Journal of Computational Biology and Bioinformatics Research Vol. 3(6), pp. 83-90, November 2011
Available online at http://www.academicjournals.org/JCBBR
ISSN-2141-2227 ©2011 Academic Journals

Full Length Research Paper

Querying formal concepts containing transcription factors: A case study using multiple databases

Mathilde Pellerin[1,2], and Olivier Gandrillon[1]*

[1]Université de Lyon, Université Lyon 1, Centre de Génétique et de Physiologie Moléculaire et Cellulaire (CGPHIMC), CNRS UMR5534, F-69622 Lyon, France.
[2]Statlife, Espace Maurice Tubiana, 39 rue Camille Desmoulins, 94805 VILLEJUIF, France.

Accepted 26 October 2011

In order to reduce the amount of information when querying from large databases, one has to develop new approaches. We present here a new way to query our SQUAT database. SQUAT contains formal concepts representing an association between a number of genes that are simultaneously overexpressed and the biological situations in which those genes are overexpressed. We explored the relevance of querying "self-explaining" formal concepts obeying a double constraint: (1) The concept should contain, within the genes of the concepts, at least one transcription factor (TF), and (2) At least one gene in the concept, should contain in its promoter a transcription factor binding site (TFBS) for the identified TF. The present work demonstrated that: (1) there are such "self-explaining" formal concepts in SQUAT. (2) Mining only those "self-explaining" formal concepts severely reduces the number of concepts that have to be analyzed. (3) Two such "self-explaining" concepts have been further analyzed, and their biological relevance has been demonstrated.

Key words: Data mining, gene expression, large database, formal concepts.

INTRODUCTION

The generation of very large gene expression databases by high-throughput technologies like microarray (Gershon, 2002), SAGE (Velculescu et al., 1995) or RNA-seq (Hanriot et al., 2008) calls for similarly high-throughput exploration tools of the possible functional links between gene expression levels and biological situations. Various techniques have been used for exploring such relationships, including global techniques like hierarchical clustering (Ng, 2001) or local techniques like local pattern extraction (Prelic et al., 2006). For the biologist, a local pattern is an association between a number of genes displaying specific expression properties and the situations in which those genes display such properties. A recent review highlights the relevance of mining local patterns with respect to clustering analyses (Madeira and Oliveira, 2004).

We have been developing local pattern extraction such as association rule discovery (Becquet et al., 2002; Creighton and Hanash, 2003; Georgii et al., 2005; Li et at., 2003) or formal concepts (Rioult et al., 2003; Blachon et al., 2007) to capture groups of genes displaying a simultaneous behavior in a number of biological situations. We have been focusing on the gene overexpression property (for a discussion about overexpression, see Becquet et al., 2002 and Pensa et al., 2004). A formal concept is a special case of a local pattern that harbors an association between genes that are simultaneously overexpressed and the biological situations in which those genes are overexpressed. We have recently described a web-available database called SQUAT containing different types of data, including raw SAGE expression values and local patterns in the form of formal concepts (Leyritz et al., 2008) allowing the biologist to query the resulting information.

One of the main drawbacks of every local pattern approach is the huge number of extracted patterns. This is especially true in noisy data, such as transcriptomic data. As an example, the human part of SQUAT database contains 532,073 formal concepts, and the murine part contains 1,141,895 formal concepts. We have therefore developed over the years a number of techniques to reduce the amount of information to be displayed to the final end-user, that is, the biologist.

This includes:

(1) a simple color-coding approach by function

*Corresponding author. E-mail: olivier.gandrillon@univ-lyon1.fr.

