Gene Co-Expression in Mouse Embryo Tissues

Simon Andrews[1] and Kenneth M^{c}Leod[2]

[1] Conceptual Structures Research Group
Communication and Computing Research Centre
Faculty of Arts, Computing, Engineering and Sciences
Sheffield Hallam University, Sheffield, UK
s.andrews@shu.ac.uk
[2] School of Mathematical and Computer Sciences,
Heriot-Watt University, Edinburgh, UK
kenneth.mcleod@hw.ac.uk

Abstract. This paper develops some existing ideas in FCA to provide an analysis of a large data set of mouse embryo gene expressions. It develops new techniques for managing complexity and visualisation in FCA to identify and approximate large groups of co-expressed genes. This work has been carried out as part the European CUBIST Project: http://www.cubist-project.eu/

1 Introduction

Formal Concept Analysis (FCA) has already proved useful in the study of gene co-expression. FCA is attractive in the field because formal concepts are natural representations of maximal groups of co-expressed genes. In [5] FCA was used to extract groups of genes with similar expressions profiles from data of the fungus *Laccaria bicolor* and in [4] human SAGE data provides the example from which clusters of concepts with similar properties are visualised. In both approaches the complexity, in terms of the large number of formal concepts present in the raw data, is managed by specifying a concept's minimum size (the well known idea of minimum support in FCA and frequent itemset mining). In [4], tools were developed to query the set of extracted concepts according to various criteria (e.g., presence of a keyword in a gene description) and then to cluster concepts according to similarity, in terms of the attributes (samples) and objects (genes above a threshold of expression) in them. They called these clusters, *quasi-synexpression-groups* (QSGs). By contrast, in [5], ranges of a measure of gene concentration were used as attributes and the genes as objects. Individual concepts that satisfied a specified minimum size were then examined by, for example, plotting the actual measures of concentration of genes together in a line plot.

In this paper we develop some of these ideas and use some freely available, open-source, tools to apply them to a set of mouse-embryo gene expression data. We employ the idea of minimum support to focus on 'large' co-expressions and use a similar notion to that of QSGs in identifying clusters of similar large co-expressions, giving rise to larger *approximate* co-expressions by using a similar

notion to that of FCA 'fault tolerance' [7]. We show that the technique of clustering co-expressions can be straightforward and is a simple way of approximating and visualising a large amount of gene expression data. We demonstrate that FCA can act as a tool for knowledge discovery and can be used to identify possible 'gaps' in knowledge; data that may be missing, erroneous or inconsistent and thus where further investigation or experimentation may be required.

2 The mouse-embryo gene expressions data set: EMAGE

A *gene* is a unit of instructions that provides directions for one essential task, i.e., the creation of a protein. Gene expression information describes whether or not a gene is expressed (active) in a location. Broadly speaking there are two types of gene expression information: those that focus on where the gene is expressed, and those whose primary concern is the strength of expression. This work concentrates on the former category, and in particular a technology called *in situ* hybridisation gene expression.

Information on gene expression is often given in relation to a tissue in a particular model organism. Here the model organism is the mouse. This organism is studied from conception until adulthood. The time window is split into 28 Theiler Stages (TS). Each stage has its own anatomy, and corresponding anatomy ontology called EMAP [3].

Gene expression information allows biologists to discover relationships between genes, in particular when genes are active in the same location. This *co-expression* information provides insights into the ways in which relationships between genes affect the development of a tissue.

The result of an *in situ* experiment is documented as an image displaying an area of a mouse (from a particular Theiler Stage) in which some subsections of the mouse are highly coloured. Areas of colour indicate that the gene is expressed in that location. Additionally, the image provides some indication of the level (strength) of expression: the more intense the colour, the stronger the expression.

Results are analysed manually under a microscope. A human expert determines in which tissues the gene is expressed, and at what level of expression. As volume information is not the main focus of the experiment, its description uses vague natural language terms such as strong, moderate, weak or present. For example, the gene *Bmp4* is strongly expressed in the future brain from Theiler Stage 15.

Completed *in situ* gene expression experiments are published online. One of the main resources in this field is EMAGE [8]. EMAGE documents the result of an experiment using a series of *textual annotations*. Each annotation is a triple: gene - tissue - level of expression. The entire collection of annotations is used as the data set for this work.

