Collective Analysis of Multiple High-Throughput Gene Expression Datasets

Novel Computational Methods & Biological Insights

A thesis submitted for the degree of

Doctor of Philosophy

by

Basel Abu Jamous

Department of Electronic and Computer Engineering

College of Engineering, Design and Physical Sciences

Brunel University London

May 2015

Abstract

Modern technologies have resulted in the production of numerous high-throughput biological datasets. However, the pace of development of capable computational methods does not cope with the pace of generation of new high-throughput datasets. Amongst the most popular biological high-throughput datasets are gene expression datasets (e.g. microarray datasets). This work targets this aspect by proposing a suite of computational methods which can analyse multiple gene expression datasets collectively. The focal method in this suite is the *unification of clustering results from multiple datasets using external specifications (UNCLES)*. This method applies clustering to multiple heterogeneous datasets which measure the expression of the same set of genes separately and then combines the resulting partitions in accordance to one of two types of external specifications; type A identifies the subsets of genes that are consistently co-expressed in all of the given datasets while type B identifies the subsets of genes that are consistently co-expressed in a subset of datasets while being poorly co-expressed in another subset of datasets. This contributes to the types of questions which can addressed by computational methods because existing clustering, consensus clustering, and biclustering methods are inapplicable to address the aforementioned objectives. Moreover, in order to assist in setting some of the parameters required by UNCLES, the *M-N scatter plots technique* is proposed. These methods, and less mature versions of them, have been validated and applied to numerous real datasets from the biological contexts of budding yeast, bacteria, human red blood cells, and malaria. While collaborating with biologists, these applications have led to various biological insights. In yeast, the role of the poorly-understood gene CMR1 in the yeast cell-cycle has been further elucidated. Also, a novel subset of poorly understood yeast genes has been discovered with an expression profile consistently negatively correlated with the well-known ribosome biogenesis genes. Bacterial data analysis has identified two clusters of negatively correlated genes. Analysis of data from human red blood cells has produced some hypotheses regarding the regulation of the pathways producing such cells. On the other hand, malarial data analysis is still at a preliminary stage. Taken together, this thesis provides an original integrative suite of computational methods which scrutinise multiple gene expression datasets collectively to address previously unresolved questions, and provides the results and findings of many applications of these methods to real biological datasets from multiple contexts.

Table of Contents

List of Figures

List of Tables

Acknowledgment

First, I would like to thank my parents for their support without which I would have not been able to achieve what I have achieved. Also, I would like to thank my supervisor Professor Asoke Nandi for his continuous supervision and guidance as well as for the several opportunities with which he has provided me, and is still providing. Additionally, I am grateful to the financial support by the National Institute for Health Research (NIHR) and the Brunel College of Engineering, Design and Physical Sciences. This support has allowed me to focus my research by covering all of the expenses of my Ph.D. studies and provided me with the opportunity to attend few conferences and research meetings nationally and internationally.

A great deal of support, which I acknowledge, has been provided by Professor David Roberts, the Professor of Haematology in the Radcliffe Department of Medicine at the University of Oxford, and the director of the UK National Institute for Health Research (NIHR) programme ‘Erythropoiesis in Health and Disease’. This support was in the form of directing, advising, explaining, reviewing, and giving feedback regarding the biological sides of my research. I have also worked closely and benefited from the commendable collaboration with many of Professor Roberts’ team and colleagues at the University of Oxford including Dr Kathryn Robson, Dr Alison Merryweather-Clarke, Dr Alex Tipping, and Dr Abigail Lamikanra.

Certainly, I owe many other individuals and bodies a lot such as the University of Jordan (Amman, Jordan), where I acquired a strong theoretical foundation, as well as my colleagues and friends.

Proclamation

The author proclaims that the entirety of this document is original and it obeys all valid documents rules and regulations of the College of Engineering, Design and Physical Sciences at Brunel University London.

Chapter 1
Introduction

1.1. Background

Neologism coinage is a natural concomitant of advancements in technology, science, and engineering. The affiliated suffixes, '-ome', '-omic', and '-omics' are examples of such neologisms that have been introduced to English as a consequence of the rise of the age of big data. A 'genome' is the complete set of genes in an organism, 'genomic' matter is that which is related to the complete set of genes in an organism, and 'genomics' refers to study of the complete set of genes in an organism. Similarly, 'proteomes', 'transcriptomes', 'glycomes', and 'metabolomes' are the complete sets of proteins, transcripts, glycans (carbohydrates), and metabolites (small molecules) in an organism, respectively. Consequently, the fields of research considering these omic datasets respectively are 'proteomics', 'transcriptomics', 'glycomics', and 'metabolomics'.

Certainly, these new terms were conceived following the realisation of their implied meanings. For example, transcriptomics was introduced only after the development of arrays of sensors (microarrays) which can measure the abundance of a large set of genetic transcripts in parallel (the abundance of any gene's transcripts, aka *gene expression*, reflects the level of activity of that gene; see Appendix I for more details). Datasets produced by such high-throughput technologies usually include large number of numeric values to the extent that they become no longer feasibly comprehensible by traditional manual means. This burst of data generation necessitates the employment of computational methods that are designed to analyse large amounts of data and guide discovery inference from them. The new interdisciplinary field of research formed by the marriage between biochemistry and computational sciences is now known as *bioinformatics*, which has delivered, and continues to deliver, various key findings in the biological and medical sciences.

As the cost of the omic high-throughput technologies is dropping rapidly while proving their usefulness, they are becoming more readily available to the biochemical community, and consequently are being increasingly utilised to produce more large omic datasets. Elaine Mardis, the Professor of Genetics in the Genome Institute at Washington University, and a collaborator in the 1000 Genomes Project, titled her "musing" published in *Genome Medicine* in 2010 as "the $1,000 genome, the $100,000 analysis?" (Mardis, 2010). Mardis discussed the tremendous drop in the cost of sequencing the complete genome of an individual human from hundreds of millions to few thousand dollars, and that it is expected to reach the line of $1,000. She predicted, based on many facts and observations, that the cost of data analysis, which does not seem to be dropping, will constitute the major part of the total cost rather than the cost of data generation.

Today, tens of thousands of gene expression (transcriptomic) microarray datasets are available, each of which is a large matrix of numbers measuring the expression of a large number of genes over many time-points or conditions (detailed in Section 2.1). Tremendous numbers of datasets have also been produced from other types of high-throughput biological assays. The pace of data generation has neither slowed down nor plateaued.

As a result of this, it is now becoming of substantial importance to design a new generation of computational methods which are not only able to analyse a single massive biological dataset meaningfully but that are also capable of analysing multiple semantically-related high-throughput datasets collectively in order to mine for those findings that are hidden in the aggregation of the datasets in contrast to their individuals. Pick a biological context like erythropoiesis, which is the production of red blood cells in humans and other mammals, many research groups have produced gene expression datasets in this context from different laboratories around the world, by adopting different technologies, and while differing in their exact conditions and environmental parameters (Keller, et al., 2006; Nilsson, et al., 2009; Merryweather-Clarke, et al., 2011). What can we learn about erythropoiesis from such collection of datasets? The new generation of methods should be able to address questions of this type.

Computational methods in bioinformatics do not belong to a single class or paradigm. Rather, they are classified based on their computational approach (e.g. supervised and unsupervised learning, network and graph analysis, statistical methods, etc.) as well as based on the types of biological datasets and questions that they pertain to (e.g. gene expression datasets, proteomics, gene regulation, DNA sequence analysis, etc.). A few methods belonging to some of those classes have already been proposed to investigate some types of multiple datasets collectively. However, not all types of relevant and important questions are targeted by the available literature of methods.

The scope of this thesis is the development and application of methods for unsupervised analysis of multiple heterogeneous gene expression datasets collectively. The main aim of this analysis is to identify the subsets of genes which consistently show similar profiles of gene expression over the given datasets while conforming to different types of external specifications. Despite its importance, no existing method possesses the ability to address this question in an unsupervised fashion.

This thesis presents a novel and thoroughly tested suite of consensus clustering methods that can scrutinise multiple heterogeneous gene expression datasets in order to identify the clusters (subsets) of genes that consistently meet specific external specifications regarding their co-expression (similarity in expression) in the given datasets. Also, this suite is well-equipped with various novel techniques to overcome typical hindrances found in the design and validation of clustering methods such as parameter setting, output validation, the selection of a single result from a set, and the synthesis of artificial datasets with a known ground-truth while faithfully reflecting the properties of real datasets.

As well as the aforementioned significant progress in the design of computational methods, the contributions of this thesis further extend to their applications in the field of biology. A series of applications targeting the molecular biology of budding yeast, human red blood cells production, *E. coli* bacteria, and the malaria disease are presented. The status of these applications varies; a few of them have already been published, some are under consideration for publication, and others represent seeds for future work and personal career development through fellowship-, grant-, and collaboration-hunting.

1.2. Structure of the thesis

The rest of this chapter details the contributions of this thesis and lists my publications. **Chapter 2** reviews the literature of consensus clustering and biclustering methods and their applications in bioinformatics. Of the vast array of literature that discusses unsupervised clustering methods, these two classes of methods are the most relevant to the scope of this thesis. Importantly, this chapter is summarised and concluded in Section 2.6 while enumerating the issues that are poorly addressed by existing methods while being addressed in this thesis. **Chapter 3** describes the methods and techniques that are employed in this thesis. Many of those methods are in reality new methods contributed herein. The introductory paragraph to this chapter explicitly names the sections that present new, in contrast to existing, methods. **Chapter 4** details the sets of experiments conducted to assess the proposed methods and to demonstrate their validity.

Chapter 5 to **Chapter 8** introduce the applications of the aforementioned methods to real biological datasets, their experimental setup, results, conclusions, and my relevant publications when applicable. More explicitly, **Chapter 5**, **Chapter 6**, **Chapter 7**, and **Chapter 8** describe sets of experiments analysing datasets from the contexts of budding yeast, human red blood cell production, *E. coli* bacteria, and malaria, respectively. Each of these chapters also provides a brief biological background regarding its context oriented to non-biological readers. The final chapter, **Chapter 9**, concludes the thesis and provides insights into future work.

The back matter includes some Appendices. **Appendix I** provides a background about cells and their molecular biology, which may prove useful to non-biologist readers. Furthermore, **Appendix II** enumerates the list of references and **Appendix III** is an index.

1.3. Summary of contributions

The original contributions of this thesis can be classified to novel computational methods and biological insights.

1.3.1. Computational methods

All of the proposed computational methods are described in Chapter 3 and assessed and validated in Chapter 4. These methods include:

1. **Bi-CoPaM:** The ***Bi****narisation of* ***Co****nsensus* ***Pa****rtition* ***M****atrices* method is a consensus clustering method which mines for the subsets of genes consistently co-expressed over multiple gene expression datasets. The introduction of Bi-CoPaM is in reality an introduction of a new paradigm in clustering with the ability to produce wide and overlapping clusters and tight and focused clusters in addition to conventional complementary clusters. It also allows a given gene to either belong to multiple clusters simultaneously, to none of the clusters, or to belong to a single cluster exclusively; the latter is the only option offered by conventional clustering methods. Within the course of applying the Bi-CoPaM to multiple datasets, it adopts multiple existing clustering methods to produce intermediate clustering results which are collectively scrutinised in order to produce the final result. Having said that, the Bi-CoPaM has the capacity to exploit existing clustering methods as well as emerging methods. The Bi-CoPaM method is described in Section 3.2 and was published in (Abu-Jamous, et al., 2013a).

2. **UNCLES:** The ***UN**ification of **CL**ustering results from multiple datasets using **E**xternal **S**pecifications* method mines multiple gene expression datasets for subsets of genes that consistently meet given external specifications regarding their co-expression. Two types of external specifications are proposed here; the aim of the first type (type A) is equivalent to the aim of the Bi-CoPaM method, while the aim of the second type (type B) is to identify subsets of genes which are consistently co-expressed in one subset of datasets while being consistently poorly co-expressed in another subset of datasets. UNCLES possesses similar capabilities to the Bi-CoPaM in terms of the flexibility of the types of generated clusters, genes' inclusion in the clusters, and the adoption of various existing clustering methods within its pipeline of steps. This method is explained in Section 3.3 and was published in (Abu-Jamous, et al., 2015c).

3. **M-N scatter plots:** The nature of the clusters generated by the Bi-CoPaM and the UNCLES methods, which vary dramatically in size, renders existing cluster validation techniques inapplicable. Moreover, both methods require setting some parameters such as the number of clusters (K) and some tuning parameters. A technique is proposed here for the validation of clusters of this nature based on my M-N scatter plots. The clusters are scattered on these 2-D plots whose horizontal axes represent a mean-squared error-based (MSE-based) metric and whose vertical axes represent the number of genes included in the clusters on a logarithmic scale. This technique not only ranks the clusters and assists in selecting the best amongst them; it also solves the issue of setting the parameters of the Bi-CoPaM and the UNCLES methods. Thus, the computational framework of the M-N scatter plots in combination with the Bi-CoPaM or the UNCLES methods is an automated parameter-free framework. These plots are described in Section 3.5 and are reported in (Abu-Jamous, et al., 2015c).

4. **F-P scatter plots:** Similar to the M-N scatter plots, F-P plots scatter the clusters over a 2-D plane in order to evaluate their quality. However, F-P plots are only applicable when the ground-truth is available as their horizontal axes represent the false-positive rate (FPR) and their vertical axes represent a modified and normalised p-value. These plots are useful in validating methods when they are tested over synthetic datasets, and have been used here to validate M-N scatter plots. F-P plots are described in Section 3.6 and are described in (Abu-Jamous, et al., 2015c).

5. **Expression data synthesis based on real measurements:** In order to validate the UNCLES method, many sets of datasets were synthesised with different sizes which are designed to have subsets of genes with the properties targeted by UNCLES. In order to do so, a new approach of data synthesis based on real measurements has been proposed to overcome the inaccuracies that other models of data synthesis have. This approach is described in Section 3.7 and in (Abu-Jamous, et al., 2015c).

1.3.2. Biological insights

The methods presented in this thesis have been applied to many real datasets from various biological contexts. Not all of those experiments were performed after the proposal of all of the aforementioned methods, and therefore they do not all exploit the most up-to-date versions of them. Some of the biological experiments were conducted mainly to validate the methods, while others led to important biological insights and findings. The experiments include:

1. **Insights into the yeast CMR1 gene:** Applying the Bi-CoPaM to two filtered yeast cell-cycle datasets revealed four focused clusters. The most focused of them, after maximum tightening, included 19 genes which are largely related to the G1/S stage of the yeast cell-cycle and DNA metabolic processes. A previously poorly understood gene, CMR1, appeared in this focused group leading to many hypotheses relating this gene to the other well-known genes in the cluster. Those biological insights and hypotheses were reported in (Abu-Jamous, et al., 2013b) and are described in Section 5.2.

2. **APha-RiB novel yeast cluster of genes:** Forty genome-wide (unfiltered) yeast datasets from various contexts and sources were collectively analysed by the Bi-CoPaM to reveal that two core clusters of genes with 257 and 47 genes, respectively, out of 5,667 input genes, are consistently co-expressed in all of the forty datasets. The first cluster is the well-known ribosome biogenesis cluster of genes. Strikingly, the second cluster includes genes with mostly unknown or unrelated biological processes and functions. Moreover, this second cluster is consistently negatively correlated with the first one. We therefore named this novel cluster of genes as '*anti-phase with ribosome biogenesis (APha-RiB)*', and drew various hypotheses regarding the function of its genes and their regulation. The findings were reported in (Abu-Jamous, et al., 2014a) and are described in Section 5.3.

3. **Analysis of eight human and murine red blood cells production (erythropoiesis) datasets:** Despite the differences, human and murine erythropoiesis, that is, the biochemical series of interactions and developmental stages leading to the production of mature red blood cells from stem cells, are largely similar. Therefore, eight different human and murine erythropoietic gene expression datasets from different sources were analysed collectively. Out of the 13,269 input genes, five clusters of genes with consistent co-expression over all of the datasets were identified. Various preliminary hypotheses, mainly regarding the transcriptional regulation of the five clusters, were drawn. This work is described in Chapter 6.

4. **Analysis of five *E. coli* bacterial datasets:** Five datasets of the model bacterium *E. coli* from different sources were collectively analysed by the Bi-CoPaM. Two focused and consistently negatively correlated clusters were identified. Although both clusters are enriched with genes with known processes, many of their genes are poorly understood or completely unknown. This experiment and its consequently drawn biological hypotheses were reported in (Abu-Jamous, et al., 2015b) and are described in Chapter 7.

5. **Analysis of malarial blood-stage datasets:** As a part of a new research interest in malaria, and as a foundation for under-construction collaborations with malarial laboratories nationally and internationally, two popular blood-stage malaria parasite's datasets have been analysed by the UNCLES method while adopting the M-N scatter plots. The results, which are presented in Chapter 8, show nine consistently co-expressed clusters of genes which represent a perfect cascade of expression peaks over the parasite's blood-stage cycle. This illustrates the applicability of the method to malarial datasets, the soundness of the biological facts regarding periodicity of expression over the malarial blood-stage cycle, and the credibility and the potential of our approaches in grant-, fellowship-, and collaboration-hunting.

1.4. List of publications

1.4.1. Books

1. Basel Abu-Jamous, Rui Fa, Asoke K. Nandi. *Integrative cluster analysis in bioinformatics*, 2015, John Wiley & Sons.

1.4.2. Journal publications

1. Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi. "UNCLES: method for the identification of genes differentially consistently co-expressed in a specific subset of datasets". *BMC Bioinformatics*, 2015, **16**: 184, doi: 10.1186/s12859-015-0614-0. HIGHLY ACCESSED.

2. Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi. "Application of the Bi-CoPaM method to five *Escherichia coli* datasets generated under various biological conditions". *Journal of Signal Processing Systems*, 2015, **79** (2): 159-166, doi: 10.1007/s11265-014-0919-7.

3. Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi. "Comprehensive analysis of forty yeast microarray datasets reveals a novel subset of genes (APha-RiB) consistently negatively associated with ribosome biogenesis". *BMC Bioinformatics*, 2014, **15**: 322, doi: 10.1186/1471-2105-15-322. HIGHLY ACCESSED.

4. Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi. "Paradigm of tunable clustering using binarization of consensus partition matrices (Bi-CoPaM) for gene discovery". *PLOS ONE*, 2013, **8**(2): e56432, doi: 10.1371/journal.pone.0056432. (>1,900 views).

5. Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi. "Yeast gene CMR1/YDL156W is consistently co-expressed with genes participating in DNA-metabolic processes in a variety of stringent clustering experiments". *Journal of the Royal Society Interface*, 2013, **10**(81): 20120990, doi: 10.1098/rsif.2012.0990.

6. Fengyu Cong, Vinoo Alluri, Asoke K. Nandi, Petri Toiviainen, Rui Fa, Basel Abu-Jamous, Li-yun Gong, Bart G. W. Craenen, et al., "Linking brain responses to naturalistic music through analysis of ongoing EEG and stimulus features". *IEEE Transactions on Multimedia*, 2013, **15**(5): 1060-1069, doi: 10.1109/TMM.2013.2253452.

1.4.2.2. Submitted journal papers

1. Alison Merryweather-Clarke*, Alex Tipping*, Abigail Lamikanra*, Rui Fa*, Basel Abu-Jamous*, H Tsang, Lee Carpenter, Kathryn Robson, Asoke K. Nandi, David J. Roberts. "Distinct gene expression program dynamics during erythropoiesis from human induced pluripotent stem cells compared with adult and cord blood progenitors". *Haematologica* (Submitted). *The first five authors share the same first-author position.

2. Chao Liu, Basel Abu-Jamous, Elvira Brattico, Asoke K. Nandi. "Towards tunable consensus clustering for studying functional brain connectivity during affective processing". *NeuroImage* (Submitted).

1.4.3. Peer-reviewed full-length conference publications

1. Rui Fa, Basel Abu-Jamous, David Roberts, Asoke Nandi. "CoCE-SMART: Consensus clustering based on enhanced splitting-merging awareness tactics". *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, Brisbane, Australia, pp. 962-966.

2. Chao Liu, Rui Fa, Basel Abu-Jamous, Elvira Brattico, Asoke Nandi. "Scalable clustering based on enhanced-smart for large-scale fMRI datasets". *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, Brisbane, Australia, pp. 2011-2015.

3. Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi. "M-N scatter plots technique for evaluating varying-size clusters and setting the parameters of Bi-CoPaM and UNCLES methods". *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, Florence, Italy, pp. 6726-6730.

4. Rui Fa, Basel Abu-Jamous, David J. Roberts, and Asoke K. Nandi. "Splitting-while-merging framework for clustering high-dimension data with component-wise expectation conditional maximisation". *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, Florence, Italy, pp. 2932-2936.

5. Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi. "Bi-CoPaM ensemble clustering application to five Escherichia coli bacterial datasets". *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2014, Lisbon, Portugal.

6. Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi. "Identification of genes consistently co-expressed in multiple microarray datasets by a genome-wide Bi-CoPaM approach". *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, Vancouver, Canada, pp. 1172-1176.

7. Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi. "Method for the identification of the subsets of genes specifically consistently co-expressed in a set of datasets". *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, Southampton, UK, doi: 10.1109/MLSP.2013.6661907.

8. Rui Fa, Basel Abu-Jamous, David J. Roberts, and Asoke K. Nandi. "Enhanced SMART framework for gene clustering using successive processing". *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, Southampton, UK, doi: 10.1109/MLSP.2013.6661964.

9. Rui Fa, Basel Abu-Jamous, and Asoke K. Nandi. "Bi-CoPaM ensemble clustering application to five Escherichia coli bacterial datasets". *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2013, Marrakech, Morocco.

10. Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi. "Hybrid binarisation technique for the Bi-CoPaM method". *Proceedings of the Constantinides International Workshop on Signal Processing (CIWSP)*, 2013, London, UK, doi: 10.1049/ic.2013.0006.

11. Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi. "Comprehensive analysis of multiple microarray datasets by binarization of consensus partition matrix". *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2012, Santander, Spain, paper no. 62, doi: 10.1109/MLSP.2012.6349787.

12. Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi. "Binarization of consensus partition matrix for ensemble clustering". *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2012, Bucharest, Romania, pp. 2193-2197.

13. Rui Fa, Basel Abu-Jamous, and Asoke K. Nandi. "Development and evaluation of kernel-based clustering validity indices". *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2012, Bucharest, Romania, pp. 634-638.

1.4.4. Abstracts and posters

1. Chao Liu, Basel Abu-Jamous, Elvira Brattico, and Asoke K. Nandi. "Consensus clustering reveals neural networks during affective music processing". *The Neurosciences and Music V*, 2014, Dijon, France.

2. Basel Abu-Jamous, Maysam Abbod, Asoke K. Nandi. "Comprehensive analysis of high throughput biological datasets". *Brunel Annual Student Research Conference (ResCon)*, 2014, London, UK.

3. Basel Abu-Jamous, Maysam Abbod, Asoke K. Nandi. "Comprehensive analysis of high throughput biological datasets". *Brunel Annual Student Research Conference (ResCon)*, 2013, London, UK.

4. Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi. "Gene clustering by the binarization of consensus partition matrices (Bi-CoPaM)". *Cognitive Signal Processing Workshop*, 2012, Liverpool, UK.

Chapter 2

Consensus Clustering and Biclustering in Bioinformatics

This chapter reviews the literature of consensus clustering and biclustering as well as their applications in bioinformatics. To start with, Section 2.1 describes the structure of gene expression datasets, which is an introductory section required to understand the data structure to which clustering techniques are applied in this work. Section 2.2 describes the concept of clustering and its relevance to the field of bioinformatics with a demonstrative example that shows some types of findings which may be obtained from such analysis. This example is not meant to be comprehensive or complete and is not included for the sake of its scientific content; rather it is included to demonstrate, in a practical and clear way, how bioinformatics may benefit from clustering. Sections 2.3 and 2.4 detail the concepts of consensus clustering and biclustering, respectively, while describing many methods which belong to them and their classification. Section 2.5 briefly enumerates some applications of consensus clustering and biclustering in bioinformatics while Section 2.6 summarises the chapter.

2.1. Gene expression data structure

Given a set of *N* genes, a gene expression dataset, represented by the matrix $\boldsymbol{X}_{N\times M}$, includes the genetic expression level for each of the *N* genes over *M* different samples (Table 2.1).

Table 2.1. Sample gene expression dataset $X_{N\times M}$

	Sample 1	**Sample 2**	**Sample 3**	**…**	**Sample *M***
Gene 1					
Gene 2					
Gene 3					
Gene 4					
…					
Gene *N*					

These samples may be taken from:

1. semantically different biological conditions, such as different tissue types like the skin, bones, blood, nerves, and others,
2. different time-points in a linearly- or nonlinearly-spaced time series, such as a series of samples taken every *T* minutes, hours, or days from a culture of cells starting from a defined biological state, or
3. different biological stages chronically ordered in a biological process, such as samples taken from initial, intermediate, and late well-defined stages of cells developing in a biological process.

The order of samples is meaningful in the latter two types of gene expression datasets while it is irrelevant in the first one. However, although the order of the samples in the third type is meaningful, absolute time measurement using time units like minutes, hours, or days is not applicable therein. In many cases, we may refer to each of the three types in this document as biological conditions or samples for simplicity as all of them can be seen as different biological conditions from some point of view.

Furthermore, in most of the studies, multiple samples, known as replicates, are obtained for the same biological condition, biological stage, or time-point in order to increase the reliability of the measured expression value. In such cases, summarisation is performed, which results in a single representative value for each biological condition. If the number of replicates is small, the median of their values can be viewed as the most convenient representative value.

2.2. Clustering in bioinformatics

Clustering is an unsupervised learning class of methods in which objects are grouped into a number of clusters such that those objects which are assigned to the same cluster are similar to each other while being dissimilar to the objects assigned to the other clusters based on a given similarity or dissimilarity criterion. Numerous methods have been proposed in the literature to perform this task such as k-means (Pena, et al., 1999), self-organising maps (SOMs) (Kohonen, 1997; Haykin, 1999; Xiao, et al., 2003), hierarchical clustering (HC) (Eisen, et al., 1998), self-organising oscillator networks (SOON) (Rhouma & Frigui, 2001; Salem, et al., 2008), information-based clustering (Slonim, et al., 2005), fuzzy clustering (Baumgartner, et al., 1998), and others.

Clustering methods have been applies to various types of datasets in bioinformatics, the most common of which are gene expression datasets. The structure of gene expression datasets is detailed in Section 2.1. Those genes which are included in the same cluster due to the similarity between their expression profiles are known as co-expressed genes. Genetic co-expression amongst a subset of genes indicates that they may well be co-regulated, that is, their expression levels are regulated by a common regulatory machinery. Moreover, those genes are expected to participate in similar biological processes and pathways.

2.2.1. Demonstrative example of cluster analysis

In order to demonstrate this with an example, we have applied k-means clustering with Kauffman's initialisation (Pena, et al., 1999) to a well-known yeast gene expression dataset of 384 cell-cycle genes measured over 17 times-points from (Cho, et al., 1998; Yeung, 2001). The number of clusters (*K*) was set to four, and therefore four clusters were obtained and respectively labelled as C1, C2, C3, and C4 with the respective numbers of genes of 149, 80, 74, and 81.

Figure 2.1 shows the normalised expression profiles of genes included in each of the four clusters over the 17 times-points. It can be clearly seen in this Figure that the profiles of the genes within any single cluster are similar to each other while being dissimilar to those in the other clusters.

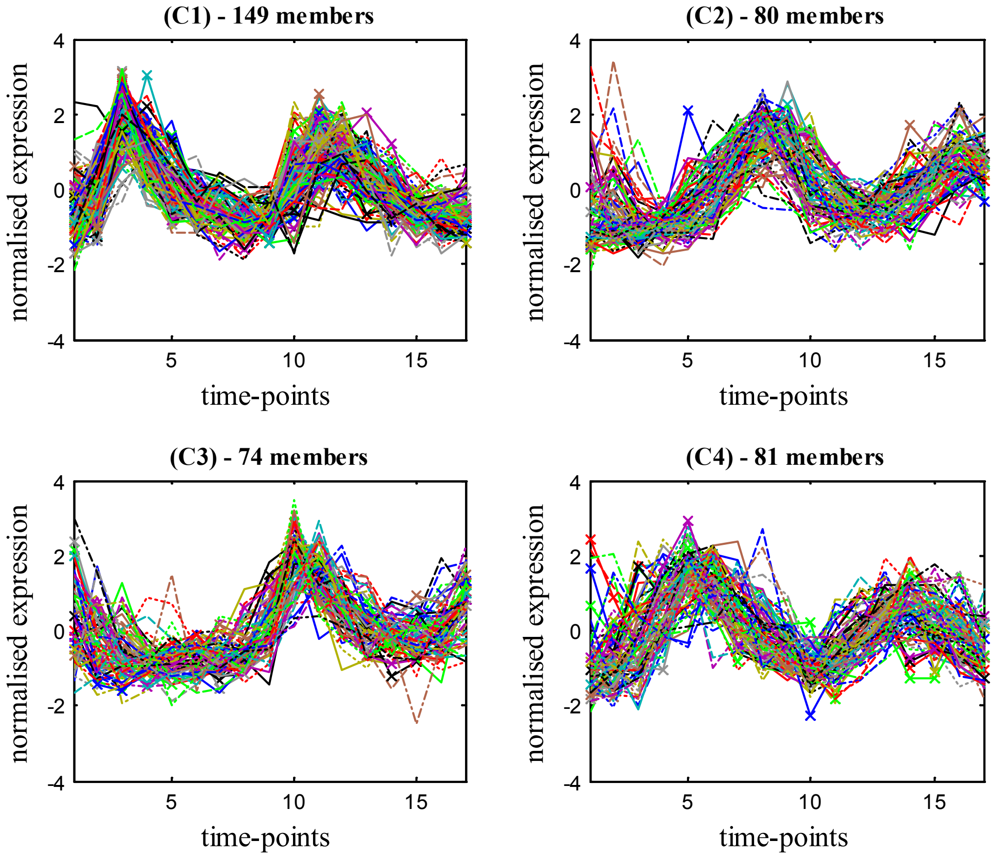

Figure 2.1. Expression profiles of the four yeast clusters from the 384 genes' dataset

By investigating the biological processes in which the genes within those clusters participate, we have found that C1 is highly enriched with DNA metabolism genes (p-value 7.8×10^{-26}), C2 is highly enriched with cell-division genes (p-value 3.5×10^{-5}), C3 is highly enriched with cell-cycle G1/S phase genes (p-value 1.3×10^{-4}), and C4 is highly enriched with chromosome organisation genes (p-value 5.4×10^{-10}). These results have been obtained by the GO Term Finder tool provided by the Saccharomyces Genome Database (SGD) (SGD, 2014) (see Section 3.8). The 17 time-points cover two complete cell-cycles where the exact stage of the cell-cycle represented by any of these time-points can be found in (Cho, et al., 1998).

By exploiting this information, it can be confirmed that the aforementioned biological processes, which are enriched in those four clusters, match the literature of yeast molecular biology (Cho, et al., 1998). For instance, the peak expression of the genes in C1 is at the entrance of the S stage of the cell-cycle in which the DNA needs to be replicated, which is a process undertaken by DNA metabolism genes (Cho, et al., 1998; Spellman, et al., 1998). Similar statements can be drawn regarding the other clusters.

Another type of analysis of the content of those clusters is the further investigation of their potential in being co-regulated in addition to being co-expressed. Upstream sequence analysis (see Section 3.9) identifies those short sequences of DNA (motifs) which are significantly abundant in the upstream sequences of a given subset of genes. Expression regulators, known as transcription factors (TFs), resemble keys which recognise different target motifs, which resemble locks. If a gene's upstream sequence includes the target motif of a TF, given that a few other conditions are met, the TF binds that motif and consequently activates (or represses) the expression of that gene. Thus, the existence of the same motif in the upstream sequences of a group of genes indicates that they may well be co-regulated by a common TF. Refer to Appendix I.H for more about upstream sequence motifs and TFs.

The DREME tool (Bailey, 2011) was used to investigate the upstream sequences of the genes in the cluster C1, and found that its genes are enriched with the MCB motif, which is the target of the TF Mbp1-Swi6; this TF is well-known for activating the expression of the DNA synthesis genes required in the S stage of the cell-cycle (Siegmund & Nasmyth, 1996). Again, a similar experiment can be conducted to investigate the genes in the other clusters.

However, the function, role, and regulation of many genes in those clusters are still unknown or poorly understood. This in reality provides more material to draw new biological hypotheses. For instance, given that few poorly understood genes are co-expressed with a group of genes, do the poorly understood genes participate in the same biological processes in which the well-understood genes in the cluster participate? Are they regulated by the same common TF, especially if that TF's binding site (target motif) was found in their upstream sequences? Those questions represent seeds for guided and focused biological hypotheses elucidating different aspects regarding poorly understood genes.

Taken together, clustering gene expression data lead to various findings and conclusions at the levels of clusters and individual genes. Usually those findings are in the form of hypotheses which need to be followed up by biological functional experiments.

2.3. Consensus clustering

It is generally observed that different clustering results are produced when clustering is applied to the same dataset while adopting different clustering methods, different sets of parameters for the same method, or the same stochastic method over multiple runs (Vega-Pons & Ruiz-Shulcloper, 2011). However, there is no one superior method which overcomes all other methods in quality in all cases. Therefore, it is a common question to ask: which of those different sets of results should be considered, or initially, which clustering method should be adopted?

One approach by which this issue has been tackled by many studies is the employment of *consensus clustering* (Vega-Pons & Ruiz-Shulcloper, 2011). In consensus clustering, the results of applying those multiple methods, or same methods with different parameters, to the same dataset are combined in order to identify a single final consensus result.

Consensus clustering methods can be classified into four different classes based on the style in which they infer the final consensus result from the individual partitions generated by applying different methods and/or sets of parameters independently to the dataset (Abu-Jamous, et al., 2015a). A partition is a clustering result consisting of a set of clusters to which the objects belong with binary or fuzzy membership values. In binary partitions, the belongingness of any object to any cluster is either 1.0 (belongs) or 0.0 (does not belong). In contrary, fuzzy partitions allow for fractional membership values between zero and unity indicating proportional belongingness.

The four classes of consensus clustering methods are:

1. **Partition-partition (P-P) comparison:** while comparing whole partitions with each other, the objective is to maximise the similarity (or to minimise the dissimilarity) between the inferred consensus partition and the individual partitions. The Mirkin distance metric and the information theoretical distance metric are examples of dissimilarity metrics between partitions.

2. **Cluster-cluster (C-C) comparison:** single clusters from different partitions are compared directly in contrast to comparing whole partitions. Graph-based clustering methods are amongst the methods which belong to this class.

3. **Member-in-cluster (MIC) voting:** after matching most-similar clusters from different partitions to each other, partitions vote for the belongingness of different objects to different clusters.

4. **Member-member (M-M) co-occurrence:** a co-association matrix is generated based on the frequency of co-occurrence of pairs of objects, that is, their co-inclusion in the same cluster in different partitions. The final consensus partition matrix is generated based on this co-association matrix.

Some representative consensus clustering methods from these four classes are detailed in the following subsections. More details can be found in the literature review by Vega-Pons and Ruiz-Schulcloper (Vega-Pons & Ruiz-Shulcloper, 2011).