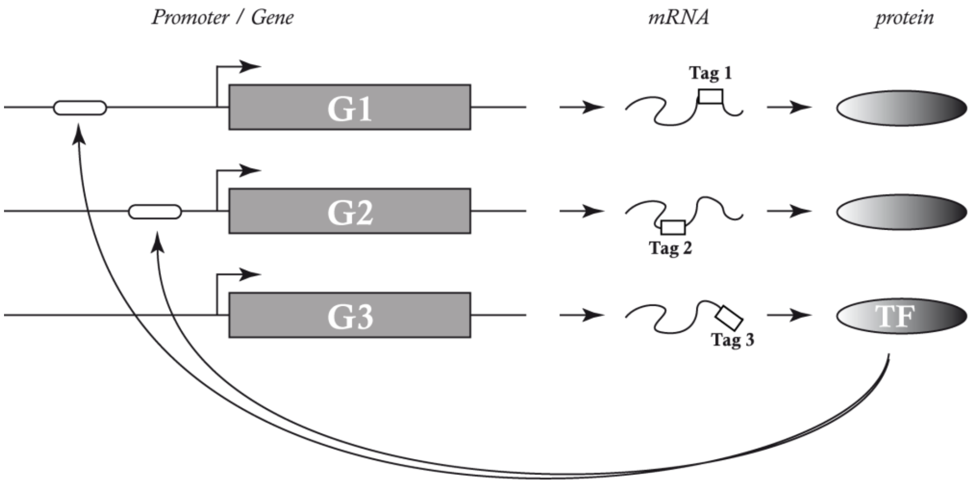


Figure 1. Schematic description of the “self-explaining” concept. SQUAT was used to establish a “tag-to-TFBS” relationship. This allowed to query for concepts containing autoregulated genes (genes harboring in their promoter at least one TFBS for the TF that is present within the concept). The percentage of autoregulated genes is then calculated (in the example it is 66.6%).

(Becquet et al., 2002).

(2) a regrouping of formal concepts using clustering techniques (Blachon et al., 2007) implemented in SQUAT (Leyritz et al., 2008).

(3) the simultaneous use of various sources of information, including text mining approaches (Klema et al., 2008).

In the present work, we decided to focus on “self-explaining” formal concepts. For this the basic idea was to find concepts responding to the following query: “find all concepts, containing within the genes of the concepts, at least one transcription factor (TF), and in which at least one gene in the concept contains in its promoter a transcription factor binding site (TFBS) for the identified TF.”

Our hope was that this TF/TFBS relationship should be able to explain at least part of the molecular link explaining why some of those genes were found in the same concept, which is why those genes are simultaneously overexpressed.

METHODS

The SQUAT database was used for performing the tag-to-transcript relationship (Keime et al., 2004), followed by the tag-to-transcript-to-TSS relationship (Leyritz et al., 2008). Starting from TSS positions, promoter sequences were defined as ranging from 5 kbp in 5’ of the TSS to 1 kbp in 3’ of the TSS. All promoters corresponding to a 1 tag – 1 transcript – 1 TSS were kept, as well as promoters corresponding to a 1 tag – 1 transcript – n TSS, if all TSS were contained within a 2 kpb distance. In the first case, the most 3’ TSS was used for further studies. This left us with a total of 12,951 human promoters.

The MATCH program (Kel et al., 2003, 2008 version 4) was run using the resources of the commercial version of TRANSFAC® for finding all TFBS on these promoter sequences. In order to reduce the number of false positives, the profile contained in the “vertebrate_non_redundant_minFP.prf” file was used.

In the end, we obtain a tag-to-TFBS relationship (Figure 1) that is the basis for future queries. The query is a two step process. All concepts containing at least one TF were isolated from SQUAT. The percentage of autoregulated genes was then calculated for each of the concepts.

The L2L-based queries were performed using the stand alone version of L2L (http://depts.washington.edu/l2l/; Newman and Weiner, 2005). This tool, given a gene list, provides categories that are statistically overrepresented as compared to a gene random sampling. We therefore took as an entry a list of genes, belonging to one concept, a well as lists belonging to the following categories:

1. The Gene Ontology organizing principle: biological process,
2. The Gene Ontology organizing principle: molecular function.
3. The L2L specific category: microarray data. For this L2L compare the list of genes contained within a concept to lists of genes that have been experimentally determined as being over expressed in response to a particular stimulus - in other words, published lists of microarray results.

The program first calculates the number of expected matches for that list, then the relative enrichment of actual matches, and finally a binomial probability for the relative enrichment. The results are logged, and written to a raw output file. The best p-values were retrieved and one therefore obtains, for each formal concept, three values: the best p value obtained when trying to find an enrichment regarding a biological process, a molecular function or a microarray experiment.

STRING was queried using the default parameters values.

RESULTS

The first purpose of this work was to find “self explaining concepts” obeying a double constraint regarding the presence of a TF in the concept and of potential target genes among the other genes of the concept (Figure 1). The second purpose was to see if that would lead to a

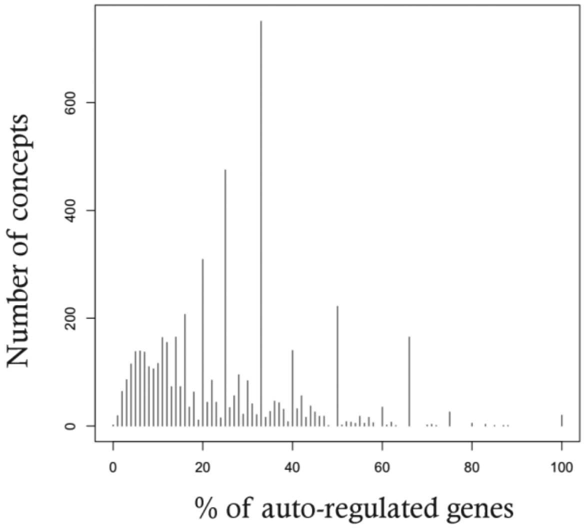


Figure 2. The number of concepts as a function of the autoregulated genes it harbors. 5013 concepts containing at least one autoregulated gene (mean size = 14.74 genes, as compared with the mean size of 5.43 for all SQUAT concepts) were obtained. 561 concepts harbored at least 50 % of autoregulated gene (mean size = 6,65 genes) and 20 concepts harbored 100 % of autoregulated gene (mean size = 3,15 genes).

