no scaling or enrichment is used. Of course, this preliminary work has only demonstrated a single loop of the exploration procedure.

For instance, Motameny et al. [6] concentrated on a binary classification task over human leukaemia. They scaled gene expression values into binary attributes and used standard extents to obtain gene sets inducing rules for classification. In related work, [7] uses *interval scaling* aided by experts to discretise expression values.

Scaling is widely acknowledged to introduce biases in the analysis and perhaps to result in loss of context information [8]. Thresholding and insensibility parameters [9] have been used to minimize these effects, but also richer, hopefully loss-free, kinds of scaling such as *interordinal scaling* [10].

With regard to noise preprocessing, since normalization by means of control conditions does not dispose of noise, practitioners either refuse to believe data too firmly—as in our work—or do a flavour noise-insensitive analysis [10,9].

Regarding the latter, Pattern Formal Concept Analysis was designed to minimize or dispose of the need for scaling [11] . The novelty in [8,10] is considering expression value intervals as pattern structures to act as “attributes” in the context. The process of lattice building accords narrow intervals to concepts lower in the lattice and wider intervals to those higher up. The wider the interval, the less reliable is the concept association between genes and conditions. This seems to be a complementary approach to our analysis based in the threshold of existence for concepts, but it has not been applied to the complementary process of gleaning up- and down-regulated genes.

Regarding the phenomena being explored, most of the work so far seems to have concentrated in over-expressed genes or up-regulation, whereas our framework also caters for down-regulation, albeit with a technique complementary to that used for up-regulation, that is $\left(\overline{\mathbb{R}}_{\min,+}\right)$- vs. $\left(\overline{\mathbb{R}}_{\max,+}\right)$-Formal Concept Analysis.

In future work we plan to attack control vs. case studies in *A. thaliana*, as well as using the different types of Galois connections of Extended Formal Concept Analysis [3] on gene expression data to widen the tools at the practitioner's disposal.

References

1. Valverde-Albacete, F.J., Peláez-Moreno, C.: Towards a generalisation of Formal Concept Analysis for data mining purposes. In: Concept Lattices. Proceedings of the International Conference on Formal Concept Analysis (ICFCA 06). Volume 3874 of LNAI. (Dec 2006) 161—176
2. Valverde-Albacete, F.J., Peláez-Moreno, C.: Further Galois connections between semimodules over idempotent semirings. In Diatta, J., Eklund, P., eds.: Proceedings of the 4th Conference on Concept Lattices and Applications (CLA 07), Montpellier (October 2007) 199–212
3. Valverde-Albacete, F.J., Peláez-Moreno, C.: Extending conceptualisation modes for generalised Formal Concept Analysis. Information Sciences (in press)
4. Stoughton, R.: Applications of DNA microarrays in biology. Biochemistry **74**(1) (2005) 53

5. Pensa, R., Besson, J., Boulicaut, J.: A methodology for biologically relevant pattern discovery from gene expression data. In Suzuki, E., Arikawa, S., eds.: Discovery Science. Volume 3245 of LNAI., Springer (2004) 230—241
6. Motameny, S., Versmold, B., Schmutzler, R.: Formal Concept Analysis for the identification of combinatorial biomarkers in breast cancer. Lecture Notes in Artificial Intelligence **4933** (Jan 2008) 229—240
7. Gebert, J., Motameny, S., Faigle, U., Forst, C., Schrader, R.: Identifying genes of gene regulatory networks using Formal Concept Analysis. Journal of Computational Biology **15**(2) (Jan 2008) 185—194
8. Kaytoue, M., Duplessis, S., Kuznetsov, S.O., Napoli, A.: Two FCA-based methods for mining gene expression data. Formal Concept Analysis (Jan 2009) 251—266
9. Pensa, R., Boulicaut, J.: Towards fault-tolerant Formal Concept Analysis. In: AI* IA 2005: Advances in Artificial Intelligence. Volume 3675 of LNAI., Springer (2005) 212–223
10. Kaytoue, M., Kuznetsov, S., Napoli, A., Duplessis, S.: Mining gene expression data with pattern structures in Formal Concept Analysis. Information Sciences (In Press.) (2011)
11. Ganter, B., Kuznetsov, S.: Pattern structures and their projections. In: Conceptual Structures: Broadening the Base. Volume 2120 of LNCS. Springer (2001) 129–142