WebGeneKFCA: an On-line Conceptual Analysis Tool for Genomic Expression Data

José María Fernández-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
jmgc,carmen,fva@tsc.uc3m.es

Abstract. In this paper we introduce a Web-based tool for the analysis of Genomic Expression (GE) data based in $\mathcal{K}$-Formal Concept Analysis (KFCA). First we present the task of analysing GE data and then we describe the tool implementing KFCA. As a second contribution, we present a mechanism to visualise a sequence of concept lattices by fixing the intents against the concept lattice of the contranominal scale of attributes $\mathfrak{B}(M, M, \neq)$. Derived from this we also propose a mechanism to explore the scope of objects in such sequences.

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.

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 (GE) data* taken to analyse their functional influence consists of the expression value of every gene R_{ij}— an expression profile—under one condition m_j in a non-explicitly given set of conditions $M = \{m_j\}_{j=1}^p$. Under these premises, *co-regulation* refers to the increment (*up-regulation*) or decrement (*down-regulation*) of the expression value in a set of genes brought about by the change in expression value of other genes.

GE data exploration using Formal Concept Analysis includes the seminal work for of [1]. Later contributions essentially adhere to this paradigm, including our own [2] where we employ $\mathcal{K}$-Formal Concept Analysis (KFCA) [3].

In this paper we introduce WebGeneKFCA, an application supporting Exploratory Analysis of GE data using KFCA that attempts to embody the iterative process of exploration based on *inductive databases* suggested by Pensa et

* Corresponding Author.

© Laszlo Szathmary, Uta Priss (Eds.): CLA 2012, pp. 345–350, 2012.
ISBN 978–84–695–5252–0, Universidad de Málaga (Dept. Matemática Aplicada)

al. [1]. Beyond the proof-of-concept nature of the work described in [2] WebGeneKFCA insists on providing tools for the practitioner to contextualise with domain knowledge as embodied in Gene Ontologies (GOs): a point-and-click interface seamlessly integrated with lattice visualisation enables the exploration of many different contextualised hypotheses.

Data procurement and normalisation is described in Sec. 2.1. Our extension of the state of the art in concept lattice (CL) visualisation that provides a representation for *sequences* of these is described in Sec. 2.2.

Relying on this property, we present yet another new visualisation feature whereby the related concepts in different CL of the sequence can be aggregated and their scope in terms of the value of a single continuous parameter explored (Sec. 2.3). The paper closes with a summary of contributions and further work.

2 Exploratory Analysis with WebGeneKFCA

In this Section we describe the Exploratory Analysis of Gene Expression Data using the inductive databases paradigm as embodied in WebGeneKFCA [1].

2.1 Data procurement and normalization

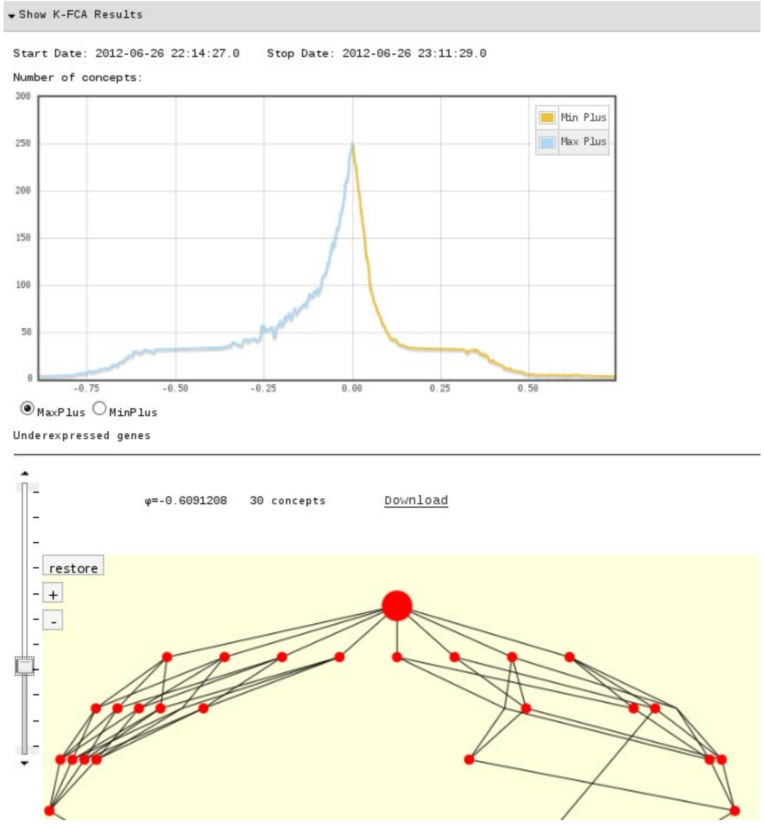
Before any new data can be analysed it must be uploaded to the platform in form of Affymetrix v4 CEL files [4]. Each experiment to analyse will consist of several tests each corresponding to a CEL file. Then, the `apt-summary-tool` is executed which is an open source tool provided by Affymetrix that creates matrix R_{ij} where each column represents a condition profile obtained from a CEL file and each row is a gene profile.