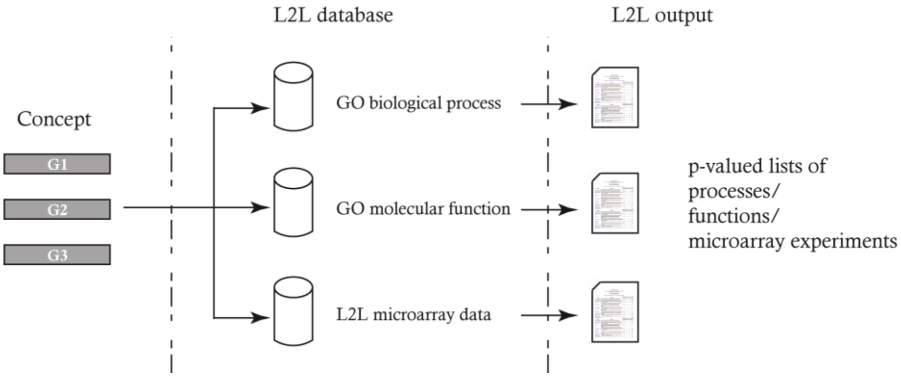


Figure 3. Schematic description of the use of L2L. Each individual concept can be seen as a list of genes. Each list was then compared to three types of lists of genes present within L2L. This results in the calculation of the p-value estimating the statistical significance of the redundancy between the two lists.

reduction in the number of formal concepts that have to be studied, and the third was to investigate their biological relevance.

We first checked for the presence of concepts that would fit such a double constraint. The results of the corresponding queries are displayed in Figure 2. Three things are readily apparent from Figure 2:

1. There indeed are concepts obeying the double constraint.
2. The percentage of auto-regulated gene can vary over the full range of 1 to 100% of the genes
3. Although the number of concepts is severely reduced from more than 500,000 to a few hundreds, it nevertheless still represents an unmanageable amount of information

We therefore decided to explore the possible biological relevance of the queried concepts, using L2L. For each identified concept, three files, representing three types of categories were retrieved from L2L (Figure 3):

1. Biological process,
2. Molecular function

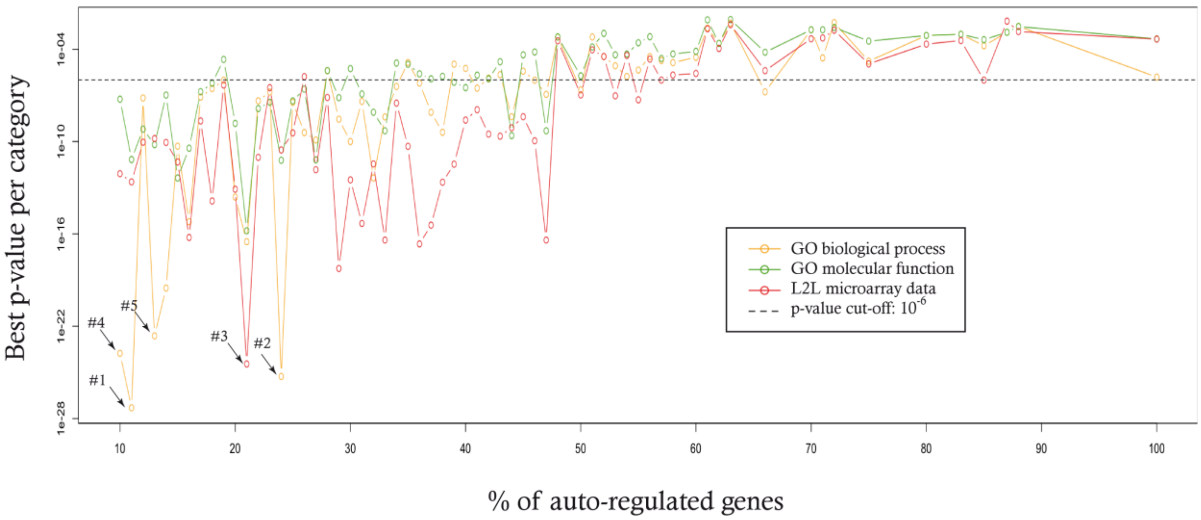


Figure 4. Most significant concepts for varying values of the percentage of auto-regulated genes.
Shown is the best p-value among all of the possible p-values for all of the gene belonging to all of the concept that were extracted at a certain value of x, the percentage of auto-regulated genes. The best p-value was selected each for of the three L2L categories examined. The five concepts displaying the lowest p-values are indicated and labeled 1 to 5.