2.3.1. Partition-partition (P-P) comparison methods

Given R partitions $\{\boldsymbol{P}_1 \ldots \boldsymbol{P}_R\}$ generated by applying various clustering methods and/or sets of parameters to a given dataset, P-P comparison methods model the problem as an optimisation problem formulated in this equation (Filkov & Skiena, 2004):

$$\boldsymbol{P}^* = \underset{\boldsymbol{P} \in \mathbb{P}}{\operatorname{argmax}} \sum_{j=1}^{R} \Gamma(\boldsymbol{P}, \boldsymbol{P}_j), \tag{2.1}$$

where $\boldsymbol{P}^*$ is the final consensus partition, $\mathbb{P}$ is the set of all possible partitions, and $\Gamma(.,.)$ is the similarity between the two argument partitions. Therefore, the problem is to identify the partition which is most similar to all of the individual partitions. Note that $\boldsymbol{P}^*$ does not have to be one of the given partitions; rather it can be any partition which belongs to the set of all possible partitions $\mathbb{P}$. Indeed, if a dissimilarity metric $\mathcal{D}(.,.)$, instead of a similarity metric, is used, the problem's formula becomes:

$$\boldsymbol{P}^* = \underset{\boldsymbol{P} \in \mathbb{P}}{\operatorname{argmin}} \sum_{j=1}^{R} \mathcal{D}(\boldsymbol{P}, \boldsymbol{P}_j). \tag{2.2}$$

A popular partition dissimilarity metric is the Mirkin distance $\mathcal{M}(.,.)$ (Mirkin, 1996), which is calculated by:

$$\mathcal{M}(\boldsymbol{P}_1, \boldsymbol{P}_2) = n_{01} + n_{10}, \tag{2.3}$$

where n_{01} is the number of pairs of objects which are included in different clusters in $\boldsymbol{P}_1$ but in the same cluster in $\boldsymbol{P}_2$, while n_{10} is the number of pairs of objects which are included in the same cluster in $\boldsymbol{P}_1$ but in different clusters in $\boldsymbol{P}_2$. In other words, the Mirkin distance is the number of pairs of objects on which there is a disagreement between the two partitions. For completion, n_{11} and n_{00} are the numbers of pairs of objects which are included and not included, respectively, in the same cluster by the agreement of both partitions.

With this, the optimisation problem can be rewritten as:

$$\boldsymbol{P}^* = \underset{\boldsymbol{P}\in\mathbb{P}}{\operatorname{argmin}} \sum_{j=1}^{R} \mathcal{M}\left(\boldsymbol{P}, \boldsymbol{P}_j\right). \tag{2.4}$$

Many methods have been proposed to solve this problem such as the trivial pick-a-cluster method, the best-of-k (BOK) method (Filkov & Skiena, 2004), which is also known as the best-clustering method (Bertolacci & Wirth, 2007), the balls algorithm (Gionis, et al., 2007), the CC-pivot algorithm (Ailon, et al., 2008), and others. Pick-a-cluster, which can be more accurately named as pick-a-partition, simply and trivially picks one of the individual partitions $\{\boldsymbol{P}_1 \dots \boldsymbol{P}_R\}$ as the final partition $\boldsymbol{P}^*$. As for the BOK algorithm, it selects the partition, amongst the individual partitions, which is most similar to all of the other partitions (Filkov & Skiena, 2004); note that it only searches in the *R* generated partitions in $\{\boldsymbol{P}_1 \dots \boldsymbol{P}_R\}$ and not in all possible partitions in $\mathbb{P}$. The balls and the CC-pivot algorithm are more sophisticated as they attempt at solving the problem while considering a graph representation for the data (Gionis, et al., 2007; Ailon, et al., 2008).

Other popular metrics are those which are based on information theory (Topchy, et al., 2005). Let the r^{th} partition be $\boldsymbol{P}_r = \left\{\boldsymbol{C}_1^{(r)} \dots \boldsymbol{C}_{K^{(r)}}^{(r)}\right\}$, where $\boldsymbol{C}_k^{(r)}$ is this partition's k^{th} cluster and $K^{(r)}$ is the number of clusters in it. The amount of information between the r^{th} and the s^{th} partitions $I(\boldsymbol{P}_r, \boldsymbol{P}_s)$ can be expressed as:

$$I(\boldsymbol{P}_r, \boldsymbol{P}_s) = \sum_{i=1}^{K^{(r)}} \sum_{j=1}^{K^{(s)}} p\left(\boldsymbol{C}_i^{(r)}, \boldsymbol{C}_j^{(s)}\right) \log\left(\frac{p\left(\boldsymbol{C}_i^{(r)}, \boldsymbol{C}_j^{(s)}\right)}{p\left(\boldsymbol{C}_i^{(r)}\right) p\left(\boldsymbol{C}_j^{(s)}\right)}\right). \tag{2.5}$$

The optimal consensus partition $\boldsymbol{P}^*$ can therefore be found by solving the following optimisation problem:

$$\boldsymbol{P}^* = \underset{\boldsymbol{P}\in\mathbb{P}}{\operatorname{argmax}} \sum_{j=1}^{R} I(\boldsymbol{P}, \boldsymbol{P}_j). \tag{2.6}$$

2.3.2. Cluster-cluster (C-C) comparison methods

The meta-clustering algorithm (MCLA), proposed by Strehl and Ghosh, is a typical C-C comparison method (Strehl & Ghosh, 2003). A graph is constructed where each cluster amongst the clusters in the *R* partitions is represented by a hyperedge connecting the objects which it includes. If the number of clusters in the r^{th} partition is K_r, the total number of hyperedges in this graph will be $\sum_{r=1}^{R} K_r$. Then, a meta-graph is constructed by considering each of the indicator vectors $\boldsymbol{h}$ for the $\sum_{r=1}^{R} K_r$ hyperedges as a vertex in this meta-graph. The edges in this undirected meta-graph have weights proportional to the similarity between the vertices. This similarity can be measured by the Jaccard measure. After that, the vertices in the meta-graph, representing the original clusters, are clustered into *K* meta-clusters of clusters. Afterwards, the hyperedges in each meta-cluster are collapsed into a single hyperedge by averaging their indicator vectors $\boldsymbol{h}$. Finally, each object is assigned to the meta-cluster with which it has the highest association value. Extra details on this method can be found in the study by Strehl and Ghosh (Strehl & Ghosh, 2003).

2.3.3. Member-in-cluster (MIC) voting methods

The general theme of MIC voting methods is that partitions vote for the inclusion of each object in clusters, and the clusters finally include those objects for which they get some sort of majority votes.

A popular method in this class is the relabelling and voting method. In its basic terms, the clusters in each of the *R* individual partitions are permuted so that the k^{th} cluster from a given partition best matches the k^{th} cluster from each of the other partition; this step is called *relabelling*. This is an essential step in this method because clustering is unsupervised, and therefore, such alignment of clusters is not guaranteed unless it is deliberately done by relabelling. Accordingly, those clusters which are mapped to each other from different partitions are considered as different versions for the same consensus cluster, and thus are assigned the same label. Consequently, each object is included in the consensus cluster (cluster label) to which more partitions assign it, that is, which is granted the majority of the votes.

The voting-merging (VM) algorithm (Dimitriadou, et al., 2001) and the cumulative voting algorithm (Ayad & Kamel, 2008) are examples of variants of relabelling and voting methods.

Other MIC voting methods include those which are based on mixture models. The mixture model-based method which was proposed by Topchy and colleagues models the memberships of genes in clusters as random variables drawn from a probability distribution described as a mixture of multivariate component densities (Topchy, et al., 2005). The model is formulated as:

$$f(\boldsymbol{y}_n|\boldsymbol{\Theta}) = \sum_{k=1}^{K} \alpha_k f(\boldsymbol{y}_n|\boldsymbol{\theta}_k), \tag{2.7}$$

where $\boldsymbol{\Theta}$ is the set of the parameters $\{\alpha_1, \dots, \alpha_K, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$ corresponding to the K clusters/labels, $\boldsymbol{y}_n$ is the random variable corresponding to the membership values of the n^{th} gene in the clusters, and $f(.\,|.)$ is the conditional probability distribution function. The problem is formulated afterwards as a minimum likelihood estimation problem which is solved by the expectation minimisation (EM) algorithm (Topchy, et al., 2005).

2.3.4. Member-member (M-M) co-occurrence methods

M-M co-occurrence methods convert the consensus clustering problem to a co-association matrix partitioning problem. The (i, j) entry of the co-association matrix, denoted as A_{ij}, is the frequency of the co-appearance of the objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in all of the R partitions; it is expressed as:

$$A_{ij} = \frac{1}{R} \sum_{r=1}^{R} \delta\left(\boldsymbol{P}_r(\boldsymbol{x}_i), \boldsymbol{P}_r(\boldsymbol{x}_j)\right), \tag{2.8}$$

where $\boldsymbol{P}_r(\boldsymbol{x})$ is the cluster to which the object $\boldsymbol{x}$ is assigned in the r^{th} partition and $\delta(a, b)$ is 1 if $a = b$ and is 0 otherwise. It is worth mentioning that, in this setup, it is not a condition to adopt the same number of clusters in all of the individual partitions.

There are various methods which have been designed to extract the consensus partition from this co-association matrix such as the evidence-accumulation method (Fred & Jain, 2005), graph-based methods (Strehl & Ghosh, 2003), hypergraph-based methods (Strehl & Ghosh, 2003), and resampling methods (Monti, et al., 2003).

Fred and Jain (2005) named the construction of the co-association matrix as *evidence accumulation* by considering that each partition assigning a given pair of objects to the same cluster as an evidence of the inclusion of this pair in the same cluster. Therefore, the entries of the co-association matrix are considered as normalised accumulated evidence of the inclusion of any given pair of objects in the same cluster. They propose obtaining the final consensus partition by applying single linkage or average linkage agglomerative hierarchical algorithm to the objects in the co-association matrix. As for the number of clusters in the consensus partition, it is identified by the range of threshold values on the dendogram that lead to the identification of the clusters (Fred & Jain, 2005).

Strehl and Ghosh proposed a cluster-based similarity partitioning algorithm (CSPA) as a graph-based algorithm, and a hypergraph partitioning algorithm (HGPA) (Strehl & Ghosh, 2003). The concatenated block of individual partition matrices $\boldsymbol{H} = [\boldsymbol{P}_1 \dots \boldsymbol{P}_R]$ defines a hypergraph $\boldsymbol{H}$. The co-association matrix $\boldsymbol{A}$ can therefore be obtained by:

$$\boldsymbol{A} = \frac{1}{R}\boldsymbol{H}\boldsymbol{H}^T. \tag{2.9}$$

CSPA adopts a graph partitioning algorithm, such as METIS[1] (Karypis & Kumar, 1995; Karypis & Kumar, 1998), to cluster the objects in this co-association matrix. On the other hand, HGPA applies clustering to the hypergraph $\boldsymbol{H}$ itself by using the hypergraph partitioning package HMETIS (Strehl & Ghosh, 2003).

Resampling techniques were employed in the design of M-M co-occurrence consensus clustering methods as in the method proposed by Monti and colleagues (Monti, et al., 2003). This method provides consensus across multiple runs of a given clustering algorithm while assessing the stability of the discovered clusters. The results of this method can be graphically visualised and incorporated in the decisions about the number of clusters and cluster membership, which is a key feature of this resampling method. This method is based on the assumption that the membership of genes in their corresponding natural clusters should not change radically when a clustering algorithm is applied repeatedly to the given dataset after resampling.

The dataset is perturbed R times to produce R perturbed datasets $\{\boldsymbol{X}^1 \dots \boldsymbol{X}^R\}$ which are clustered by a given clustering algorithm to produce R partitions $\{\boldsymbol{P}^1 \dots \boldsymbol{P}^R\}$. The R co-association matrices formed based on those partitions are normalised and combined to produce a consensus co-association matrix which is clustered based on an agglomerative hierarchical clustering (HC) method to produce a tree (dendogram) of clusters. Summary statistics are calculated for the clusters in the dendogram to quantify their stability, rank them accordingly, and determine the best number of clusters. Further details can be found in (Monti, et al., 2003).

[1] 'METIS' refers to wisdom as derived from ancient Greek mythology.

2.4. Biclustering

Biclustering methods aim at clustering a gene expression data matrix based on each of its two dimensions, that is, based on the expression profiles of genes over samples and the expression profiles of samples over genes. By this, genes which are co-expressed over a subset of samples, or samples which have similar profiles over a subset of genes can be identified. A *bicluster* is therefore defined by a specific subset of genes expressed over a specific subset of samples, and represents a submatrix of the original expression data matrix.

Biclustering has gained much interest since Cheng and Church proposed their biclustering algorithm (CC) in 2000 (Cheng & Church, 2000). Now there are more than thirty different biclustering algorithms in the literature, many surveys and performance comparison papers (Madeira & Oliveira, 2004; Prelić, et al., 2006; Eren, et al., 2013; Oghabian, et al., 2014; Tchagang, et al., 2011), and many toolboxes in many different platforms available (Barkow, et al., 2006; Kaiser & Leisch, 2008; Eren, 2012).

Oghabian and colleagues proposed a classification of biclustering methods based on the criteria of identifying the biclusters (Oghabian, et al., 2014). This taxonomy classifies biclustering methods into the following four classes:

1. **Variance-minimisation biclustering methods (VMB):** VMB searches for biclusters in which expression values have low variance throughout the selected genes, samples, or the whole submatrix.
2. **Correlation-maximisation biclustering methods (CMB):** CMB mines for the subsets of genes and samples for which the expression values of the genes correlate highly among the samples.
3. **Two-way clustering methods (TWC):** TWC discovers the homogeneous subsets of genes and samples; that is, biclusters, by iteratively performing one-way clustering on the genes and samples.
4. **Probabilistic and generative methods (PGM):** PGM employs probabilistic techniques to discover genes (or, respectively, samples) that are similarly expressed across a subset of samples (or, respectively, genes) in the data matrix.

Some details on some methods belonging to these four classes of biclustering methods are presented in the following subsections.

2.4.1. Variance-minimisation biclustering methods (VMB)

VMB methods search for biclusters in which expression values have low variance throughout the selected genes, samples, or the whole submatrix. Examples of VMB methods are the Cheng and Church (CC) method (Cheng & Church, 2000) and spectral biclustering (Kluger, et al., 2003).

Let a bicluster (submatrix of the data matrix) be defined by a subset of genes (rows of the data matrix) $\boldsymbol{I}$ and a subset of samples (columns of the data matrix) $\boldsymbol{J}$. The CC algorithm defines a *mean-squared residue (MSR)* metric as:

$$\text{MSR} = \frac{1}{|\boldsymbol{I}||\boldsymbol{J}|} \sum_{i \in \boldsymbol{I}, j \in \boldsymbol{J}} \left(x_{ij} - \bar{x}_{i\boldsymbol{J}} - \bar{x}_{\boldsymbol{I}j} + \bar{x}_{\boldsymbol{IJ}} \right), \tag{2.10}$$

where x_{ij} is the expression value at the i^{th} row and the j^{th} column, $\bar{x}_{i\boldsymbol{J}}$ is the mean expression of the i^{th} row over all of the $\boldsymbol{J}$ columns, $\bar{x}_{\boldsymbol{I}j}$ is the mean expression of the j^{th} column over all of the $\boldsymbol{I}$ rows, and $\bar{x}_{\boldsymbol{IJ}}$ is the mean expression of the submatrix defined by the $\boldsymbol{I}$ rows and the $\boldsymbol{J}$ columns.

CC starts with the whole data matrix and removes the rows and columns that have high residues gradually. Once the MSR of the bicluster reaches a given threshold, δ, the rows and columns that produce smaller residue values than the bicluster residue are added back to the bicluster. The found biclusters are masked with random values and then the process repeats until no biclusters can be found.

The spectral biclustering algorithm (Kluger, et al., 2003) assumes that different subsets of genes have high expression values at different subsets of samples, and, if the rows and columns of the data matrix are reordered appropriately, the data matrix will have a checkerboard-like appearance with blocks of high expression values and blocks of low expression values. The objective of spectral biclustering is to identify this checkerboard-like structure.

Biclustering consists of several steps: (i) simultaneous normalisation of genes and samples, (ii) post-processing of eigenvectors to find partitions, and (iii) probabilistic interpretation. The first step is performed by independent scaling of rows and columns iteratively until convergence, which is defined by having all of the rows sum to a constant and all of the columns sum to another constant; this process is known as *bistochatisation* (Kluger, et al., 2003). Singular value decomposition (SVD) is applied afterwards to the normalised matrix producing a set of eigenvectors and eigenvalues. The largest non-trivial eigenvectors are then clustered, for example by k-means. Finally, the degrees of membership of different genes and samples to the biclusters identified by partitioning the eigenvectors are ranked.

2.4.2. Correlation-maximisation biclustering methods (CMB)

CMB methods mine for the subsets of genes and samples for which the expression values of the genes correlate highly amongst the samples. BiMine (Ayadi, et al., 2009), bimax (Prelić, et al., 2006), and the robust biclustering algorithm (ROBA) (Tchagang & Tewfik, 2006) are examples of CMB methods.

BiMine is a typical CMB method which relies on the average Spearman's rho (ASR) evaluation function which guides effective exploration of the search space. Spearman's rank correlation is formulated as:

$$\rho_{ij} = 1 - \frac{6\sum_{k=1}^{m}\left(r_k^i\left(x_k^i\right) - r_k^j\left(x_k^j\right)\right)^2}{m(m^2-1)}, \tag{2.11}$$

where $r_k^i\left(x_k^i\right)$ is the rank of x_k^i, and m is the size of the data vector. Thereafter, ASR is formulated as:

$$\mathrm{ASR}(\boldsymbol{I},\boldsymbol{J}) = 2 \cdot \max\left\{\frac{\sum_{i\in\boldsymbol{I}}\sum_{j\geq i+1, j\in\boldsymbol{J}}\left(\rho_{ij}\right)}{|\boldsymbol{I}|(|\boldsymbol{I}|-1)}, \frac{\sum_{k\in\boldsymbol{J}}\sum_{l\geq k+1, l\in\boldsymbol{I}}\left(\rho_{kl}\right)}{|\boldsymbol{J}|(|\boldsymbol{J}|-1)}\right\}. \tag{2.12}$$

The values of $\mathrm{ASR}(\boldsymbol{I},\boldsymbol{J}) \in [-1, 1]$ which are closer to 1.0 indicate higher correlation between the given vectors within the bicluster. BiMine uses a tree-structure called the bicluster enumeration tree (BET) to represent the hierarchy of the discovered candidate biclusters throughout the process of maximising the ASR value.