For the sake of brevity, both genes and tissues will be referred to by short names or identifiers rather than their full name. For example, the gene "bone morphogenetic protein 4" will be referred to as "*Bmp4*". Likewise, the tissue

"mouse.embryo.skeleton.cranium.viscerocranium.orbito-sphenoid from TS 23" will be known by its unique EMAP identifier "EMAP:8385".

3 An approach using freely available FCA tools

The approach was to convert the EMAGE data into a formal context, mine the context for concepts satisfying a specified minimum size and then approximate the results using FCA 'fault tolerance'. To do this, three tools that are open source and freely available at *Sourceforge* were used:

- **FcaBedrock** [1] to convert the EMAGE data into a formal context by converting *(tissue, level, gene)* triples into *(tissue-level, gene)* pairs.
- **In-Close** [2] to mine to context for concepts satisfying a specified size and produce a corresponding 'reduced' context.
- **Concept Explorer** (ConExp) [10] to visualise the 'large' concepts and apply 'fault tolerance' to produce even larger, 'approximate' concepts.

In addition to their main tasks, In-Close was used to sort the formal context to allow easy identification of clusters of similar concepts (a simple way of finding QSG-type groupings [4]) and ConExp was used as a context editor to extract these clusters and to provide a simple manual method of producing the larger approximate concepts.

3.1 Converting and concept-mining the raw EMAGE data

The EMAGE data set was obtained in the form of csv triples. FcaBedrock was used to automatically convert the data set into a formal context in the standard Burmeister .cxt format. The context contained 6838 attributes (tissue-levels) and 4627 objects (genes). In-Close was used to mine the context generating 208,377 concepts. By a process of trial and error, a minimum size of concept of 14 tissue-levels and 18 genes was determined that produced a reduced context that was possible to visualise in ConExp (Figure 1). Note that the process of visualising the reduced context shows concepts additional to those satisfying the minimum size because where two concepts that satisfy the minimum size 'overlap' in the context grid (share relations), smaller concepts will exist.

3.2 Identification of co-expression clusters

There are two large concepts at the bottom of the lattice in Figure 1 giving a suggestion of two distinct clusters of concepts. A clear visualisation of the two groups is shown in the reduced context produced by In-Close (Figure 2). Because In-Close, as part of its processing, sorts context rows to reduce the difference between them, patterns that would otherwise be difficult to detect become clear. It is apparent that there are two disjoint clusters of concepts, i.e., two disjoint clusters of gene co-expression.

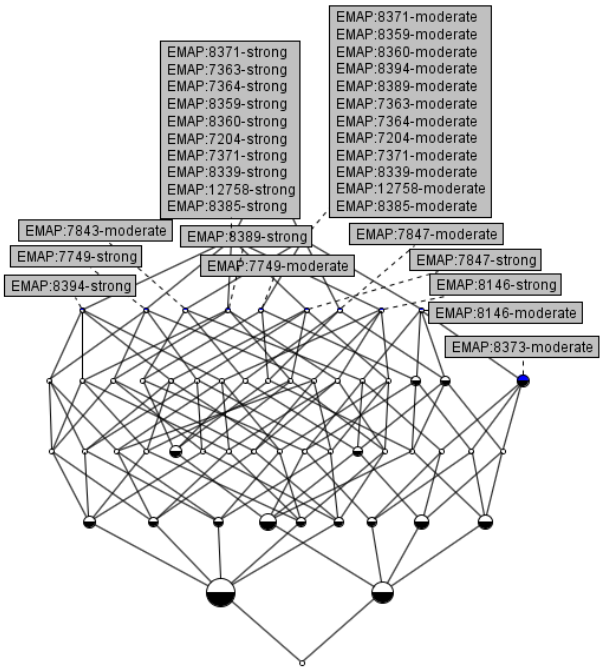

Fig. 1. Concept lattice produced from EMAGE gene expression data (for clarity, only the tissue-levels are displayed)

3.3 Using fault tolerance to produce large approximate concepts

Figures 3 and 4 show the concept clusters as separate context grids. They now appear as dense grids of crosses with only a few crosses 'missing'. The notion of fault tolerance in FCA [7] is that a certain amount of missing information can be tolerated as being errors of omission, or that at least a sensible approximation is possible by adding a limited number of relations to 'complete' a concept. In Figure 3, for example, there is only one cross missing from the column of EMAP:8394-strong. A fault tolerance level of one gene would add that cross and the ones missing for EMAP:7749-strong, EMAP:8389-strong and EMAP:7847-strong. A fault tolerance level of two would also complete the column for EMAP:8146-strong. It is perhaps equally legitimate to apply fault tolerance to missing attributes, thus a fault tolerance level of three would supply all the missing crosses in both grids. Such an approximation results in the lattice in Figure 5.