Table 1. The four CRX concepts.

Rank	List Name	Binomial p-value	Number of libraries	Number of tags	% of CRX-regulated genes
1	visual perception	4.94 10 -28	2	168	11,00
2	visual perception	5.46 10-26	2	150	11,00
4	visual perception	1.73 10-24	3	84	10,00
5	visual perception	2.34 10-23	2	139	13,00

Column 1: The rank of the concept among the 5 best p-values (Figure 4); Column 2: L2L list name providing the best p-value; Column 3: the actual p-value; Columns 4 and 5: the number of libraries (i.e. biological situations) and tags (i.e. genes) in each concept; Columns 6: the percentage of auto-regulated genes (via CRX binding sites) in each concept.

3. Microarray data.

All concepts for values of x (the percentage of auto-regulated gene) ranging from 1 to 100% were extracted. Then for one value of x, one obtains a large amount of concepts. All of the genes belonging to those concepts were processed through L2L, and the best p-value obtained for a given value of x, for the three categories chosen, for any concept, was selected. The best p-value, for the three categories chosen, is displayed in Figure 4.

Although this might be counterintuitive, it is nevertheless clear that there are much more biologically significant concepts arising for the lower values of x. For concepts where more than half of the genes are auto-regulated, there is almost no concept displaying a significant p-value using L2L.

The five most significant extracted concepts were then examined. Among those, 4 harbored the same transcription factor, named CRX (Table 1). It is immediately apparent that the best p-values were all obtained for the “Visual perception” category of the GO biological processes.

We then interrogated SQUAT in order to estimate the global number of concepts which contained CRX. We found a total of 7 concepts, 4 of them containing more than one gene. It therefore appears that among the 5 best p-values we obtained all four CRX concepts among the 532073 concepts contained within our SQUA database. This demonstrates the power of the approach to extract a very small number of closely related concepts.

We then investigated the nature of the three libraries found within the concepts. It turned out that all three libraries were made from normal retina. At that stage we investigated the nature of CRX. Using the hyperlink from SQUAT to Entrez Gene, we could find the following description of the function of CRX: “The protein encoded by this gene is a photoreceptor-specific transcription factor which plays a role in the differentiation of photoreceptor cells. This homeodomain protein is necessary for the maintenance of normal cone and rod function”. It was therefore clear that we had extracted information regarding the overexpression for a

A

Gene product	TFBS	Description
RAX2	1	Homo sapiens retina and anterior neural fold homeobox 2
RDH8	1	Homo sapiens retinol dehydrogenase 8 (all-trans)
GUCA1A	1	Homo sapiens guanylate cyclase activator 1A (retina)
RLBP1	1	Homo sapiens retinaldehyde binding protein 1
PDE6G	2	Homo sapiens phosphodiesterase 6G, cGMP-specific, rod, gamma
RCVRN	1	Homo sapiens recoverin
PPEF2	1	Homo sapiens protein phosphatase, EF-hand calcium binding domain 2