In contrast to the previous methods, the bimax method considers binary expression data in which the expression value of a given gene at a given sample is either one (expressed) or zero (not expressed). Therefore, non-binary data is binarised before consequent bimax steps are performed. Ideally, a bicluster, as defined by this method, is that submatrix of the data matrix which only includes ones, and that is not entirely a sub-bicluster of another larger bicluster. Bimax adopts a divide-and-conquer incremental procedure proposed by Alexe and colleagues in order to identify those biclusters (Alexe, et al., 2004).

ROBA aims at identifying all perfect biclusters in a dataset in a timely manner by employing linear algebraic and arithmetic, contrary to heuristic, tools and methods (Tchagang & Tewfik, 2006).

2.4.3. Two-way clustering methods (TWC)

TWC methods mine for homogeneous biclusters by iteratively performing one-way clustering on genes and samples. Coupled two-way clustering (CTWC) (Getz, et al., 2000) and interrelated two-way clustering (ITWC) (Tang, et al., 2001) are two examples of TWC methods.

CTWC provides an efficient heuristic approach which restricts the number of candidate biclusters, that is, submatrices that can be formed based on a given dataset to a feasible range instead of exponentially increasing with the size of the dataset. This is done by starting with the entire dataset as a single bicluster and then performing iterative two-way clustering to the two dimensions in order to find those genes and samples that form stable biclusters which are further clustered to form child sub-biclusters. When no further biclusters can be found based on given criteria, the algorithm terminates (Getz, et al., 2000).

ITWC, also, is an iterative method where each iteration starts by clustering the dataset in the genes dimension by any clustering method. Then, each produced cluster is clustered into two clusters in the samples dimension. After that, the clustering results from the previous two steps are combined and heterogeneous groups, that is, pairs of groups whose samples are not grouped in any cluster, are identified. Finally, the most distant third of genes belonging to heterogeneous groups are selected as a cluster while the rest of the genes are forwarded to the following iteration.

2.4.4. Probabilistic and generative methods (PGM)

PGM methods apply probabilistic techniques to discover genes (or, respectively, samples) that are similarly expressed across a subset of samples (or, respectively, genes) in the expression data matrix. Plaid (Lazzeroni, et al., 2002), Bayesian Plaid (Caldas & Kaski, 2008), and CMonkey (Reiss, et al., 2006) are examples of biclustering methods belonging to this class.

Plaid aims at reordering the genes and the samples so that the data matrix, visualised as a heat map, shows K rectangular blocks of high expression values. Each of the blocks represents a bicluster whose genes are only expressed in its samples. The data is modelled as the superposition of a background layer ($k = 0$) and K layers ($k = 1 \dots K$) representing K clusters/blocks:

$$x_{ij} = \sum_{k=0}^{K} \theta_{ijk} \rho_{ik} \kappa_{jk}, \tag{2.13}$$

where x_{ij} is the expression value of the i^{th} gene in the j^{th} cluster, θ_{ij0} is the background colour, θ_{ijk} is the colour in the k^{th} block in addition, if needed, to the specific response of the i^{th} gene over a subset of samples and/or the specific response of the j^{th} sample over a subset of genes, ρ_{ik} is one if the i^{th} gene belongs to the k^{th} cluster/block and zero otherwise, and κ_{jk} is one if the j^{th} sample belongs to the k^{th} cluster/block. Plaid aims at minimising the cost function:

$$Q = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(x_{ij} - \theta_{ij0} - \sum_{k=1}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk}\right)^2. \qquad (2.14)$$

Bayesian Plaid models all of the variables assumed by the Plaid model as random variables following appropriate distributions (Caldas & Kaski, 2008). The components of the θ_{ijk} parameter, namely the colour of k^{th} block, the specific response of the i^{th} gene, and the specific response of the j^{th} sample, are assumed as Gaussian variables while the variables ρ_{ik} and κ_{jk} are binomial variables.

The CMonkey method combines gene expression data, DNA-sequence data and associated network data to produce biclusters based on a probabilistic model (Reiss, et al., 2006). Each bicluster is modelled by the Markov chain process, in which the bicluster is iteratively optimised, and its state is updated based on conditional probability distributions computed using the cluster's previous state. Biclusters are initialised by one of different seeding methods and are consequently iteratively optimised by adding/removing genes and samples.

2.5. Applications in bioinformatics

2.5.1. Consensus clustering methods

As discussed above, Monti and colleagues developed a resampling method of class discovery and clustering validation tailored to the task of analysing gene expression data (Monti, et al., 2003). They applied their resampling-based consensus clustering to six real gene expression datasets, namely leukaemia dataset (Golub, et al., 1999), Novartis multi-tissue (Su, et al., 2002), St. Jude leukaemia (Yeoh, et al., 2002), lung cancer (Bhattacharjee, et al., 2001), central nervous system tumours (Pomeroy, et al., 2002), and normal tissue (Ramaswamy, et al., 2001). They found that, in general, adopting hierarchical clustering as an underlying basic method while applying this resampling method to gene expression data outperforms adopting the SOM method (Monti, et al., 2003).

Swift and colleagues developed a consensus clustering algorithm which, according to their investigation, improves confidence (Swift, et al., 2004). They used the weighted-kappa metric, which was originally proposed by Cohen (Cohen, 1968), as a direct measure of similarity of partitions. A consensus strategy was applied to produce both robust and consensus clustering of gene expression data and assign statistical significance to these clusters from known gene functions. The method is different from the afore-discussed resampling method (Monti, et al., 2003) in that different clustering algorithms are used rather than perturbing the gene expression data for a single algorithm. Using consensus clustering with probabilistic measures of cluster membership derived from external validation with gene function annotations, specific transcriptionally co-regulated genes from microarray data of distinct B-cell lymphoma types (Jenner, et al., 2003) was identified accurately and rapidly.

Brannon and colleagues analysed gene expression microarray data using software that implements iterative unsupervised consensus clustering algorithms to identify the optimal molecular subclasses, without clinical or other classify information (Brannon, et al., 2010). Clear cell renal cell carcinoma (ccRCC) is the predominant RCC subtype, but even within this classification, the natural history is heterogeneous and difficult to predict. ConsensusCluster was proposed by Seiler and colleagues (Seiler, et al., 2010), for the analysis of high-dimensional single nucleotide polymorphism (SNP) and gene expression microarray data. The software implemented the consensus clustering algorithm and PCA to stratify the data into a given number of robust clusters. The robustness is achieved by combining clustering results from data and sample resampling as well as by averaging over various algorithms and parameter settings to achieve accurate, stable clustering results. Several different clustering algorithms have been implemented, including k-means, PAM, SOMs, and hierarchical clustering (HC) methods. After clustering the data, ConsensusCluster generates a consensus matrix heat map to give a useful visual representation of cluster membership, and automatically generates a log of selected features that distinguish each pair of clusters. Such consensus clustering analysis identified two distinct subtypes of ccRCC, designated clear cell type A and B. In each subtype, logical analysis of data defined a small, highly predictive gene set that could then be used to classify additional tumours individually. The subclasses were corroborated in a validation data set of 177 tumours and analysed for clinical outcome. Based on individual tumour assignment, tumours designated type A had markedly improved disease-specific survival compared to type B. Using patterns of gene expression based on a defined gene set, ccRCC was classified into two robust subclasses based on inherent molecular features that ultimately corresponded to marked differences in clinical outcome. This classification schema thus provided a molecular stratification applicable to individual tumours that may have implications to influence treatment decisions, define biological mechanisms involved in ccRCC tumour progression, and direct future drug discovery.

2.5.2. Biclustering methods

Most of the available biclustering methods have been applied to bioinformatic datasets leaving us with a rich literature of such applications. For instance, Tchagang and colleagues employed ROBA biclustering to identify group biomarkers using microarray gene expression data of ovarian cancer (Tchagang, et al., 2008). Huttenhower and colleagues proposed the combinational algorithm for expression and sequence-based cluster extraction (COALESCE) system for regulatory module prediction (Huttenhower, et al., 2009). Bryan and colleagues were the first to apply biclustering techniques to model functional modules within an integrated microRNA (miRNA)-messenger RNA (mRNA) association matrix (Bryan, et al., 2014).

Other applications include the analysis of yeast cell cycle datasets (Cho, et al., 1998; Spellman, et al., 1998), yeast stress datasets (Gasch, et al., 2000; Gasch, et al., 2001), yeast compendium (Hughes, et al., 2000), yeast galactose utilisation (Ideker, et al., 2001). Other algorithms were used for human breast tumour (Pawitan, et al., 2005; Miller, et al., 2005; Loi, et al., 2007), lymphoma (Alizadeh, et al., 2000), and leukaemia (Golub, et al., 1999).

2.6. Discussion

Clustering methods group a given set of objects (e.g. genes) into a number of clusters such that those objects which are included in the same cluster are similar to each other while being dissimilar to the objects included in the other clusters. In the context of gene expression data clustering, genes are clustered into groups based on their co-expression, that it, their expression profiles similarity over a number of time-points or samples from different conditions.

It is well-known that applying different clustering methods to the same dataset does not produce identical results. The same observation is true when the same stochastic clustering method is applied to the same dataset multiple times or with different sets of parameters. Many consensus clustering methods were designed to tackle this issue by collectively scrutinising the different results produced by such multiple clustering applications in order to produce a single consensus result.

Another aspect which is not tackled by conventional clustering methods is that some genes may be co-expressed over a subset of samples only, or that some samples might have similar expression values over a subset of genes only. Biclustering methods were proposed in order to address this aspect by mining for *biclusters*. A bicluster is defined by a subset of genes and a subset of samples where this subset of genes is specifically co-expressed over that subset of samples.

However, this vast literature of conventional clustering methods, consensus clustering methods, and biclustering methods, does not attend, or partially does, to other issues and aspects that are raised while analysing gene expression datasets, especially when multiple datasets are considered collectively. This is a list of some of those aspects:

1. The collective cluster analysis of multiple homogeneous or heterogeneous gene expression datasets. For example, what are the subsets of genes that are not only co-expressed in a given dataset, but are also consistently co-expressed over multiple gene expression datasets produced, possibly, under similar or different conditions and biological contexts?

2. The ability to relax conventional clustering constraints by allowing genes to have any of the three eventualities, to be exclusively included in a single cluster, to be simultaneously included in multiple clusters, or to be not included in any cluster at all. This better matches the biological reality that a gene may participate in a single biological process or in multiple biological processes with different groups of genes, or, as most of the genes do given any particular biological context, a gene may be irrelevant to the context and should not be included in any of the clusters.

3. In general terms, those methods require the dataset to be filtered prior to clustering by eliminating those genes which are expected not to be significantly relevant to the study. Gene selection and gene differential expression analysis methods are commonly used for this purpose. This is because such clustering methods assume that all of the genes that reach the clustering step are relevant and will be included in some clusters.

4. Some subsets of genes may show consistent co-expression in some datasets which were generated under specific conditions while being poorly co-expressed in other datasets (Wade, et al., 2006; Nilsson, et al., 2009). The problem of identifying such subsets of genes in an unsupervised manner from multiple datasets is not achieved by any previously proposed method. The closest to this aspect in relevance are biclustering methods, but they require the datasets to be grouped into a single

dataset first, which implies that they should either be homogeneous, or should be thoroughly statistically manipulated to be combinable; also, biclustering methods do not allow the question to be specific about consistent co-expression in a specific subset of samples/conditions with poor co-expression (in contrast to low expression) in another specific subset of samples/conditions.

5. The correct number of clusters (K) included in a dataset is a very common question in this field. It can either be manually set based on *a priori* problem-specific knowledge, or be automatically determined by the method, or be selected from a range of tested and validated values. Although some methods address this issue with different levels of accuracy, any new method which is proposed to address the aforementioned points should also address this key aspect.

6. Another aspect which has been investigated widely in clustering, yet needs more consideration, is the validation of clustering results. Different clustering methods and applications may produce results of different structures and attributes. Therefore, a validation technique which targets a specific type of results may not be a valid choice to validate other types of clustering results. For instance, many clustering validation techniques do not assess the quality of individual clusters; they rather assess whole partitions only. Moreover, when the generated clusters are of significantly different sizes in terms of the numbers of genes included in them, most of the available clustering validation indices tend to favour smaller clusters. In addition to that, any new clustering validation technique has to be validated itself, most likely by using data with known ground-truth. This forms another layer of consideration in this area.

7. The use of synthetic datasets, for which the ground-truth is known to the researcher, is a common practice to test new methods. Various models have been proposed to synthesise datasets which aim at being valid approximations of real expression data (Yeung, et al., 2001; Zhao, et al., 2001; Liu, et al., 2004). Many of these models include parameters to control levels of noise and other aspects. However, synthetic modelling of noise and other deficiencies that naturally occur in real datasets may not be very accurate because the actual level of expression values in real datasets without the noise is not readily available, and the nature of such noise and defects differs, sometimes significantly, between different gene expression measurement technologies.

Taken together, the literature of clustering is rich with many growing paradigms which target the problem of clustering from very different angles. Conventional clustering algorithms (e.g. k-means and hierarchical clustering), consensus clustering algorithms (e.g. graph-based methods and relabelling and voting methods), and biclustering algorithms (e.g. CC and plaid) are three different root paradigms of clustering. However, various aspects have either not been visited by clustering yet or are repeatedly raised, and have to be addressed, whenever a new clustering method or paradigm is proposed (e.g. setting the number of clusters (*K*) and clustering validation).

Chapter 3
Methods

This chapter details the methods and techniques which have been adopted to produce the results presented and discussed in the subsequent chapters. The Sections 3.2, 3.3, 3.5, 3.6, and 3.7 introduce novel methods, namely the Bi-CoPaM method, the UNCLES method, the M-N scatter plots technique, the P-F scatter plots technique, and a method for expression data synthesis based on real measurements. These methods mark the novel contribution of this thesis in designing new computational methods for collective analysis of multiple high-throughput biological datasets. On the other hand, the rest of the sections in this chapter explain some methods and techniques commonly used in the literature of bioinformatics research.

3.1. Gene expression data normalisation

Reliable quantification of gene expression is that which faithfully reflects the true mRNA levels in a given sample. However, much variability exists in the available technologies (e.g. microarrays) which perturbs the measurements so that they are no longer reliable in their raw form. Such variability can be caused by the preparation of the biological sample, fluorescent labelling, specific hybridisation, non-specific hybridisation, scanning, image processing, and other sources (Calza & Pawitan, 2010). Moreover, in most of the microarray datasets, many mRNA samples are taken and measured by multiple microarray chips/slides; these samples can be from different types of tissues (e.g. cancer and normal tissues), at different chronological stages or time points within a biological process, or from different samples contained in different biological conditions. Thus, not only the comparability of intensities of different genes within one slide is questioned, but also the comparability of intensities of a single gene across different slides (samples) is questioned.

Normalisation aims at eliminating these technical variations within a single slide or between multiple slides. This is so that the remaining variations of intensities reliably represent actual biological variations, which are what such experiments desire to measure. The necessity of the normalisation step was reported as one of the six important issues listed by the Minimum Information about a Microarray Experiment (MIAME) protocol (Brazma, et al., 2001).

Numerous normalisation methods have been proposed in the literature. Amongst the most commonly adopted ones are quantile normalisation for one-channel microarrays (Bolstad, et al., 2003) and the locally weighted scatter plot smoothing (lowess) method for two-channel microarrays (Yang, et al., 2002). This application of these methods to these specific types of microarray datasets is recommended by relevant reviews such as the one by Roberts (Roberts, 2008). As these two methods were adopted in many of our sets of experiments presented in this thesis, their details are illustrated in the following two subsections.

3.1.1. Quantile normalisation

This method, which was proposed by Bolstad and colleagues (Bolstad, et al., 2003), has become the most popular method for normalising one-channel microarray datasets (Roberts, 2008; Cahan, et al., 2007). This method is based on the assumption that all of the arrays have a similar signal distribution, which is typical for most of the microarray datasets (Bolstad, et al., 2003; Roberts, 2008; Calza & Pawitan, 2010). Though, for the cases in which different samples are taken from very different tissue types, quantile normalisation should be avoided as the underlying assumption would not be valid anymore (Roberts, 2008; Wang, et al., 2012; Calza & Pawitan, 2010).

The steps of quantile normalisation are summarised as:

(1) Given M arrays/chips of length (number of elements) N, form $\boldsymbol{X}$ of dimension $N \times M$ where each column represents an array.

(2) Sort each column to get $\boldsymbol{X}_{sorted}$.

(3) Take the means across rows of $\boldsymbol{X}_{sorted}$ and assign this mean to each element of that row to get $\boldsymbol{X}'_{sorted}$.

(4) Rearrange the elements in the columns of $\boldsymbol{X}'_{sorted}$ to have the same order as in $\boldsymbol{X}$. This results in the normalised array $\boldsymbol{X}_{normalised}$.

Bolstad and collaborators discussed then that this forces the quantiles to be equal in all of the given arrays, which might not be very accurate at very high intensities. Though, Bolstad and collaborators followed up by mentioning that, because probe-set expression values were calculated by considering multiple probes, this problem did not seem to be a major problem anymore (Bolstad, et al., 2003).

Irizarry and colleagues used some controlled datasets to show the importance of normalisation for oligonucleotide microarrays (Irizarry, et al., 2003a). They used a 'dilution dataset' in which a range of six known proportions of the cRNA taken from human liver tissues were considered; five replicates were taken per each of these six samples and were scanned by five different scanners. Another dataset which they used was a spike-in data in which all genes are expected to be non-differentially expressed except for 20 genes from which fragments with known concentrations were added. These datasets are real (not simulated) datasets but with controls that provide the ground-truth information. Their analysis of the distribution of intensities and log-ratios in these datasets demonstrated the necessity of normalisation and showed that quantile normalisation meets the requirements as needed (Irizarry, et al., 2003a).