Prior to data analysis, the user must choose how to normalise the data to make it suitable for $\mathcal{K}$-Formal Concept Analysis. Currently we support four different types of normalisation schemes that make use of a special kind of profile called *control*: no normalisation, by the arithmetic mean, the geometric mean or the maximum value of the control profile.

2.2 Lattice Exploration and visualisation

As explained in [2], lattice exploration consists in sweeping all possible values of the data matrix R' in the $\overline{\mathbb{R}}_{\max,+}$ and $\overline{\mathbb{R}}_{\min,+}$ domains. The result of these two different exploring strategies can be shown together as in Fig. 1. The Structural Context that gives rise to the CL shown can also be obtained by clicking on the "Download" link just on the top of the graph in a standard CSV format.

[1] The server was built using Java 1.6 and the Spring framework (v. 3.1), and runs on Tomcat 6. The web view makes extensive use of *javascript* and *html5* and has been optimised for Chrome web browsers. To store the data the application currently uses *mysql* but it can be easily ported to any other SQL database. At least 1.5GB of free RAM is required. The tool is currently only available for in-house usage.



Start Date: 2012-06-26 22:14:27.0 Stop Date: 2012-06-26 23:11:29.0

Number of concepts:

● MaxPlus ○ MinPlus

Underexpressed genes

φ=-0.6091208 30 concepts

Fig. 1: (Colour online) First data exploration screen. The top pane shows the number of concepts vs. φ (light blue, to the left of 0.0) and ϕ (drab green, right of 0.0) for the context being explored. The bottom pane shows the φ slider and (part of) the CL thus selected. In-between, the toggle for $\overline{\mathbb{R}}_{\max,+}$ or $\overline{\mathbb{R}}_{\min,+}$.

Several algorithms exist for the visualisation of CL, each with its advantages and disadvantages (see [5] for a review). However, since the use of $\mathcal{K}$-Formal Concept Analysis requires the visualisation of a *sequence* of CL, we propose a scheme having the distinctive feature that the Formal Concepts with the same intent belonging to different CL are always plotted in the same position. This means that the user can change the value of ϕ (φ) and will easily see how the extent of each concept evolves, increasing or decreasing until disappearing.

To ensure this property, the CL corresponding to a particular ϕ (φ) is drawn *over* the silhouette of the CL of a (virtual) contranominal scale involving all possible attributes, $\mathbb{N}^c_M = \underline{\mathfrak{B}}(M, M, \neq)$.

The rationale for this overlay is as follows: in the boolean lattice $\underline{\mathfrak{B}}(M, M, \neq)$, the top is a concept with no attributes. The next level of concepts from the top are those which only have one attribute, and so on. Call $|M| = p$, the number of conditions. Clearly, the total number of levels is the number of conditions plus one, and the number of concepts in each level l is $\binom{p}{l}$, whence $\underline{\mathfrak{B}}(M, M, \neq)$ has the maximum possible number of Formal Concepts for a given set of attributes M, a well-known result (see [6], p.48).

The important fact is that *any possible intent in any lattice with p attributes appears in this virtual order diagram.* Thus the locations of these Formal Concepts of the most complex CL related to M can serve as locations for those Formal Concepts of *any other CL with the same attribute set M.* Figure 2.(a) shows an example of such an overlay of a CL with relatively few concepts, while another example is shown in Fig. 2.(b) of a CL whose set of intents approaches that of the boolean lattice (despite conspicuous absences, e.g. in the atom set).

In this visualisation scheme, the size of each concept node is directly proportional to its extent size (but see 2.3). Making the mouse hover over each node, a dialog box showing the attributes of that concept and its number of objects appears. On clicking on the node, a floating window appears showing the full extent. Besides, since these objects are genes, each of them can be selected causing its associated information, obtained from the corresponding NetAffx Annotation file [7], to be displayed in a pane on the right side. This information pane also links with other external web pages that offer more information.