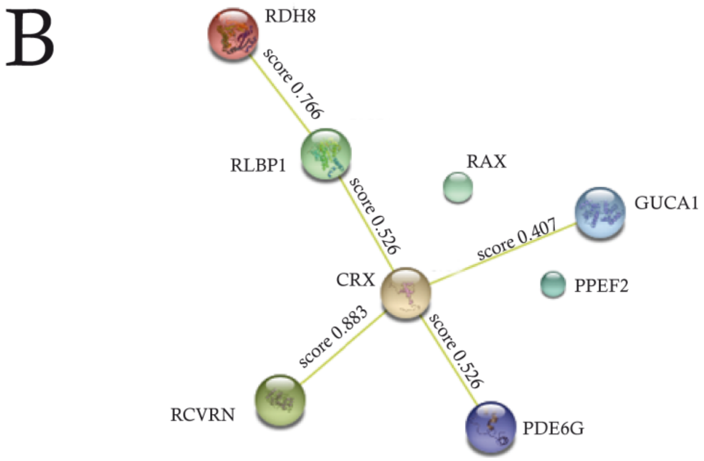
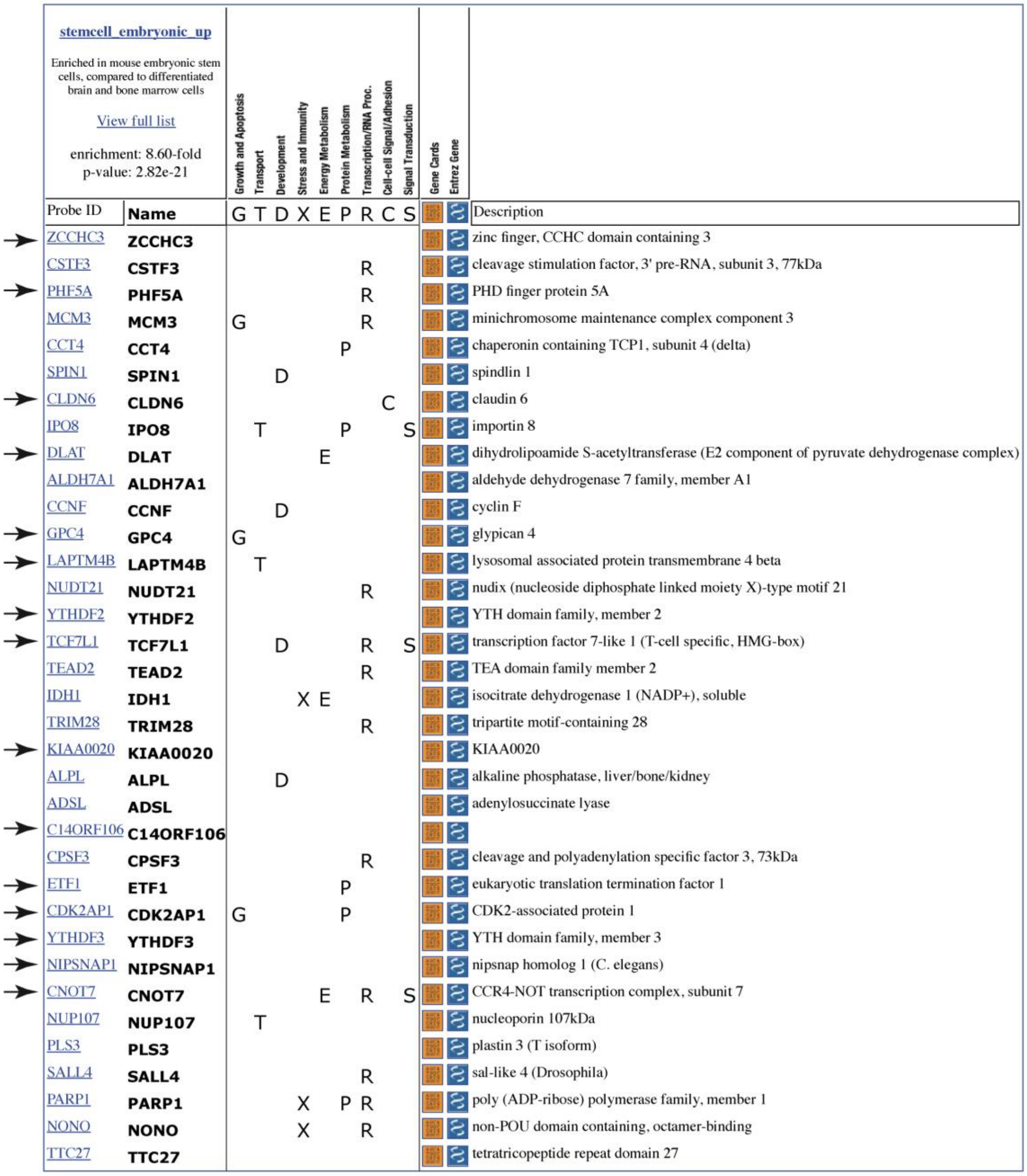


Figure 5. Analysis of the biological meaning of the CRX concepts. **A**: Shown are all the genes that 1. were common to all four concepts and 2. harbored at least one TFBS for CRX in their promoter. For each gene is shown its HUGO name, the number of CRX-binding sites in its promoter, and its full name. **B.** STRING output showing the known relationship between the 8 genes, indicated by a yellow line representing text-mining-based evidence. The combined score are computed as the joint probability of the probabilities from the different evidence channels, correcting for the probability of randomly observing an interaction (see the STRING website for more information).

photoreceptor-specific transcription factor in retinal cells Finally, we identified all genes that (1) were common to all four concepts and (2) harbored at least one TFBS for CRX in their promoter. This resulted in a list of 7 genes (Figure 5A). Those were still linked to "visual perception" with a very highly significant score (p= 2.48 10-12), which is due to the fact that all individual gene products could be shown to be related to eye development and vision (not shown).