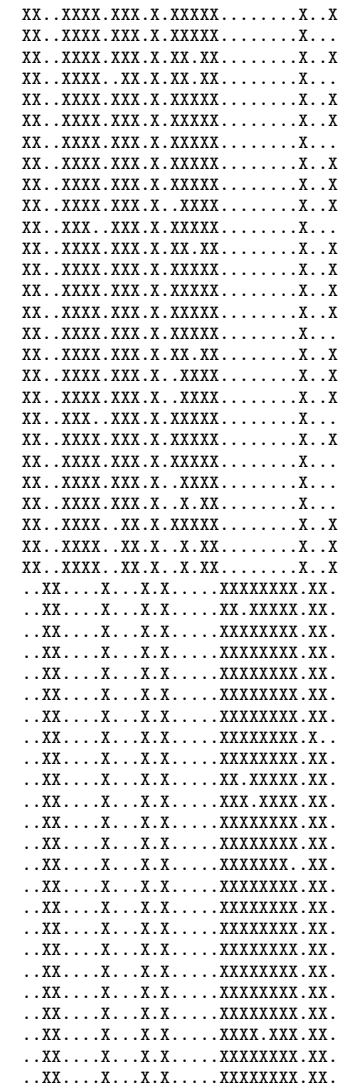

Fig. 2. Two distinct clusters in the reduced context

Gene Co-Exp 1	EMAP:8385-strong	EMAP:12758-strong	EMAP:8339-strong	EMAP:7371-strong	EMAP:7204-strong	EMAP:8360-strong	EMAP:8359-strong	EMAP:8146-strong	EMAP:7847-strong	EMAP:8389-strong	EMAP:7364-strong	EMAP:7363-strong	EMAP:7749-strong	EMAP:8371-strong	EMAP:8394-strong
Mapk8ip2	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Tgfbi	×	×	×	×	×	×	×		×	×	×	×	×	×	×
Zcchc6	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Brpf3	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Caly	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Bcl9l	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Zc3h18	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Dnajc18	×	×	×	×	×	×	×	×	×	×	×	×	×	×	
H2-T22	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Colec12	×	×	×	×	×	×	×		×	×	×	×	×	×	×
Wwp2	×	×	×	×	×	×	×	×		×	×	×	×	×	×
Emp3	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Tcam1	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Papss2	×	×	×	×	×	×	×	×	×	×	×	×		×	×
Ubxn10	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Cytl1	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
BC024814	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Plekhb1	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Apitd1	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Cebpz	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
1110017D15Rik	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Copb1	×	×	×	×	×	×	×	×	×		×	×	×	×	×
Unc5cl	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Haus4	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×

Fig. 3. Cross-table for cluster 1

Gene Co-Exp 2	EMAP:8385-moderate	EMAP:12758-moderate	EMAP:8339-moderate	EMAP:7371-moderate	EMAP:7204-moderate	EMAP:8146-moderate	EMAP:7847-moderate	EMAP:7364-moderate	EMAP:7363-moderate	EMAP:8389-moderate	EMAP:7749-moderate	EMAP:8394-moderate	EMAP:7843-moderate	EMAP:8360-moderate	EMAP:8359-moderate	EMAP:8371-moderate	EMAP:8373-moderate
Sult1c2	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Klk7	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	
Ing4	×	×	×	×	×	×	×	×	×	×	×	×		×	×	×	×
Mir300	×	×	×	×	×	×		×	×	×	×	×		×	×	×	
Plcg2	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
U2af1	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Fzr1	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	
Gm22	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Gm10232	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Tnfsf4	×	×	×	×	×	×	×	×	×	×		×	×	×	×	×	×
Acot6	×	×	×	×	×		×	×	×	×	×	×	×	×	×	×	
Cit	×	×	×	×	×	×	×	×	×	×	×	×		×	×	×	×
Pop4	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Atp6v0a2	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Ebi3	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Fcgrt	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	
Mvk	×	×	×	×	×	×	×	×	×	×	×	×		×	×	×	×
Mdfi	×	×	×	×	×	×	×	×	×	×		×	×	×	×	×	×
Adrm1	×	×	×	×	×	×	×	×	×	×		×	×	×	×	×	×
Clca1	×	×	×	×	×		×	×	×	×	×	×	×	×	×	×	
Tnfaip1	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Arhgap27	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	
Tmc5	×	×	×	×	×	×	×	×	×	×		×	×	×	×	×	
Cops7b	×	×	×	×	×	×	×	×	×	×		×		×	×	×	
Itpr3	×	×	×	×	×	×		×	×	×	×	×	×	×	×	×	×
Tsga10ip	×	×	×	×	×	×		×	×	×		×		×	×	×	×
Upk2	×	×	×	×	×	×		×	×	×		×		×	×	×	×