Calza and Pawitan (2010) included quantile normalisation in their recent review as one of the most commonly used techniques for normalising one-channel arrays (Calza & Pawitan, 2010). They mentioned that it can deal with non-linear intensity distributions, is simple to understand and implement, and is fast to run. They also mentioned that it is usually performed over the entire set of probes before summarisation as to exploit as much information as possible (Calza & Pawitan, 2010). However, this method did not perform well in some other studies, such as, for instance, a study in which it was compared with some less popular methods over DNA methylation microarray datasets (Adriaens, et al., 2012).

3.1.2. Locally weighted scatter plot smoothing (lowess) normalisation

It was proposed by Yang and colleagues (Yang, et al., 2002) based on the statistical regression model proposed in (Cleveland, 1979). The excellent review of normalisation methods in *Nature Genetics* by Quackenbush also presented this method in a very clear way and showed its strength in normalisation (Quackenbush, 2002). It takes non-linearity into consideration and it is the most commonly used method in the case of within-slide two-channel normalisation (Sievertzon, et al., 2006; Calza & Pawitan, 2010; Smyth & Speed, 2003; Roberts, 2008).

The method was motivated by the obvious bias between the red Cy5-dye and the green Cy3-dye in two-channel microarrays, that is, intensity-dependant effects (Sievertzon, et al., 2006; Irizarry, et al., 2003a). The MA plot, which plots the log-ratio $M_i = \log_2(R_i/G_i)$ versus the abundance $A_i = \log_2 \sqrt{R_i G_i}$ (Dudoit, et al., 2002), shows this bias clearly, where

R_i and G_i are the red and the green intensities of the i^{th} probe / gene. An example of an MA plot is shown in Figure 3.1 (a). The lowess normalisation method aims at correcting this bias (Yang, et al., 2002; Quackenbush, 2002; Xie, et al., 2004).

Assume that $x_i = A_i$ and $y_i = M_i$, an MA plot would plot y versus x. Robust lowess smoother is used for regression in order to estimate $y(x_k)$ which represents the best-fit average based on the experimentally observed values (Quackenbush, 2002; Sievertzon, et al., 2006). While estimating the y value for the point x, a fraction of points closest to the point x can be considered instead of the entire sample set; this fraction of points is called the *span*. If the span is too small, it leads to over-fitting, while if it is too large it leads to inefficient normalisation (Sievertzon, et al., 2006). Spans of about 0.3 (30%) are usually used (Sievertzon, et al., 2006; Quackenbush, 2002; Yang, et al., 2002).

Then, log-ratio correction is applied in a point-by-point manner by subtracting the best-fit estimate from the original log-ratio. This is represented by

$$log_2(T_i') = log_2(T_i) - y(x_i) = log_2(T_i) - log_2(2^{y(x_i)}), \tag{3.1}$$

or

$$\log_2(T_i') = \log_2(T_i \times \frac{1}{2^{y(x_i)}}) = \log_2(\frac{R_i}{G_i} \times \frac{1}{2^{y(x_i)}}). \tag{3.2}$$

In terms of intensity correction, this is equivalent to

$$G_i' = G_i \times 2^{y(x_i)} \text{ and } R_i' = R_i. \tag{3.3}$$

An example of lowess normalised data is shown in Figure 3.1 (b). Lowess normalisation was used successfully by many other studies (Quackenbush, 2002; Xie, et al., 2004; Önskog, et al., 2011).

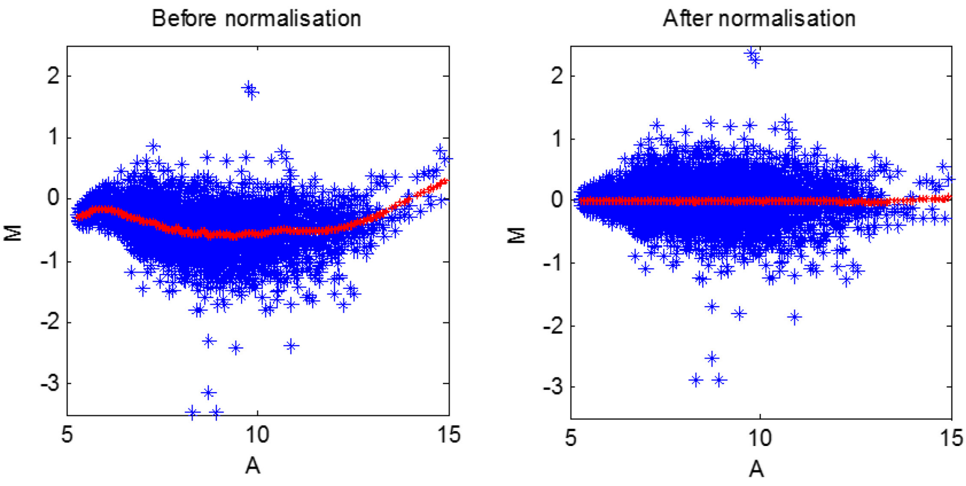

Figure 3.1. MA plots for the yeast genome sample with the NCBI accession number GSM81075 (a) before and (b) after lowess normalisation.

3.2. Bi-CoPaM

The binarisation of consensus partition matrices (Bi-CoPaM) method is a novel contribution of this thesis. Bi-CoPaM is a consensus clustering method which accepts multiple gene expression datasets as an input and produces focused clusters of genes with consistency in co-expression in those datasets as an output (Abu-Jamous, et al., 2013a). It is a condition that all of the considered datasets measure the expression profiles of the same set of genes. However, they can differ in terms of the number of time-points/conditions/features, biological context, year, laboratory, technology, underlying statistical distribution, noise, and the like. In reality, and due to the consensus, that is, collective, nature of the Bi-CoPaM, adopting datasets which are different in such attributes supports the robustness, reliability, and confidence in the results and conclusions.

The Bi-CoPaM method relaxes the conventional binary clustering constraint that each gene (object) has to be exclusively assigned to a single cluster. Bi-CoPaM rather allows each gene to have any of the three eventualities – (i) to be assigned exclusively to a single cluster, (ii) to be assigned simultaneously to multiple clusters, or (iii) not to be assigned to any cluster at all. This is done in a tuneable manner. Such tuneable relaxation in gene assignment leads to tuneable relaxation in the structure of the produced clusters; clusters may be (i) complementary, as conventional binary clustering produces, where they include all genes without overlaps, (ii) wide and overlapping, or (iii) tight and focused while leaving many genes unassigned to any of the clusters.

The application of the Bi-CoPaM method constitutes of four main steps summarised in Figure 3.2 and explained in the following subsections; they are respectively (i) individual partitions' generation, (ii) relabelling, (iii) fuzzy consensus partition matrix (CoPaM) generation, and (iv) binarisation. The following subsections describe those four steps in detail.

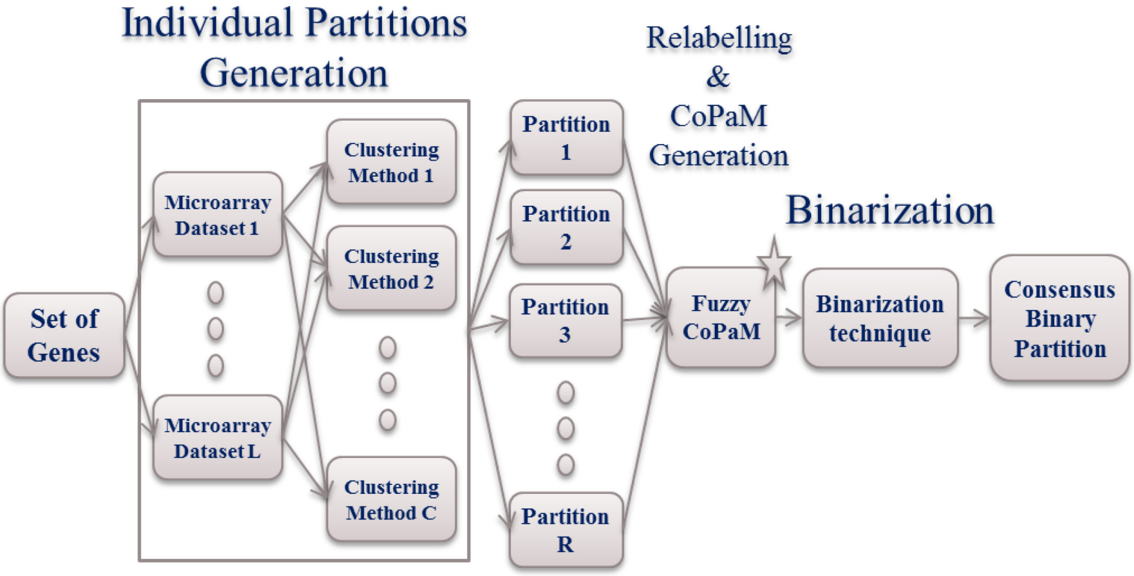

Figure 3.2. Flowchart summarising the steps of the Bi-CoPaM method

The expression profile of a specific set of genes is measured in L different microarray datasets. Each one of those datasets is exposed to gene clustering by each of the C considered clustering methods. The $R = L \times C$ resulting partitions (clustering results) are relabelled and then combined to form a single fuzzy consensus partition matrix (CoPaM) which is then binarised by one of the six proposed binarisation techniques to produce the final binary consensus partition (Abu-Jamous, et al., 2013b).

3.2.1. Individual partition generation

Given L gene expression datasets and C different clustering methods / sets of parameters, clustering the genes in each of the datasets into K clusters by adopting each of the clustering methods generates $R = L \times C$ partitions (clustering results). Let each partition be denoted by a partition matrix $U^r_{K\times N}$, where $r \in [1 \ldots R]$, K is the number of clusters, and N is the number of genes. An element in the rth partition matrix, $u^r_{k,i} \in [0,1]$, represents the membership of the ith gene in the kth cluster, where the membership of zero indicates that this gene does not belong to that cluster, the membership of unity indicates full belongingness of that gene in that cluster, and the membership values between zero and unity indicate proportionate level of belongingness. Crisp clustering, also known as binary clustering, produces binary membership values only, that is, it can only be zero or unity and cannot have any intermediate value. Note that the rows of a partition matrix represent clusters while the columns represent genes.

Conventional partition matrices fulfil the following three conditions:

(1) $$u^r_{k,i} \in [0,1], \qquad \forall r, \forall k, \forall i, \tag{3.4}$$

(2) $$\sum_{k=1}^{K} u^r_{k,i} = 1, \qquad \forall r, \forall i, \tag{3.5}$$

(3) $$0 < \sum_{i=1}^{N} u_{k,i}^{r} < N, \qquad \forall r, \forall k. \tag{3.6}$$

The second condition necessitates that the total membership of any given gene in all of the clusters should be unity because such membership values represent the probability of belongingness.

3.2.2. Relabelling

Because clustering is unsupervised, the *k*th cluster in a given partition may not correspond to the *k*th cluster in other partitions. Therefore, it is essential to reorder the clusters in all of the partitions such that they are aligned. Thereafter, the *k*th cluster in a given relabelled partition corresponds to the *k*th cluster in each one of the other partitions.

Depending on the objectives of the application under consideration, the priorities of relabelling may differ. In some applications, all of the resulting clusters are of interest to the investigator, and the priority in this case is to optimise the overall relabelling accuracy. On the other hand, many applications aim at producing few focused high-quality clusters while ignoring the rest of the clusters; in this latter case, the priority is to maximise the quality of the promising clusters while paying no attention to poor clusters. Thus, two relabelling techniques are described here, namely min-max (Abu-Jamous, et al., 2013a; Abu-Jamous, et al., 2013b) and min-min (Abu-Jamous, et al., 2013d; Abu-Jamous, et al., 2014a; Abu-Jamous, et al., 2015b), which respectively adopt the two different aforementioned priorities.

Consider reordering the clusters in a partition U aiming to align them with the clusters in a reference partition U^{ref}. The first step in either technique is to construct a K×K pairwise distance matrix whose rows represent the clusters in U, columns represent the clusters in U^{ref}, and elements represent pairwise distance/dissimilarity values between corresponding clusters. Two sample pairwise distance matrices, for the K values four and ten, are shown in Figure 3.3 (a) and (b) respectively.

The second step in both relabelling techniques is to find the minimum value in each of the columns in the matrix; these minima are shown in the last rows of the matrices in Figure 3.3.

The two techniques diverge at the third step; min-max identifies the row and the column whose intersection includes the maximum of the afore-calculated minima (shaded in Figure 3.3 (a)) while the min-min technique identifies those which own the minimum of the minima (shaded in Figure 3.3 (b)).

(a)

	1	2	3	4
1	2	7	6	7
2	5	0	3	4
3	9	5	7	1
4	6	1	8	6
Min	**2**	**0**	**3**	**1**

(b)

	1	2	3	4	5	6	7	8	9	10
1	5	7	9	9	10	7	8	6	11	10
2	0	13	8	6	8	12	7	9	8	5
3	9	1	6	8	9	6	11	7	17	8
4	4	5	8	13	16	7	9	9	8	9
5	7	9	15	6	9	9	6	12	18	7
6	4	5	10	4	11	15	8	8	9	7
7	10	11	9	6	9	18	16	1	12	6
8	6	5	8	9	15	9	8	7	9	12
9	8	6	13	2	11	7	7	6	10	8
10	12	5	7	7	9	9	10	9	7	9
Min	**0**	**1**	**6**	**2**	**8**	**6**	**6**	**1**	**7**	**5**

Figure 3.3. Min-max and min-min cluster dissimilarity matrices

This is a demonstration of the first iteration of relabelling by the (a) min-max and the (b) min-min techniques. The last row of the matrix shows the minima of the columns, and the highlighted cell therein is the (a) maximum or the (b) minimum of those minima.

The fourth step is to map the clusters in U and U^{ref} which are respectively represented by the identified row and column to each other, and then to remove those row and column from the matrix. After that, the minimum/maximum of the column's minima in the reduced matrix is identified leading to mapping a second pair of clusters from U and U^{ref}. These steps are repeated until all clusters from U are mapped to their corresponding clusters in U^{ref}.

Let us discuss how each of these two techniques meets its priorities. If we apply min-min instead of min-max to the example in Figure 3.3 (a), the cluster represented by the second row will be mapped to the cluster represented by the second column because the distance between them is the perfectly minimum distance of 0.0. By preserving the second row early in this iterative process as such, the third column will have no descent partner to be mapped to as the next closest partner, which is the first row, is very distant from it with a large distance of 6.0 units. On the other hand, the min-max technique gives the column containing the maximum of the minima priority in assignment to ensure that it will not be eventually assigned to a very distant partner. In this case, while the second column will not be paired with its closest row, which is the second row, it can still be paired with the forth row with an acceptable distance of 1.0. Although assigning such cluster to its second closest cluster might not always result in acceptable distances as in the given example, the min-max approach still show more fairness in the distribution of care over the different clusters by prioritising those that are under a higher risk of not finding acceptable partners if delayed.

Moving our focus to the example in Figure 3.3 (b), one can notice that there are four pairs of clusters (row-column pairs) which have very low distance values, namely the (row, column) pairs (2, 1), (3, 2), (7, 8), and (9, 4), with the distances of 0.0, 1.0, 1.0, and 2.0, respectively. In contrast, the rest of the six rows and six columns are very far from each other with distances of 5.0 or higher. This indicates that the two partitions have consensus over four clusters and disagreement over six clusters. Being generated by using different methods or under different biological conditions, their agreement on some clusters is a hint that those are genuine clusters with relatively higher quality and are robustly identified under different experimental setups or biological conditions. Therefore, the min-min approach aims at ensuring that these four pairs of clusters are correctly associated while ignoring the rest of the clusters. If the min-max approach had been used here instead, the second row would have been paired with the fifth column in the first iteration depriving it from being paired with its genuine match, which is the first column. This behaviour would have been due to prioritising the poor cluster represented by the fifth column, which is seemingly not a cluster of interest in this application.

Because most of our applications consider large datasets with the objective of producing focused clusters that do not include all of the input objects (e.g. genes), we tend to apply the Bi-CoPaM method with relatively high values of *K* while adopting the min-min relabelling technique.

3.2.3. Fuzzy CoPaM generation

Once all of the *R* partitions $\{U^1 \ldots U^R\}$ have been relabelled, a fuzzy consensus partition matrix (CoPaM) is generated by averaging the *R* partition matrices in an element-by-element fashion. However, our adopted implementation of this considers and iterative approach in which relabelling and fuzzy CoPaM generation are done in a partition-by-partition manner. Let the function $\text{Relabel}(U^r, U^{ref})$ perform relabelling to the clusters in the partition U^r while considering the partition U^{ref} as a reference. The generation of the final fuzzy CoPaM U^* is therefore performed according to the following algorithm:

```
$U^{int(1)} = U^1$
For $r = 2$ to $R$
    $\hat{U}^r = \text{Relabel}(U^r, U^{int(r-1)})$
    $U^{int(r)} = \frac{1}{r}\sum_{\acute{r}=1}^{r} \hat{U}^{\acute{r}} = \frac{1}{r}\hat{U}^r + \frac{r-1}{r} U^{int(r-1)}$
$U^* = U^{int(R)}$
```

Here, the final fuzzy CoPaM U^* is produced through the accumulative evolution of intermediate fuzzy CoPaM (ICoPaM) matrices, where $U^{int(r)}$ is an ICoPaM at the *r*th iteration. The first ICoPaM, $U^{int(1)}$, is set to the first individual partition, U^1. Then, the iterative mode of the algorithm starts. In every iteration (r), the next individual partition to be relabelled and merged, U^r, is relabelled by considering the most recent ICoPaM, $U^{int(r-1)}$, as a reference. We denote the relabelled version of this partition as $\widehat{U}^r$. The ICoPaM is then updated by weighted element-by-element averaging as shown in the algorithm above. After all of the R partitions have been relabelled and merged with the ICoPaM, the last ICoPaM, $U^{int(R)}$, is considered as the final fuzzy CoPaM, U^*.