2.3 The Exploration of an Object Scope

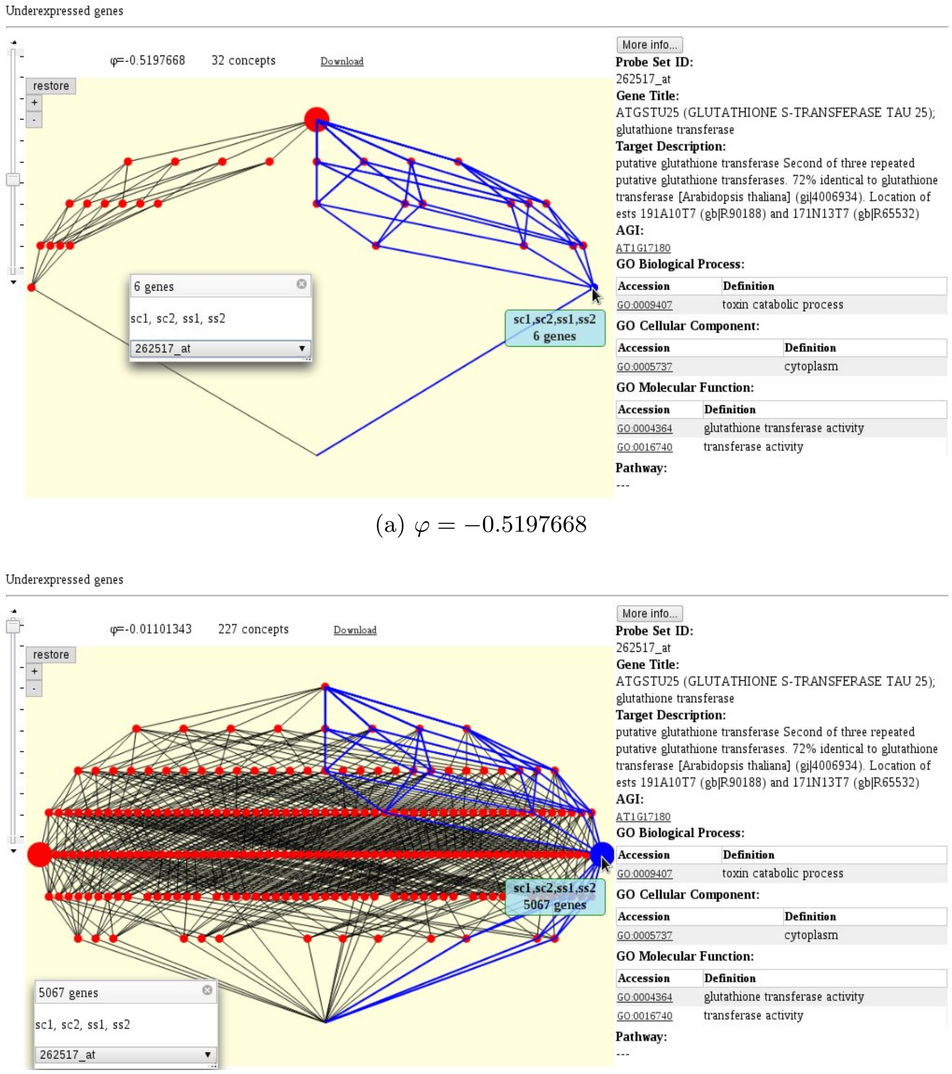
An additional property of each object (gene) can be observed when clicking the button with the legend "More info...". The new view that comes out (Fig. 3) displays the set of concepts the gene appears in through the CL full sequence. We call this the *scope* of the object (gene).

Against the backdrop of the boolean lattice, every concept appearing in any of the CL is rendered and a blue path connects all the concepts that the selected gene has ever belonged to. Hovering with the mouse over one of the blue dots from this path makes a tooltip appear showing the scope in φ (or ϕ) for the gene and that concept. The size of the dot is proportional to the width of the scope, the size of the φ interval. This information is also complemented with the data from the NetAffx Annotation file.

3 Contributions and Further Work

We have presented a Web-based tool to analyse GE data obtained from microarrays and to cluster the genes and test conditions by their similarity in either up- or down-regulation. The system sifts through many CL and saves its results in a database for later reading. The user can inspect through a visual, point-and-click interface the gene expression and related information in Gene Ontology-contextualised CL. Thanks to direct links, it is very easy to gather gene information from other Genomics sites.

Since the basic exploration mechanism is $\mathcal{K}$-Formal Concept Analysis, the user has a wealth of CL that make it difficult to propose and validate data-suggested hypotheses. To curb this exploration complexity we have proposed a



Underexpressed genes

φ=-0.5197668 32 concepts Download

More info...