Fig. 4. Cross-table for cluster 2

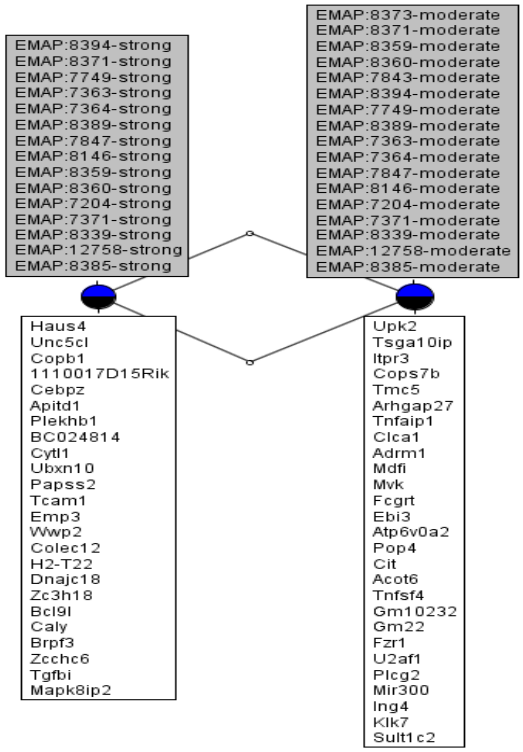

Fig. 5. Original lattice with 'fault tolerance' applied

4 Analysis of the gene co-expression results

When analysing the output of the FCA process, the first task was to convert the EMAP identifiers back into tissue names to determine which locations were flagged. This revealed that with the exception of EMAP:8146 and EMAP:7749, which are both cartilages, all the tissues are bones.

Intuitively all bones will have a similar expression profile, as they are essentially very similar structures. The list of tissues obtained through FCA demonstrates this as it covers the majority of the mouse including the limbs, body, and head. However, interesting gaps remain: for example, in the list there are no bones from the tail. Why not? Answering this question requires further study.

Looking at the output of the above processes, a reader may ask why this expression profile is found in only one of the twenty-eight Theiler Stages? The answer to this question is that TS 23 has the most experimental data; there are many experiments performed on TS 23 that are not repeated on other stages. As such, this pattern of expression may, or may not, be realised in other stages. Until the requisite experiments have been performed it is impossible to tell.

In a similar vein, many of the "missing" crosses from the cross table are a consequence of EMAGE having no experimental result discussing the gene - tissue pair. As such, it is unknown at what level the gene is expressed in the tissue, or if it is expressed at all. Accordingly, the process has revealed future experiments to perform.

Additionally, observe that EMAGE is only one of a number of resources that serve the current domain. Some of these resources provide proprietary *in situ* gene expression information that is not available to EMAGE, whilst others publish the results of different types of gene expression experiment. By reviewing just one extra resource, GXD [9], it is possible to add a cross missing from the initial lattice: *Acot6* - EMAP:8146 - moderately expressed. This leads to the conclusion that if the data from the other resources were integrated with the data from EMAGE it may be possible to add further crosses. Doing so may produce a "better" cross table, and thus a "truer" lattice. Unfortunately, there are significant difficulties in integrating such data [6], and this has been left as future work.

Future work may also investigate whether or not FCA can help resolve inconsistent information. Unfortunately, due to the nature of biology, a small number of textual annotations are inconsistent, i.e., they suggest different levels of expression for the same gene in the same tissue. Perhaps the process documented in this paper can help identify the most likely level. Furthermore, it might be possible to suggest the probable level of expression when EMAGE contains no data.

5 Conclusion

This paper explored FCA within a biological use case. In particular it demonstrated how FCA can be used to analyse *in situ* gene expression data for the developmental mouse.