In order to explore possible known relationship between CRX and any of those gens, we turned to the STRING database (http://string-db.org/; Figure 5B). Among the 7 genes, 2 (RAX and PPEF2) had no known relationship with CRX,1 (RDH8) had an indirect relationship, and the 4 left displayed weak text-mining-based relationship. When explored in details, the most relevant relationship was between Recovering and CRX, whereas only anecdotal co-occurrence-based linked CRX to RLBP-1, GUCA1 andPDE6G.

We therefore have isolated 7 new putative direct CRX target genes involved in the visual ability of retinal cells, previously uncharacterized as CRX target genes. As a next step in the analysis, we decided to relax the stringency of the p-value constraint. For this, we analyzed the concepts harboring the 54 best p-values. Among those it was immediately apparent that the TEAD2 (TEA domain family member 2) transcription factor was the most prominent one, since it appeared in 40% of the concepts. One should note that 466 concepts containing TEAD2 and with more than one gene and more than one situation appear in SQUAT. So this is different from the previous situation: here we selected a subpart of all of the



stemcell_embryonic_up

Enriched in mouse embryonic stem cells, compared to differentiated brain and bone marrow cells

View full list

enrichment: 8.60-fold
p-value: 2.82e-21

			Growth and Apoptosis	Transport	Development	Stress and Immunity	Energy Metabolism	Protein Metabolism	Transcription/RNA Proc.	Cell-cell Signal/Adhesion	Signal Transduction	Gene Cards	Entrez Gene	
	Probe ID	Name	G	T	D	X	E	P	R	C	S			Description
→	ZCCHC3	ZCCHC3												zinc finger, CCHC domain containing 3
	CSTF3	CSTF3							R					cleavage stimulation factor, 3' pre-RNA, subunit 3, 77kDa
→	PHF5A	PHF5A							R					PHD finger protein 5A
	MCM3	MCM3	G						R					minichromosome maintenance complex component 3
	CCT4	CCT4						P						chaperonin containing TCP1, subunit 4 (delta)
	SPIN1	SPIN1			D									spindlin 1
→	CLDN6	CLDN6								C				claudin 6
	IPO8	IPO8		T				P			S			importin 8
→	DLAT	DLAT					E							dihydrolipoamide S-acetyltransferase (E2 component of pyruvate dehydrogenase complex)
	ALDH7A1	ALDH7A1												aldehyde dehydrogenase 7 family, member A1
	CCNF	CCNF			D									cyclin F
→	GPC4	GPC4	G											glypican 4
→	LAPTM4B	LAPTM4B		T										lysosomal associated protein transmembrane 4 beta
	NUDT21	NUDT21							R					nudix (nucleoside diphosphate linked moiety X)-type motif 21
→	YTHDF2	YTHDF2												YTH domain family, member 2
→	TCF7L1	TCF7L1			D				R		S			transcription factor 7-like 1 (T-cell specific, HMG-box)
	TEAD2	TEAD2							R					TEA domain family member 2
	IDH1	IDH1				X	E							isocitrate dehydrogenase 1 (NADP+), soluble
	TRIM28	TRIM28							R					tripartite motif-containing 28
→	KIAA0020	KIAA0020												KIAA0020
	ALPL	ALPL			D									alkaline phosphatase, liver/bone/kidney
	ADSL	ADSL												adenylosuccinate lyase
→	C14ORF106	C14ORF106												
	CPSF3	CPSF3							R					cleavage and polyadenylation specific factor 3, 73kDa
→	ETF1	ETF1						P						eukaryotic translation termination factor 1
→	CDK2AP1	CDK2AP1	G					P						CDK2-associated protein 1
→	YTHDF3	YTHDF3												YTH domain family, member 3
→	NIPSNAP1	NIPSNAP1												nipsnap homolog 1 (C. elegans)
→	CNOT7	CNOT7					E		R		S			CCR4-NOT transcription complex, subunit 7
	NUP107	NUP107		T										nucleoporin 107kDa
	PLS3	PLS3												plastin 3 (T isoform)
	SALL4	SALL4							R					sal-like 4 (Drosophila)
	PARP1	PARP1				X		P	R					poly (ADP-ribose) polymerase family, member 1
	NONO	NONO				X			R					non-POU domain containing, octamer-binding
	TTC27	TTC27												tetratricopeptide repeat domain 27

Figure 6A. Analysis of the biological meaning of the TEAD2 concepts.L2L output of the 35 genes that match between our TEAD2 targets and the L2L microarray list “stemcell_embryonic_up”. The first column displays the name of the genes, the second, the GO functional category to which they are related, and the third indicates their complete name. Arrows points toward those genes harboring at least one TEAD2-binding site in their promoter.