3.2.4. Binarisation

The fuzzy CoPaM matrix includes membership values for all of the considered genes (data objects) in each of the *K* clusters. A membership value of unity indicates full belongingness of the given gene to the given cluster, a zero membership indicates absolutely no belongingness, and a fractional membership value indicates respective partial belongingness. If all of the *R* individual partitions have consensually assigned a given gene to the same cluster, the membership of this gene in that cluster will be unity while being nil in all of the other clusters. However, if the individual partitions have disagreement in assignment of that gene, its membership value is distributed over all of the clusters in which some partitions included it. Indeed, the membership of the gene in any of these partitions is set to be proportionate with the number of individual partitions which assigned it to it.

The next step is to binarise the membership values of the genes in the fuzzy CoPaM. It is clear that a given gene should be assigned to the cluster to which it has been assigned consensually by all partitions. However, in the case of disagreement, should this gene be simultaneously assigned to all of the clusters to which some partitions assign it, or should it be left unassigned from all of the clusters given the dispute? Below are six proposed binarisation techniques addressing this issue in different ways.

3.2.4.1. Maximum value binarisation (MVB)

MVB assigns the gene exclusively to the cluster in which it has its largest membership value, and therefore it generates complementary clusters. Let the resulting binary CoPaM be B^* with *K* rows and *N* columns, where $b^*_{k,i} \in \{0, 1\}$ is an element in this matrix representing the binary membership of the i^{th} gene in the k^{th} cluster. Similarly, the corresponding fuzzy CoPaM is U^* with the elements $u^*_{k,i} \in [0, 1]$. Given that, the MVB technique can be expressed as:

$$b_{k,i}^{*} = \begin{cases} 1, & u_{k,i}^{*} = \max\limits_{1 \le j \le K} u_{j,i}^{*} \\ 0, & otherwise \end{cases} \tag{3.7}$$

3.2.4.2. Top binarisation (TB)

The TB technique moves from the MVB technique towards producing wider clusters. This is done by assigning the given gene to multiple clusters simultaneously if its membership values in them are not farer than the value of the tuning parameter δ below its maximum membership value. The TB technique is expressed as:

$$b_{k,i}^{*} = \begin{cases} 1, & u_{k,i}^{*} \ge \max\limits_{1 \le j \le K} u_{j,i}^{*} - \delta \\ 0, & otherwise \end{cases} \tag{3.8}$$

3.2.4.3. Difference threshold binarisation (DTB)

In contrast to TB, and in a symmetric manner, the DTB technique moves from the MVB technique towards producing tighter clusters. This is performed by assigning a gene to the cluster in which it has its maximum membership value only if this value is far from the closest competitive cluster at least by the value of the tuning parameter δ; it is not assigned to any of the clusters otherwise. The DTB technique is expressed as:

$$b_{k,i}^{*} = \begin{cases} 1, & u_{k,i}^{*} \ge \max\limits_{\substack{1 \le j \le K, \\ j \ne k}} u_{j,i}^{*} + \delta \\ 0, & otherwise \end{cases} \tag{3.9}$$

We group the aforementioned three techniques in a track of binarisation and we name it as the TB-MVB-DTB track (Figure 3.4). When δ is equal to zero in TB or DTB, they become identical to the MVB technique. When δ increases, TB or DTB start widening or tightening the clusters, respectively. The maximum value of δ is unity. When this value is reached, the TB technique reaches the extreme case of wide clusters in which each one of the clusters includes all of the genes. Also, at the δ value of unity, the DTB produces the tightest clusters in which a gene is assigned to a cluster only if its fuzzy membership value is equal to unity in that cluster and is equal to zero in all of the other clusters, i.e. if all of the R individual partitions have consensually assigned that gene to that cluster. Although DTB may generate many empty clusters at δ = 1.0, this result would not be trivial if some of the clusters still preserved some genes up to this tightest level, as opposed to the TB technique's results at such δ value.

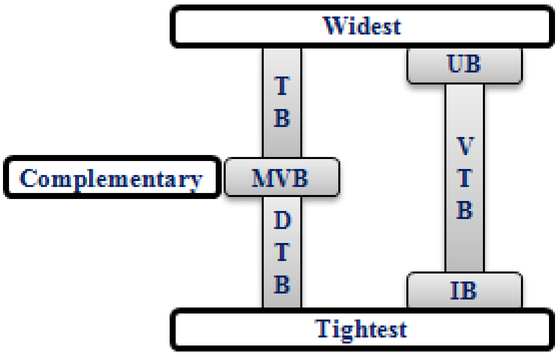

Figure 3.4. Binarisation tracks

The (left) TB-MVB-DTB track and the (right) UB-VTB-IB track can produce clusters which range from very wide to very tight. Also, the MVB technique of the TB-MVB-DTB track can also provide complementary clusters.

3.2.4.4. Union binarisation (UB)

UB assigns each gene to all of the clusters in which it has non-zero fuzzy membership values, i.e. to all of the clusters in which at least one of the *R* individual partitions has assigned it. This generates wide and overlapping clusters. UB is expressed as:

$$b_{k,i}^{*} = \begin{cases} 1, & u_{k,i}^{*} > 0 \\ 0, & otherwise \end{cases} \tag{3.10}$$

3.2.4.5. Intersection binarisation (IB)

Contrary to the UB technique, IB assigns a gene to a cluster only if all of the *R* individual partitions have consensually assigned that gene to it, i.e. if its fuzzy membership value in it is unity while being zero elsewhere. This technique generates the tightest and most focused clusters, and is equivalent to the tightest clusters generated by the TB-MVB-DTB track, namely by the DTB technique at δ = 1.0. The IB technique is expressed as:

$$b_{k,i}^{*} = \begin{cases} 1, & u_{k,i}^{*} = 1.0 \\ 0, & otherwise \end{cases} \tag{3.11}$$

3.2.4.6. Value threshold binarisation (VTB)

VTB assigns a gene to a cluster if its membership in it is larger than or equal to the value of the tuning parameter α. The VTB technique is expressed as:

$$b_{k,i}^{*} = \begin{cases} 1, & u_{k,i}^{*} \geq \alpha \\ 0, & otherwise \end{cases} \tag{3.12}$$

We group the latter three techniques into a second track of binarisation, namely the UB-VTB-IB track (Figure 3.4). When α is equal to zero, the VTB assigns each gene to all of the clusters, which is a trivial and useless result. At α = ε, where ε is an arbitrarily small real positive number, the VTB technique becomes identical to the UB technique, and at α = 1.0, it becomes identical to the IB technique. As the value of α increases from ε to unity, the clusters are tightened.

The main semantic difference between the two tracks is that the TB-MVB-DTB track considers comparative criteria based on the competition amongst the clusters over the genes. On the other hand, the UB-VTB-IB track considers the absolute membership values of genes in individual clusters. Nonetheless, the fact that fuzzy membership values are normalised such that their sum for a single gene over all of the clusters is unity implies that those values implicitly consider certain levels of competition between the clusters, even when considered by the UB-VTB-IB track. However, the TB-MVB-DTB track is more explicit in basing gene-cluster assignments on such competitions.

3.3. UNCLES

The objective of the Bi-CoPaM method, while using tightening binarisation techniques, can be summarised as: it aims at identifying the subsets of genes (or any other types of objects) which are consistently co-expressed (highly correlated in profiles) over all of the given datasets and when analysed by all of the adopted clustering methods and setups. Indeed, binarisation parameters control how much tolerance is accepted.

However, other research questions can be answered by other ways of unifying individual clustering results. We therefore propose a more general paradigm of multiple-dataset mining which we call the *unification of clustering results from multiple datasets using external specifications* (UNCLES) (Abu-Jamous, et al., 2015c). Bi-CoPaM serves as a special case of the UNCLES method as it unifies clustering results from multiple datasets while considering consistency in co-expression over all/most of the datasets as external specifications. We name this type of external specifications as “type A”.

Here we propose another type of external specifications, labelled “type B”, which aims at identifying the subsets of genes which are consistently co-expressed in a subset of datasets (S^+) while being poorly consistently co-expressed in another subset of datasets (S^-).

To apply UNCLES type B, the type A (Bi-CoPaM) is applied to each of the two subsets of datasets S^+ and S^- separately while adopting DTB binarisation with the δ values of δ^+ and δ^- respectively. After that, the genes that are included in the results of processing the S^+ datasets and not included in the results of processing S^- datasets, indeed after binarisation, are included in the final result. Therefore, the UNCLES type B method utilises a pair of parameters (δ^+, δ^-) in order to achieve its results. The parameter δ^+ controls how well co-expressed the genes should be in the S^+ datasets to be included in the final result, while the parameter δ^- controls how well co-expressed the genes should be in the S^- datasets to be excluded from the final result. Note that at the pair (δ^+, 0) empty clusters are generated because at $\delta^- = 0$ all of the genes will be excluded from the final result.

3.4. Mean squared error (MSE) metric

The mean squared error (MSE) metric has been used in many studies to evaluate the quality of clusters by quantifying the dispersion within the cluster (Lam & Tsang, 2012; Zhu, et al., 2012). The normalised *per gene* MSE measure for the k^{th} cluster is defined as:

$$\mathrm{MSE}_{\mathrm{cluster(k)}} = \frac{1}{M \cdot N_k} \sum_{\boldsymbol{x}_i \in \boldsymbol{C}_k} \|\boldsymbol{x}_i - \boldsymbol{z}_k\|^2, \tag{3.13}$$

where M is the number of dimensions (time-points) in the dataset, N_k is the number of genes in the k^{th} cluster, $\boldsymbol{C}_k$ is the set of genetic expression profiles $\{\boldsymbol{x}_i\}$ for the genes in the k^{th} cluster, and $\boldsymbol{z}_k$ is the mean expression profile for the genes in the k^{th} cluster.

If multiple datasets were used in clustering, genes' profiles and the clusters centroids will vary from one dataset to another for the same partition. In this case, the MSE metric can be calculated multiple times for each dataset and then averaged over them.

3.5. M-N scatter plots

UNCLES types A and B generate clusters with varying levels of wideness / tightness depending on the values of the tuning parameters fed. Such clusters largely vary in size, which significantly affects the validity of known validation techniques rendering them unreliable in this particular context. Therefore, we propose a customised and sophisticated cluster evaluation and validation technique, based on our proposed *M-N scatter plots*, where *M* refers to a modified version of the MSE metric (Section 3.4), and *N* refers to the number of genes included in the cluster, or more specifically, the logarithm of that number (Abu-Jamous, et al., 2015c). The objective of the M-N scatter plots technique is to maximise the size of the cluster while minimising the mean square error. This multi-objective technique suites the tuneable nature of the clusters generated by the UNCLES method.

The M-N scatter plot is a 2-D plot on which the clusters are scattered, where the horizontal axis represents the MSE-based metric (MSE*) defined below, and the vertical axis represents the 10-based logarithm of the number of genes included in the cluster. The clusters closer to the top-left corner of this plot, after scaling each axis to have a unity length, are those that include more genes while maintaining lower MSE* values, and are considered as better clusters based on this technique.

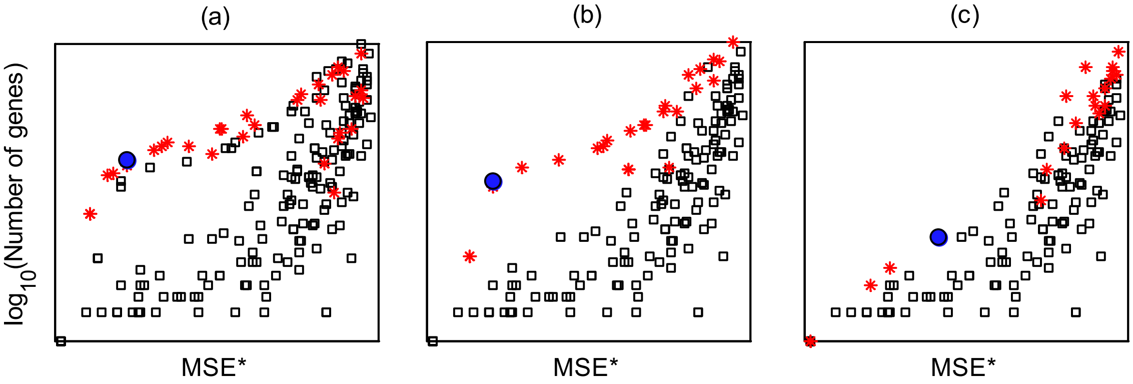

Figure 3.5. Three iterations of cluster selection based on M-N scatter plots

Figure 3.5 (a) shows a sample M-N scatter plot. Each point on this plot, regardless of its shape and colour, represents a single non-empty cluster. The one closest to the top left corner in Euclidean distance is marked with a big solid circle, and is selected as the best cluster. The stars represent all of those clusters which have significant overlap in terms of gene content with the selected best cluster. Here we consider any overlap, even with a single gene, as significant. Therefore, we can consider the clusters represented by stars as other versions of that best cluster. The clusters represented by squares are all of the rest of the clusters. Before selecting the second best cluster, those clusters with similarity to the first best cluster are removed from the plot, and the resulting updated M-N plot, in this case, is shown in Figure 3.5 (b). The same step is repeated iteratively to select many clusters until the M-N plot has no more clusters or a specific termination criterion is met. For example, after twenty iterations, the M-N plot in Figure 3.5 (a) becomes totally empty; the first three iterations are shown in Figure 3.5. The selected twenty clusters are ordered in quality from the closest to the top-left corner to the farthest, and those twenty distances are shown in Figure 3.6. Although twenty clusters are found in this example, the grace of having the clusters ordered allows selecting few top clusters only. As in Figure 3.6, there is a large gap in distances between the second and the third clusters, which would lead the researcher to restrict oneself to the first two clusters only for further biological analysis.

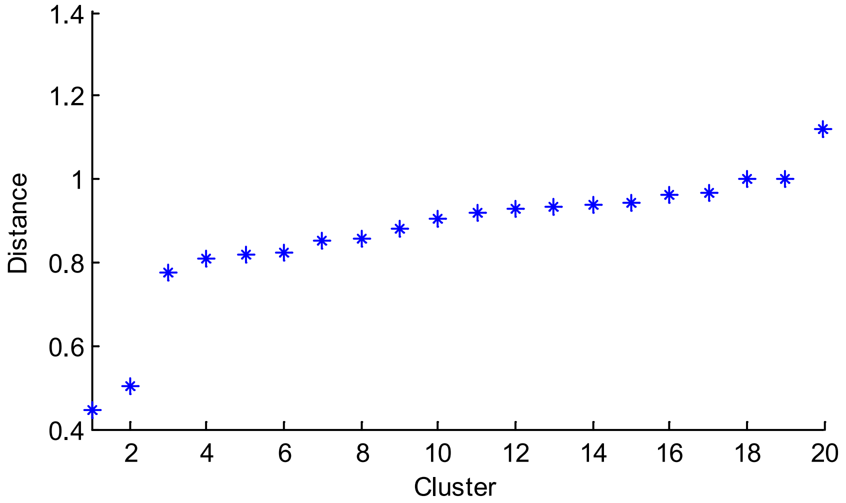

Figure 3.6. Distances of the twenty ordered clusters selected by the M-N plots from the top-left corners of those plots

The MSE-related metric (MSE*) is defined differently for UNCLE types A and B to meet their different objectives. For type A, the MSE* metric is the average of the MSE values for the considered cluster across all of the given datasets, where for type B, it is the signed difference between the average of the MSE values across the positive subset of datasets (S^+) and that average across the negative subset of datasets (S^-).

3.6. F-P scatter plots

Alongside the unsupervised M-N scatter plots described above, we propose a supervised cluster validation technique based on our proposed F-P scatter plots (Abu-Jamous, et al., 2015c). Supervised validation in this context is that which is based on the available ground-truth (external validation). On the other hand, unsupervised validation is based on the dispersion and size of the clusters themselves (internal validation).

In a similar fashion to the M-N scatter plots, the clusters are scattered on a 2-D plot whose horizontal axis represents the false positive rate (FPR or *F*), and whose vertical axis represents a scaled version of p-values (*P*). The FPR (*F*) of a cluster is the ratio between the number of genes that are wrongly included in the clusters as per the ground truth (false positives) and the total number of genes in the cluster. This ranges between zero, when no false positives are included in the clusters, and unity, when all of the genes in the cluster are false positives.

The scaled p-value (*P*) is based on a p-value calculated by modelling the problem with a hypergeometric distribution. Let there be *N* genes in the complete considered dataset, where *M* of them belong to the considered ground-truth cluster. If the cluster being validated includes *n* genes, out of which *m* genes belong to the ground-truth cluster, the true positives will be m, the false positives will be $n - m$, the false negatives will be $M - m$, and the true negatives will be $N - n - M + m$. The p-value is therefore the probability of obtaining m or more true positives in a cluster of n genes randomly selected from a pool of N genes which includes M positives. This is mathematically expressed as:

$$p - value = \sum_{j=n}^{N} \left(\frac{M!}{j!\,(M-j)!}\right) \times \left(\frac{n}{N}\right)^{j} \times \left(\frac{N-n}{N}\right)^{M-j} \qquad (3.14)$$

This logarithm of this p-value is then scaled by the logarithm of the best theoretically possible p-value; this is expressed as:

$$scaled\; p - value\; (P) = \frac{\log(p - value)}{\log\left({}^{N}/_{M}\right)^{N}} \qquad (3.15)$$

where $\left({}^{N}/_{M}\right)^{N}$ is the best theoretically possible p-value resulting from producing the perfect cluster which capture all the M genes in the ground-truth cluster and only the M genes in the ground-truth cluster. In this case, $n = M = m$, leading to the p-value of $\left({}^{N}/_{M}\right)^{N}$ when substituted in Equation (3.14). The scaled p-value (P) ranges from zero, when the cluster include no true positives at all, to unity, when the cluster is the perfect cluster.

Taken together, better clusters, that is, the clusters which better match the ground-truth, are those which maximise P while minimising F.