Probe Set ID:
262517_at

Gene Title:
ATGSTU25 (GLUTATHIONE S-TRANSFERASE TAU 25); glutathione transferase

Target Description:
putative glutathione transferase Second of three repeated putative glutathione transferases. 72% identical to glutathione transferase [Arabidopsis thaliana] (gi|4006934). Location of ests 191A10T7 (gb|R90188) and 171N13T7 (gb|R65532)

AGI:
AT1G17180

GO Biological Process:

Accession	Definition
GO:0009407	toxin catabolic process

GO Cellular Component:

Accession	Definition
GO:0005737	cytoplasm

GO Molecular Function:

Accession	Definition
GO:0004364	glutathione transferase activity
GO:0016740	transferase activity

Pathway:

(a) $\varphi = -0.5197668$

Underexpressed genes

φ=-0.01101343 227 concepts Download

More info...

Probe Set ID:
262517_at

Gene Title:
ATGSTU25 (GLUTATHIONE S-TRANSFERASE TAU 25); glutathione transferase

Target Description:
putative glutathione transferase Second of three repeated putative glutathione transferases. 72% identical to glutathione transferase [Arabidopsis thaliana] (gi|4006934). Location of ests 191A10T7 (gb|R90188) and 171N13T7 (gb|R65532)

AGI:
AT1G17180

GO Biological Process:

Accession	Definition
GO:0009407	toxin catabolic process

GO Cellular Component:

Accession	Definition
GO:0005737	cytoplasm

GO Molecular Function:

Accession	Definition
GO:0004364	glutathione transferase activity
GO:0016740	transferase activity

Pathway:

(b) $\varphi = -0.01101343$

Fig. 2: (Colour online) CL and gene description view for two different values of φ. Note the similarity of shapes.

novel visualisation scheme which amounts to the visual embedding of CL in the representation of the most complex lattice pertaining to a set of attributes, viz. the contranominal scale of attributes. It is conceivable that this representation could cater to some other interval-valued parameters like intervals of temporal continua.

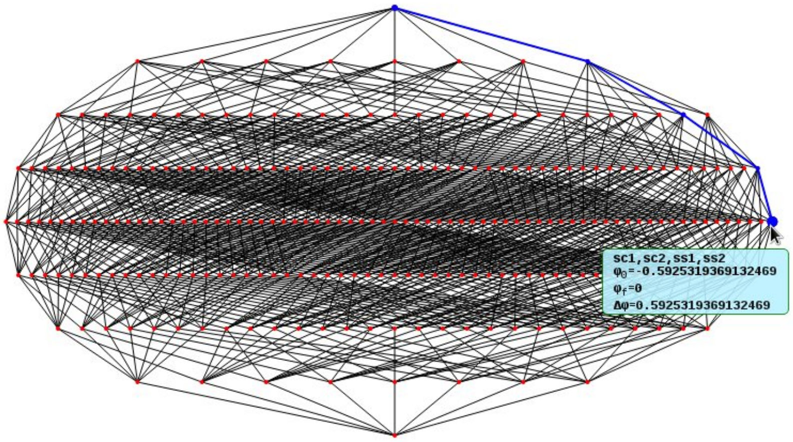

sc1,sc2,ss1,ss2
$\varphi_{\ominus}$=-0.5925319369132469
φ_{f}=0
$\Delta\varphi$=0.5925319369132469

Fig. 3: (Colour online) Gene evolution in a sequence of CLs

We have also proposed a mechanism to aggregate the visualisation mechanisms above that enables the study of the robustness of gene appearance in concepts vs. the variation of thresholds, the scope. We believe that gene scope is an important tool to ascertain the quality of the CL with respect to the “noise” in gene expression values and will explore it in further work.

References

1. 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
2. González-Calabozo, J., Peláez-Moreno, C., Valverde-Albacete, F.J.: Gene expression array exploration using $\mathcal{K}$-Formal Concept Analysis. In Valtchev, P., Jäschke, R., eds.: Proceedings of the ICFCA11. Volume 6628 of LNAI., Springer (2011) 119–134
3. Valverde-Albacete, F.J., Peláez-Moreno, C.: Extending conceptualisation modes for generalised Formal Concept Analysis. Information Sciences **181** (2011) 1888–1909
4. Affymetrix: Affymetrix CEL Data File Format (2009) http://www.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/cel.html.
5. Eklund, P., Villerd, J.: A survey of hybrid representations of concept lattices in conceptual knowledge processing. In: Formal Concept Analysis:8^{t}h International Conference, ICFCA 2010, Agadir, Morocco (2010) 296–311
6. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin, Heidelberg (1999)
7. Affymetrix: NetAffx Annotation file for *Arabidopsis thaliana* (2012) http://www.affymetrix.com/support/technical/byproduct.affx?product=arab.