Analysis was based on large concepts (14 by 18), leaving smaller concepts to be considered as future work. Additionally, further research will be required to understand the full significance of the cross tables documented in this paper.

The list of tissues contained within the cross tables is comprised of a wide selection of bones covering the vast majority of the mouse's skeleton. Yet certain anatomical structures are missing, e.g., the tail. Why are the absent structures not present? What unique features of tail bones prevent them being included in the cross tables?

A further biological question arises in that all the expression levels in each group are the same, i.e., there is a group of genes expressed strongly and a group expressed moderately. There is no reason from an FCA point of view why this should be the case. There may be a biological explanation, perhaps either to do with the nature of the experiments or the nature of the mouse embryo.

From an FCA perspective there are a number outstanding questions too. The appropriateness, and reliability, of fault tolerance needs to be investigated. Additionally, within the context of CUBIST, there is a requirement to improve

the user friendliness of FCA to the extent that a biologist is able to perform the analysis independently of an expert.

Manifestly, the work documented here is at an early stage. Nevertheless, this paper demonstrates there is significant potential that can be exploited for the benefit of both the biological and FCA communities.

Acknowledgement This work is part of the CUBIST project ("Combining and Uniting Business Intelligence with Semantic Technologies"), funded by the European Commission's 7th Framework Programme of ICT, under topic 4.3: Intelligent Information Management.

References

1. Andrews, S., Orphanides, C.: FcaBedrock, a Formal Context Creator. In: Croitoru, M., Ferre, S. and Lukose, D. (eds.): Proceedings of ICCS 2010, Kuching, Malaysia. LNAI 6208, Springer-Verlag (2010)
2. Andrews, S.: In-Close2, a High Performance Formal Concept Miner. In: Hill, R., Andrews, S., Polovina, S., Akhgar, B. (eds.): Proceedings of ICCS 2011, Derby, UK. Springer-Verlag (in press)
3. Baldock, R., Davidson, D.: Anatomy ontologies for bioinformatics: principles and practise, chap. The Edinburgh Mouse Atlas, pp. 249–265. Springer Verlag (2008)
4. Blachona, S., Pensab, R. G., Bessonb, J., Robardetb, C., Boulicautb, J-F., Gandrillona, O.: Clustering Formal Concepts to Discover Biologically Relevant Knowledge from Gene Expression Data. *In Silico* Biology 7, pp. 467–483 467, IOS Press (2007)
5. Kaytoue-Uberall, M., Duplessis, S., Napoli, A.: Using Formal Concept Analysis for the Extraction of Groups of Co-expressed Genes Mehdi Kaytoue-Uberall. In Le Thi, H. A., Bouvry, P., Pham Dinh, T. (eds.): Proceedings of MCO 2008, CCIS 14, pp. 445–455, Springer-Verlag, Berlin Heidelberg (2008)
6. M^{c}Leod, K., Ferguson, G., Burger, A.: Argudas: arguing with gene expression information. In: Paschke, A., Burger, A., Splendiani, A., Marshall, M.S., Romano, P. (eds.) Proceedings of the 3rd International Workshop on Semantic Web Applications and Tools for the Life Sciences (December 2010)
7. Pensa, R. G., Boulicaut, J-F.: Towards Fault-Tolerant Formal Concept Analysis. In: Banidini, S., Manzoni, S. (eds.) AI*IA 2005, LNAI 3673, pp. 212–223, Springer-Verlag, Berlin Heidelberg (2005)
8. Richardson, L., Venkataraman, S., Stevenson, P., Yang, Y., Burton, N., Rao, J., Fisher, M., Baldock, R.A., Davidson, D.R., Christiansen, J.H.: EMAGE mouse embryo spatial gene expression database: 2010 update. Nucleic Acids Research 38, Database issue, D703–D709 (2010)
9. Smith, C.M., Finger, J.H., Hayamizu, T.F., M^{c}Cright, I.J., Eppig, J.T., Kadin, J.A., Richardson, J.E., Ringwald, M.: The mouse gene expression database (GXD) : 2007 update. Nucleic Acids Research 35, D618–D623 (2006)
10. Yevtushenko, S. A.: System of data analysis "Concept Explorer". (In Russian). Proceedings of the 7th national conference on Artificial Intelligence KII-2000, p. 127–134, Russia (2000)