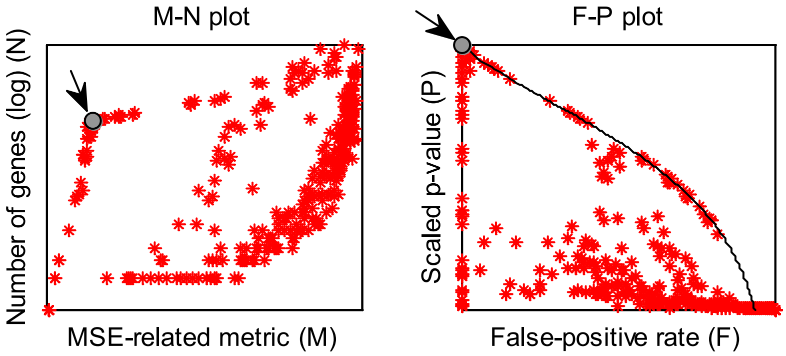

Figure 3.7. Sample M-N and F-P plots for the same set of clusters

In both plots, the cluster shown as a solid grey circle and point at by an arrow is the one closest to the top-left corner. The best cluster identified by the M-N plot was found to be the same as the one identified by the F-P plot.

A sample F-P plot is shown next to a corresponding M-N plot in Figure 3.7. Both plots scatter the same set of clusters, and both plots were found to identify the same cluster as the best one, that is, the one closest to the top-left corner. This best cluster is distinguishably marked by a solid grey circle and is pointed at by an arrow. The black continuous curve in the F-P plot marks the maximum theoretically possible P value at any given F value. In reality, the clusters lying on this curve are those with zero false-negatives, that is, they include all of the true-positives (the *M* genes), in addition to some false-positives, which might be as few as zero (the top left end of the curve), or as many as all of the genes outside the ground truth cluster (the bottom right end of the curve).

3.7. Expression data synthesis based on real data

In order to validate new proposed clustering methods, datasets with known ground-truth are needed to verify if the proposed method can produce what is expected from it or not. Because real datasets tend not to be fully understood and consequently there is not well-defined ground-truth for them in the context of clustering, it is common to synthesise datasets according to a pre-defined ground-truth, expose them to analysis by the proposed method, and then compare the results with the known ground-truth.

Rather than synthesising gene expression profiles based on mathematical models which approximate real expression (Yeung, et al., 2001; Zhao, et al., 2001; Liu, et al., 2004), we propose an approach to form a set of datasets with profiles from real datasets but in a controlled manner in order to have the ground truth available (Abu-Jamous, et al., 2015c). For this to be achieved, some real gene expression datasets which have been exposed to cluster analysis by previous studies were selected. These datasets' GEO accession numbers are GSE18057 (Fujii, et al., 2010), GSE10124 (Hayata, et al., 2009), GSE12736 (Limb, et al., 2009), and GSE9386 (Liu, et al., 2008), and belong to the species *Oryza sativa* (Asian rice), *Xenopus laevis* (African clawed frog), *Homo sapiens* (human), and *Zea mays* (maize), respectively. The respective numbers of samples in the datasets are 36, 6, 16, and 24.

Based on those four datasets we have formed six datasets, named as P1, P2, P3, N1, N2, and N3. P1 and P2 are respectively based on the first eighteen and the last eighteen samples of GSE18057, P3 is based on GSE10124, N1 is based on GSE12736, and N2 and N3 are based on the first and the last twelve samples of GSE9386 respectively.

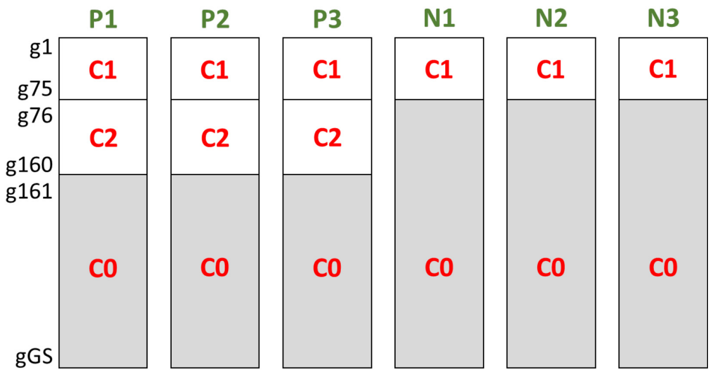

Figure 3.8. Structure of the six synthetic datasets formed based on real expression data measurements

The gene names / probe identifiers of the original datasets were omitted and the artificial gene names g1 to g*GS* were used instead, where *GS* is the artificial genome size (the total number of genes in the dataset). Here we claim that the *i*th gene (g*i*) in each of the six synthetic datasets refers to the same artificial gene. Therefore, the six datasets are seen as datasets which measure the expression profiles of the same set of genes (g1 to g*GS*) but under different conditions. In each of the six datasets, the artificial genes g1 to g75 were selected from one of the clusters identified in the relevant study, i.e. the profiles of those 75 genes in each of the datasets were previously confirmed to be co-expressed in the literature; these genes have been labelled as the cluster C1 (Figure 3.8 and Figure 3.9). The 85 genes g76 to g160 were selected in the same way but only for the positive datasets P1, P2, and P3, and have been labelled as the cluster C2 (Figure 3.8 and Figure 3.9). The rest of the genome, i.e. g161 to g*GS* in P1, P2, and P3, and g76 to g*GS* in N1, N2, and N3, was randomly selected from those genes which were considered as poorly co-expressed in the relevant studies due to being non-differentially expressed; these have been labelled as C0 (Figure 3.8). We have generated five sets of such dataset with the genome sizes (*GS*) of 1200, 2000, 3000, 5000, and 7000 genes respectively; each of these sets includes six datasets as described, with the same C1 and C2 genes shown in Figure 3.9.

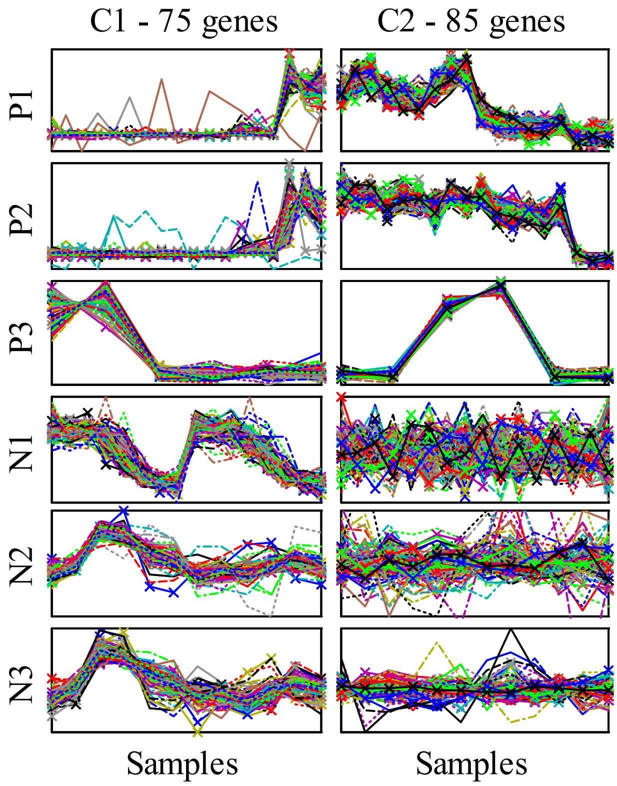

Figure 3.9. Profiles of the ground-truth clusters C1 and C2 in each of the six datasets

The genes in C1 are consistently co-expressed in all of the six datasets, meeting the specifications of UNCLES type A, while the genes in C2 are consistently co-expressed in the positive datasets only while being poorly co-expressed in the negative datasets, meeting the specifications of UNCLES type B.

3.8. GO term analysis

The Gene Ontology (GO) Consortium has taken the responsibility of unifying and standardising gene attribute association. In a more explicit statement, the Consortium has identified three families of attributes with which a gene may be associated, namely the *biological process* in which the gene's product participates, the *molecular function* which the gene's product undertakes, and the *cellular component* in which the gene's product localise (The Gene Ontology Consortium, 2000; The Gene Ontology Consortium, 2013). Each *GO term* referring to a process, function, or component in this database has a unique *GO term identifier* (GOID) starting with the prefix 'GO:' followed by seven numerical digits, for example, the identifier 'GO:0051301' refers to the biological process term 'cell division'.

In a regulated and actively revised and updated manner, the association of genes with their corresponding GO terms is done based on the existence of sufficient evidence supporting this association from the studies in the literature. As new studies emerge, gene-term associations are updated. Indeed, the general case is that any single process, function, or component can be performed, undertaken, or host many genes. For example, 'cell division' is performed by the collaboration of many genes, and surely all of them are associated with it. Similarly, any single gene, generally speaking, may participate in many processes, be involved in different functions, and localise in various cellular components. Therefore, the relation between genes and GO terms is the well-know *many-to-many* relation.

GO term enrichment analysis aims at the identification of the GO terms which are highly represented in a given cluster of genes, that is, the terms with which the genes in the given cluster are significantly associated. Let the complete set of genes (the genome) include N genes, M of which are associated with a given GO term 'X' as per the GO Consortium databases; from those N genes, we have fetched a cluster of n genes, m of which are associated with the GO term 'X'; the question is: "is this cluster highly enriched with the GO term 'X'?" We answer this question by calculating the p-value of such observation based on the hypergeometric distribution in an analogous way to what has been discussion in the previous section, Section 3.6. If the p-value is very small, for example < 0.001, we consider that this cluster is indeed highly enriched with this term.

Given a single cluster, we repeat that question while considering each one of the available terms in the database. Consequently, we obtain a Table of GO terms and their corresponding p-values, usually in an ascending order based on the p-values. Then, a threshold is adopted to filter out the terms whose p-values are higher than a given threshold (e.g. 0.001), to maintain a list of terms which are highly enriched in the cluster.

An important note should be raised here; because that test/question is asked repetitively over a large number of terms, a proportion of those tests is expected to pass due to chance, resulting in a significant number of false-positives. The p-values are therefore corrected by one of different techniques in order to compensate for this problem, which is referred to as the *multiple-hypothesis testing* problem.

GO term analysis helps in finding the biological context of a given cluster of genes by noticing those GO terms with which the cluster is enriched. Also, many genes in those clusters are expected not to be associated any GO terms yet, due to human's incomplete knowledge herein. Such unknown or poorly understood genes, which appear in clusters of many known genes, represent candidates for further biological investigation in light of the biological context in which this GO analysis places the cluster.

Various freely-available tools are available online while holding up-to-date databases of term associations. Some of those tools are generic to various species, like the Princeton University tool at http://go.princeton.edu/cgi-bin/GOTermFinder, and some of them are species-specific like the *Saccharomyces Genome Database (SGD)* tool at http://www.yeastgenome.org/cgi-bin/GO/goTermFinder.pl. Using these tools is intuitive as they accept a list of input genes and few parameters; then they provide an ordered list of GO terms with their associated p-values.

3.9. Upstream sequence analysis

The expression of genes is regulated positively and negatively by proteins known as transcription factors (TFs). When a TF regulates a gene, it recognises and binds to a specific short motifs (DNA sequences) found upstream of the DNA sequence of that target gene. Different TFs have different binding sites, that is, the sequences of the motifs which they recognise are different.

It is known that a single TF may regulate many genes because their upstream sequences include its binding site. In some cases, this TF would be semantically regulating a complete biological process (like cell division) by regulating the tens of genes which participate in it. Moreover, it is also known that many binding sites are collaboratively bound by multiple TFs forming a TF complex.

Identifying a cluster of genes which are *co-expressed*, that is, their expression increases and decreases correlatively, implies that it is likely that they are also *co-regulated*, that is, regulated by the same TF gene or TF complex machinery. Although they might be co-expressed for other reasons, co-expression is still enough for the subset of genes to be a strong candidate for co-regulation investigation.

Upstream sequence analysis mines the upstream sequences of the genes included in a cluster for those motifs that are significantly and repetitively found therein. After that, those motifs are compared with libraries of binding sites for known TFs. If the upstream sequences of a cluster are highly enriched with a motif that significantly matches a known TF's binding site, we may hypothesise that those genes are co-regulated by that TF.

Many tools are available to perform this analysis like the MEME (Multiple Em for Motif Elucitation) suite at http://meme.nbcr.net/.

Chapter 4
Methods Assessment and Validation

Before utilising the proposed methods, they need to be assessed and validated in order to have it on good authority that they can be reliably used in extracting relevant biological findings. Here we present some experiments that we have conducted in order to validate both types of the proposed UNCLES method as well as the proposed M-N scatter plots cluster validation and selection technique. This is achieved by the analysis of the sets of synthetic datasets generated based on our proposed approach and explained in Section 3.7. The Bi-CoPaM method is practically equivalent to UNCLES type A, and therefore is validated as its equivalent is validated. Extra detailed validation for the Bi-CoPaM method can be found in (Abu-Jamous, et al., 2013a).

4.1. Experimental setup

UNCLES has been applied to each of the five sets of synthetic datasets that were generated with five different genome sizes (*GS*). Each of those sets of datasets has been considered with all of the numbers of clusters (K) of 4, 8, 12, 16, 20, and 25 clusters. Both types of external specifications, types A and B, have been considered. Type A aims at identifying the subsets of genes consistently co-expressed over all of the datasets, and type B aims at identifying the subsets of genes specifically consistently co-expressed in the positive set of datasets P1, P2, and P3, while being poorly consistently co-expressed in the negative set of datasets N1, N2, and N3. The used DTB δ values for UNCLES type A were zero to unity with steps of 0.1, and the (δ^{+}, δ^{-}) pair values for UNCLES type B were all possible pairs while ranging each of the δ values from zero to unity with steps of 0.1.

4.2. UNCLES and M-N plots validation

The perfect result of 100% specificity and 100% sensitivity would be obtained if the cluster C1 is discovered by type A of UNCLES, and the cluster C2 is discovered by type B. For any single set of datasets (for a specific genome size (*GS*)), there are 935 individual clusters generated by type A by considering all of the used *K* and δ values, and there are 10,285 individual clusters generated by type B by considering all of the used *K*, δ^+, and δ^- values.

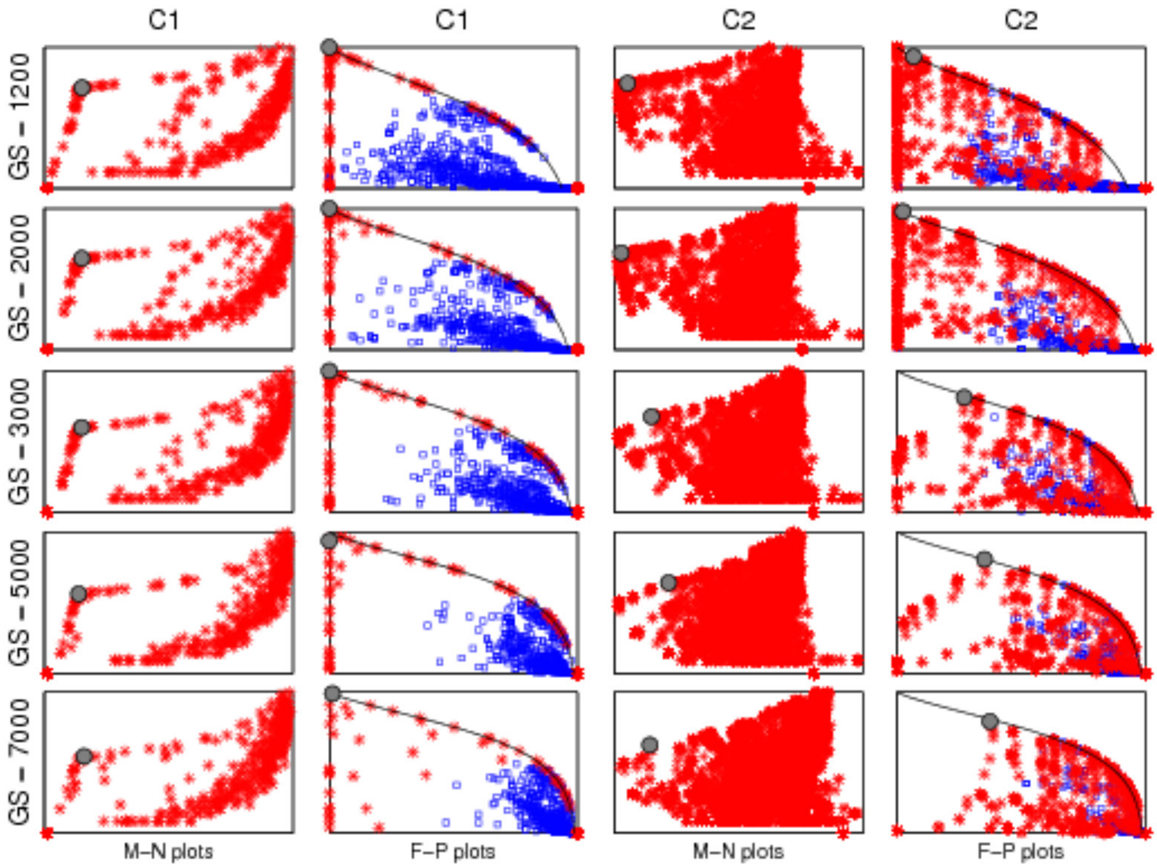

Figure 4.1. M-N and F-P scatter plots of the synthetic data clusters C1 and C2 generated by UNCLES and other methods

The selected clusters in the M-N plots are marked by solid grey circles, and their corresponding points in the F-P plots are marked by solid grey circles as well. The red stars represent all of the other clusters generated by the UNCLES method while the blue squares in the F-P plots represent the clusters generated by the other four clusters methods with which we compare UNCLES (discussed below in Section 4.3).

M-N scatter plots for each of the considered genome sizes for UNCLES types A and B are shown in Figure 4.1 (the first and the third columns) while marking the selected best cluster in each case with a solid grey circle. To validate the usage of those novel M-N scatter plots in validation, we have also shown the ground-truth-based F-P scatter plots for each of these cases in the second and the fourth columns (Figure 4.1). The selected clusters based on the M-N plots are also marked on the F-P plots with solid grey circles.