TEAD2-containing concepts.

Among the 21 concepts containing TEAD2 and appearing among the 54 best p-values, 15 were harboring homogeneous situations consisting of Embryonic Stem Cells. Pubmed was then searched using as an entry “TEAD2 embryonic stem cells”. Such a query returned 4 papers, mostly non relevant for establishing a link between TEAD2 and ES cells.

We then analyzed the function of the genes contained in the concepts, by making a complete list of all the genes appearing in the 15 concepts. This left us with a list of 116 genes that we submitted to L2L. We obtained a very significant match (p= 2.82 10-21) with a microarray-based list called “stemcell_embryonic_up (Enriched in mouse embryonic stem cells, compared to differentiated brain and bone marrow cells)”. The 35 genes that match between our TEAD2 targets and the L2L microarray list are displayed in the Figure 6A. We also performed a

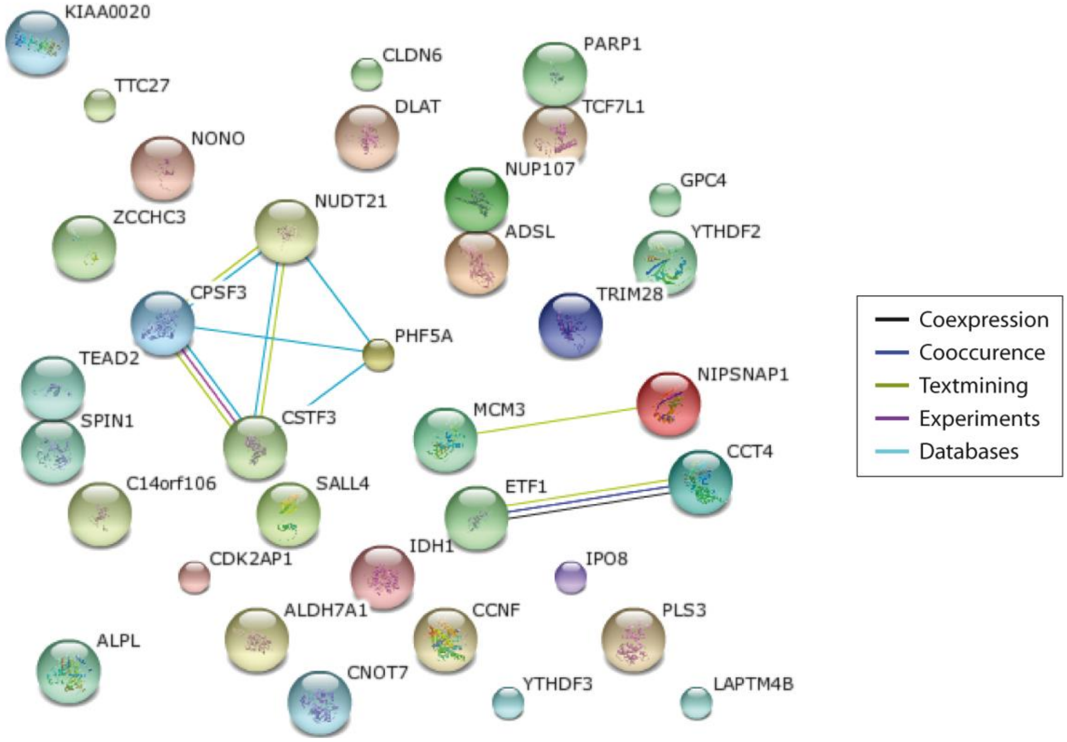


Figure 6B. Analysis of the biological meaning of the TEAD2 concepts.
STRING output showing the known relationship, indicated by a line, between the 35 genes. The color code of the lines is shown in the box on the right (see the STRING website for more information).

STRING analysis, that revealed that none of those genes are known to be connected to TEAD2, and are mostly non connected with each other (Figure 6B).

Altogether our analysis suggest a role for TEAD2 in embryonic stem cells which has until now not been described, together with a list of TEAD2-target genes that might be relevant for its function in human ES cells, half of them being putative direct TEAD2 targets.

DISCUSSION

We have developed a querying process of our SQUAT database, which allows querying simultaneously various sources of information. This allowed us to search for “self-explaining” concept containing a TF together with putative target genes of that TF. We further refined our search by relying on an automatic L2L-based indexing. We finally analyzed two groups of concepts. Both of those were found to be biologically significant with a mixture of both known and new information that indicates a successful data mining quest.

In this work various databases have been used in a sequential fashion, to progressively reduce the amount of extracted information. One possible future direction would consist in mining simultaneously different sources of information. Such a process could be viewed as computing all maximal homogeneous clique sets from different subgraphs. Preliminary evidence that this could be feasible has been obtained recently (Mougel et al., 2010).

Furthermore, it would be of interest to automate the search for interesting concepts. This would require the combination in a single solver of various information sources, an effort that is presently the subject of intense research (Medina et al., 2010; Cao et al., 2011 and references therein).

ACKNOWLEDGEMENTS

This work has been funded by the ANR (French Research National Agency) project BINGO2 (https://bingo2.greyc.fr/) ANR-07-MDCO-014 which is a follow-up of the first BINGO project (2004-2007). We thank all members of the BINGO2 project for stimulating discussion.

REFERENCES

Becquet C, Blachon S, Jeudy B, Boulicaut JF, Gandrillon O (2002). Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. Genome Biol., 3, RESEARCH0067.

Blachon S, Pensa RG, Besson J, Robardet C, Boulicaut J-F, Gandrillon O (2007). Clustering formal concepts to discover biologically relevant knowledge from gene expression data. In Silico Biol., *7*: 0033.

Cao L, Zhang H, Zhao Y, Luo D, Zhang C (2011). Combined mining: discovering informative knowledge in complex data. IEEE Trans Syst Man Cybern B Cybern, 41: 699-712.

Creighton C, Hanash S (2003). Mining gene expression databases for association rules. Bioinformatics, 19: 79-86.

Georgii E, Richter L, Ruckert U, Kramer S (2005). Analyzing microarray data using quantitative association rules. Bioinformatics, 21 Suppl 2, ii123-ii129.

Gershon D (2002). Microarray technology: an array of opportunities. Nature, 416: 885-891.

Hanriot L, Keime C, Gay N, Faure C, Dossat C, Wincker P, Scote-Blachon C, Peyron C, Gandrillon O (2008). A combination of LongSAGE with Solexa sequencing is well suited to explore the depth and the complexity of transcriptome. BMC Genomics, 9: 418.

Keime C, Damiola F, Mouchiroud D, Duret L, Gandrillon O (2004). Identitag, a relational database for SAGE tag identification and interspecies comparison of SAGE libraries. BMC Bioinform., 5: 143.

Kel AE, Gößling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E (2003). MATCHTM : A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res., 31: 3576-3579.

Klema J, Blachon S, Soulet A, Crémilleux B, Gandrillon O (2008). Constraint-Based Knowledge Discovery from SAGE Data. ISB, 8: 0014.

Leyritz L, Schicklin S, Blachon S, Keime C, Robardet C, Boulicaut J-F, Besson J, Pensa RG, Gandrillon O (2008). SQUAT: a web tool to mine human, murine and avian SAGE data. BMC Bioinform., 9: 378.

Li J, Liu H, Downing JR, Yeoh AE, Wong L (2003). Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. Bioinformatics, 19: 71-78.

Madeira SC, Oliveira AL (2004). Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Transactions on Computational Biol. Bioinform., 1: 24-45.

Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A, Tarraga J, Pascual-Montano A, Nogales-Cadenas R, Santoyo J (2010). Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Res., 38, W210-213.

Mougel PN, Plantevit M, Rigotti C, Gandrillon O, Boulicaut JF (2010). Constraint-based Mining of Sets of Cliques Sharing Vertex Properties. In Proc Workshop on the Analysis of Complex Networks ACNE 2010 co-located with ECML PKDD 2010 (Barcelona, M. Berlingerio, B. Bringmann, A. Nürnberger), pp. 48-62.

Newman JC, M Weiner AM (2005). L2L: a simple tool for discovering the hidden significance in microarray expression data. Genome Biol., 2005, 6:R81.

Ng TR, Sander J, Sleumer M (2001). Hierarchical Cluster Analysis of SAGE Data for Cancer Profiling. workshop on Data Mining in BioInformatics with SIGKDD '01.

Pensa R, Leschi C, Besson J, Boulicaut JF (2004). Assessment of discretization techniques for relevant pattern discovery from gene expression data. Paper presented at: 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics BIOKDD'04 co-located with ACM SIGKDD'04 (Seattle, USA).

Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinform., *22*: 1122-1129.

Rioult F, Robardet C, Blachon S, Crémilleux B, Gandrillon O, Boulicaut JF (2003). Mining concepts from large SAGE gene expression matrices. Paper presented at: 2nd Int Workshop Knowledge Discovery in Inductive Databases KDID'03 co-located with ECML-PKDD 2003 (Cavtat-Dubrovnik (Croatia)) STRING. http://string.embl.de/.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995). Serial analysis of gene expression. Sciences, 270:484-487.