The first, most relevant and most interesting observation is that in both types of external specifications A and B, that is, for clusters C1 and C2, and for all of the considered genome sizes (*GS*), the clusters selected based on the ground-truth-independent approach scored the best (M-N plots), or very close to the best, scores in the ground-truth-dependent approach (F-P plots) (Figure 4.1). This not only proves the ability of UNCLES to find the clusters of genes that meet each of the proposed types of external specifications A and B, but also proves the validity of using the M-N scatter plots approach to select the best clusters from the methods' results.

Most of the clusters generated by our method in both cases, A and B, are irrelevant to the target clusters, that is, they include no true-positives, and they are shown as dense points at the bottom right corners of the F-P plots. Having high densities on the vertical axis, the black continuous carve, and the bottom right corner, with low densities elsewhere, indicates that the results clearly separate the relevant cluster with its different tightness levels from the rest of the irrelevant clusters. To review the relevance of the black curve in these plots please refer to Section 3.6.

There is general agreement between the ground-truth-independent approach (M-N plots) and the ground-truth-dependent approach (F-P plots). Slight perturbations in the ground-truth-independent approach (M-N plots) were seen to lead to such slight perturbations in the ground-truth-dependent approach (F-P plots). This demonstrates the robustness of our approach in selecting the best cluster in an independent manner of the known ground-truth, that is, by the M-N plots approach.

4.3. Comparison with other clustering methods

We have also applied other methods to the same datasets for the sake of comparison with UNCLES. We have tested k-means with Kauffman's initialisation (Pena, et al., 1999), self-organising maps (SOMs) (Xiao, et al., 2003), hierarchical clustering (HC) with Ward's linkage (Eisen, et al., 1998), and the ensemble clustering method relabelling and voting (Vega-Pons & Ruiz-Shulcloper, 2011). These methods were applied separately to each of the six datasets within each of the five sets of datasets at the adopted genome sizes (*GS*) 1200 to 7000 and by considering the ten *K* values 4, 8, 12, 16, 20, 25, 50, 75, 100, and 125. The reason for using high K values for those methods, as opposed to UNCLES, is that those methods do not possess the unique feature of our method, which is the ability to tune the results to obtain tighter clusters while leaving most of the genes unassigned to any cluster. For those methods to obtain clusters of sizes that are comparable to the sizes of the ground truth clusters (75 and 85), high *K* values are needed. In total, each of these four clustering methods has generated 2,610 individual clusters by considering all of the *K* values; remembering that those methods have been applied to the six datasets separately, not collectively. All of these clusters are scattered as blue squares on the F-P plots shown in the second and the fourth columns in Figure 4.1.

In order to statistically measure this observation, we have conducted a pair-wise statistical test between UNCLES and each one of the four methods, and between every possible pair amongst the four methods themselves. While comparing two methods, clusters that include at least one true positive member are identified. Then, the closest 50% of these clusters to the top-left corner of the corresponding F-P plot are considered for a *t*-test. This *t*-test is applied to test if the two subsets of distances are significantly different from each other. The generated statistics are the mean (μ) of the signed differences between distances, its standard deviation (σ), and the p-value. The mean of the signed differences ranges from $-\sqrt{2}$ to $\sqrt{2}$ because the diameter of the F-P plot is $\sqrt{2}$. Closer values to $-\sqrt{2}$ indicate that the clusters generated by the first method have smaller distances from the top left corner of the F-P plot and therefore are better, while the opposite is true when the values are closer to $\sqrt{2}$. Mean values closer to zero indicate that both methods' results are similar to each other.

Table 4.1 shows the results of this statistical test for both clusters C1 and C2, and for all of the considered *GS* values. The third column of the Table shows μ, σ, and the p-value of the comparison between UNCLES and the closest competitor method; the method that was found as the closest competitor is named therein. The fourth column of the Table shows similar metrics for the comparison between the most separated pair of other methods while naming those methods therein.

Table 4.1. Clustering methods' performance comparison

C	*GS*	UNCLES versus closest competitor*	Most separated pair of other methods*
C1	1,200	-0.81 ± 0.15 (9.3 × 10^{-61}) [HC]	-0.13 ± 0.17 (1.5 × 10^{-10}) [HC, RV]
	2,000	-0.88 ± 0.17 (7.3 × 10^{-55}) [HC]	-0.15 ± 0.18 (1.7 × 10^{-11}) [HC, RV]
	3,000	-0.93 ± 0.15 (1.6 × 10^{-68}) [HC]	-0.12 ± 0.16 (2.5 × 10^{-11}) [SOMs, RV]
	5,000	-0.92 ± 0.15 (7.6 × 10^{-66}) [HC]	-0.09 ± 0.14 (1.9 × 10^{-8}) [SOMs, RV]
	7,000	-0.77 ± 0.15 (3.6 × 10^{-54}) [SOMs]	-0.08 ± 0.12 (2.9 × 10^{-9}) [SOMs, RV]
C2	1,200	-0.93 ± 0.15 (< 10^{-255}) [SOMs]	-0.04 ± 0.14 (5.8 × 10^{-7}) [SOMs, RV]
	2,000	-0.92 ± 0.17 (< 10^{-255}) [HC]	-0.04 ± 0.12 (5.0 × 10^{-7}) [HC, RV]
	3,000	-0.60 ± 0.15 (6.3 × 10^{-244}) [HC]	-0.03 ± 0.11 (6.7 × 10^{-5}) [HC, RV]
	5,000	-0.55 ± 0.13 (1.1 × 10^{-234}) [HC]	-0.02 ± 0.09 (2.0 × 10^{-4}) [HC, RV]
	7,000	-0.48 ± 0.13 (4.8 × 10^{-219}) [HC]	-0.02 ± 0.09 (1.3 × 10^{-3}) [HC, RV]

* The format of the entries in these two columns is: μ ± σ (p-value) [method(s)]. The closest competitor to UNCLES is the one with the largest p-value while the most significantly separated pair of other clustering methods is the pair with the smallest p-value.

For both C1 and C2, all of the clusters generated by the other four methods, even at their best, lag significantly behind many of the clusters generated by the UNCLES method including the ones selected by the M-N plot approach as can be seen in the F-P plots in Figure 4.1 and the very negative μ values accompanied with extremely low p-values in the third column of Table 4.1. On the other hand, there is no similarly significant difference between any pair of methods amongst these four as can be seen by the close-to-zero μ values and the p-values that are relatively not very low in the fourth column of Table 4.1.

4.4. Comparison with biclustering methods

Biclustering methods aim at finding genes that are co-expressed, not necessarily in all of the provided data samples, but at least in some of them. A bicluster is a cluster defined by a subset of genes and a subset of data samples (data matrix columns). Here, we compare our UNCLES analysis of the synthetic datasets with eight different biclustering methods.

Biclustering methods can be applied only to a single dataset. Therefore, and given any genome size (*GS*), we have concatenated the six synthetic datasets horizontally to provide a single data matrix with *GS* rows and 82 columns, where this number of columns is the total number of columns (samples) in all of the six datasets. The profiles of the two ground-truth clusters C1 and C2 in the combined dataset are shown in Figure 4.2. The first 42 columns belong to the three positive datasets P1, P2, and P3, while the last 40 columns belong to the three negative datasets N1, N2, and N3, and it can be clearly seen in this Figure that C1 genes are consistently co-expressed in all of the 82 columns (samples) while C2 genes are distinctly co-expressed in the first 42 ones.

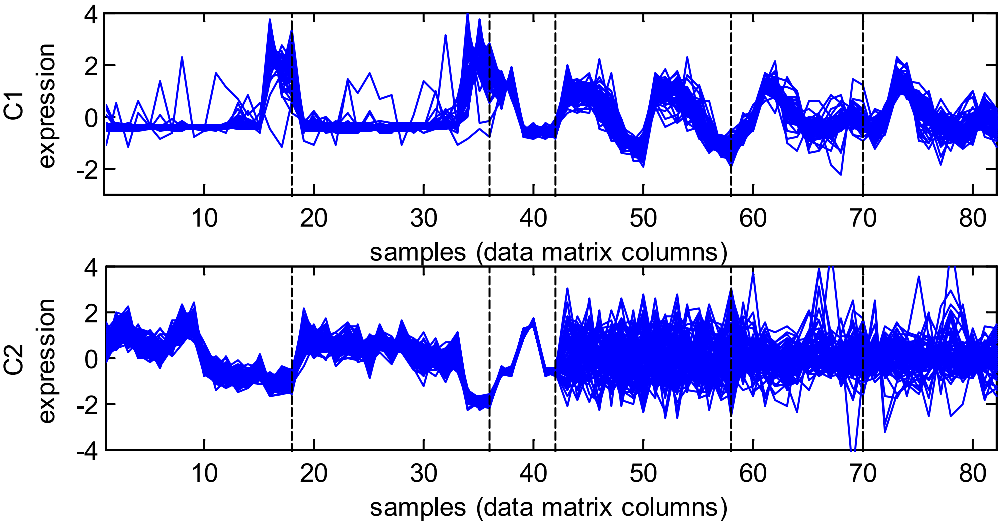

Figure 4.2. Synthetic data ground truth clusters C1 and C2 combined expression profiles from all of the six datasets

The vertical dashed lines show the boundaries between the samples belonging to each of the six datasets in their respective order of P1, P2, P3, N1, N2, and N3. C1 shows consistent co-expression over all of the combined 82 samples (data matrix columns), while C2 shows consistent co-expression only over the first 42 samples.

Eight different biclustering methods were applied to the combined datasets, namely Cheng and Church (CC) (Cheng & Church, 2000), Plaid (Lazzeroni, et al., 2002), bimax (Prelić, et al., 2006), spectral (Kluger, et al., 2003), FLOC (Yang, et al., 2005), XMOTIFS (Murali & Kasif, 2003), large average submatrices (LAS) (Shabalin, et al., 2009), bipartite spectral graph partitioning (BSGP) (Dhillon, 2001). At all genome sizes, Spectral and XMOTIFS produced no clusters, while CC produced a single trivial cluster that encompasses the entire genome and all of the data samples. On the other hand, each one of the remaining five biclustering methods, namely Plaid, Bimax, FLOC, LAS, and BGSP, produced more than one non-empty cluster. Comparison between the UNCLES method and those five biclustering methods is shown in Table 4.2.

Table 4.2 shows two metrics for each method's results considering the clusters C1 and C2 based on each of the five different considered genome sizes (*GS*). The first metric is the shortest distance from the top left corner of the F-P scatter plot; this ranges from 0.0 for the ideal cluster to $\sqrt{2} \cong 1.41$ for the worst possible cluster. The second metric is the number of correctly identified data matrix columns (data samples) out of the total number of correct data matrix columns; for type A, all of the 82 samples (combined from the six datasets) represent the correct samples, while for type B, the 42 samples originally belonging to the positive datasets P1, P2, and P3, are the correct ones.

Table 4.2. Comparison between UNCLES and eight biclustering methods

Cluster and *GS*	UNCLES *#	Plaid *	Bimax *	FLOC *	LAS *	BGSP *
C1 1200	**0.00**	0.10	1.00	1.35	0.13	0.61
	82/82	20/82	4/82	6/82	21/82	1/82
C1 2000	**0.00**	0.64	1.06	1.38	0.16	0.75
	82/82	22/82	4/82	6/82	21/82	2/82
C1 3000	**0.00**	0.95	1.12	1.39	0.29	0.90
	82/82	37/82	4/82	6/82	18/82	0/82
C1 5000	**0.04**	1.28	1.21	1.40	0.45	0.06
	82/82	5/82	3/82	6/82	18/82	0/82
C1 7000	**0.02**	0.97	0.95	1.40	0.59	0.09
	82/82	30/82	4/82	6/82	19/82	0/82
C2 1200	**0.00**	0.76	1.21	1.36	0.31	0.96
	42/42	5/42	3/42	2/42	15/42	0/42
C2 2000	**0.00**	0.92	1.26	1.37	0.28	0.91
	42/42	16/42	3/42	3/42	15/42	0/42
C2 3000	0.33	0.99	1.29	1.38	**0.32**	1.00
	42/42	5/42	3/42	5/42	15/42	0/42
C2 5000	**0.40**	1.07	1.32	1.40	0.71	1.14
	42/42	5/42	3/42	2/42	13/42	0/42
C2 7000	**0.43**	1.18	1.30	1.40	0.70	1.17
	42/42	5/42	3/42	4/42	13/42	0/42

* Each cell in these columns includes two values – the first is the distance from the top-left corner of the ground-truth-based F-P plots for the best cluster found by each method; the ideal is zero and the maximum is $\sqrt{2} \cong 1.41$; the second value is the number of data samples (data matrix columns) which the biclustering algorithms correctly found for the corresponding clusters out of the total number of correct samples (82 for type A and 42 for type B).

# The number of data matrix columns (samples) are prefixed for UNCLES while being variable for biclustering methods.

At all genome sizes, and for both types, type A (cluster C1) and type B (cluster C2), the UNCLES results showed the best performance (minimising the distance and maximising correctly identified data matrix columns / samples). The only exception is for C2 at the gnome size (*GS*) of 3,000 genes, where the LAS method scores a subtly smaller distance than UNCLES. However, even at that latest case, UNCLES' F-P distance is 0.33 compared to 0.32 for LAS, which indicates an insignificant difference between the two distances.

Moreover, LAS and all of the other biclustering methods have identified only few data matrix columns out of the total number of correct columns.

Although all of the biclustering methods lag behind UNCLES, it can be seen that Plaid, LAS, and BSGP, perform relatively better than FLOC and Bimax. In general, LAS shows more consistent quality across varying genome sizes (*GS*) compared to Plaid and BSGP.

4.5. Summary and conclusions

Our validation experiments have demonstrated the unique ability of our proposed method, UNCLES, in answering two research questions with both of its types A and B in an unsupervised and robust manner. We have also validated a novel M-N scatter plots technique for cluster evaluation. This technique was successful in selecting the best clusters while varying the number of clusters (*K* value) as well as the δ and (δ^+, δ^-) values. Therefore, by integrating this technique with the UNCLES method, the method becomes automated and can proceed from the input set of datasets and individual clustering methods to the final few focused clusters without the need to set any critical parameter. The Bi-CoPaM method is equivalent to UNCLES type A, and has accordingly been validated by the validation of UNCLES type A. UNCLES has the potential to be expanded by producing more types of external specifications for the unification of clustering results to meet other research requirements. It is also now ready to be adopted by biologists and other scientists to analyse diverse types of datasets.

Chapter 5
Budding Yeast Data Analysis

Although different to the human cell in many regards, the budding yeast (*Saccharomyces cerevisiae*) cell is orders of magnitude more similar to human cells than species like bacteria are (Duina, et al., 2014). The budding yeast cell is considered as a simple eukaryotic model organism. Eukaryotes include animals, plants, and fungi. Due to this and to the fact that it is relatively easy to grow yeast cells and apply biological experiments to them, budding yeast has become one of the most studied and relatively understood species, leading to many discoveries that have deepened our understanding of eukaryotic cells in general.

In the first section of this chapter, a brief introduction on the molecular biology of budding yeast is presented to pave the way for the reader from a computational background to understand and appreciate the biological experiments and findings presented in the forthcoming sections. Non-biologist readers are encouraged to read Appendix I as well in order to have a sufficient background on cells and their molecular biology.

Two main experiments of application of the Bi-CoPaM method to budding yeast datasets are presented here. Section 5.2 details our analysis of two filtered yeast cell-cycle datasets leading to revealing novel insights into the poorly understood gene CMR1. After that, we developed an approach of applying the Bi-CoPaM method to unfiltered datasets, that is, genome-wise datasets, demonstrating the ability of the Bi-CoPaM to extract focused and meaningful subsets of genes even from unfiltered datasets. Section 5.3 explains a realisation of this approach in which the Bi-CoPaM is applied to forty different genome-wide yeast datasets leading to the discovery of a novel cluster of genes. Each of those studies' findings has been published in journals, namely and respectively in the *Journal of the Royal Society Interface* (Abu-Jamous, et al., 2013b) and *BMC Bioinformatics* (Abu-Jamous, et al., 2014a).

5.1. Introduction to budding yeast molecular biology

Budding yeast, *Saccharomyces cerevisiae*, is a unicellular eukaryotic species, that is, its organism is composed of a single cell which has a real nucleus bound by a nuclear envelope. This species' genome was the first eukaryotic genome to be completely sequenced in 1996 (Goffeau, et al., 1996), and includes about 6,000 different genes distributed over 16 chromosomes (Goffeau, et al., 1996).

Depending on nutrient abundance, budding yeast cells may reproduce asexually by mitosis, that is, cell division to two daughter cells identical to the mother cell, or sexually by cell fusion (Herskowitz, 1988). The former is of importance and relevance to the experiments presented in the rest of this chapter.

Budding yeast mitotic cell-cycle can be divided into four main stages, namely, the first gap (G1), DNA synthesis (S), the second gap (G2), and mitosis, that is, nuclear division (M). When nutrients are not abundant, the cells arrest the cell-cycle in the G1 stage, in which they maintain their cells without further growth. Once nutrients become abundant and several other criteria are checked, the cells proceed to the S stage. The G1/S checkpoint is a key part of the cell-cycle at which a large number of genes are involved and is controlled by a complex regulatory network (Bertoli, et al., 2013). A small bud starts to appear at one side of the cell during the S stage and grows gradually; this bud will become eventually a daughter cell. Also in this stage the genetic material (the DNA packaged by the chromosomes) is replicated to two identical copies of the original one (Omelyanchuk, et al., 2004). The cells enter after that into the second gap (G2). Before entering into the M stage, the cell has to satisfy the G2/M checkpoint's requirements which are checks that guarantee the genetic material's integrity and the readiness to undergo nuclear and cellular division (Bertoli, et al., 2013). In the M stage, the nucleus is divided into two identical daughter nuclei while one of them, which carries one of the two copies of the original genetic material, is pulled towards the growing bud. The bud eventually separates from the mother cell to form a new budding yeast cell.

Many other key processes take place in yeast cells as well as in any other eukaryotic cell. One notable process is protein production. Proteins are produced by ribosomes (protein factories). As detailed in Appendix I, the information required for the synthesis of any protein are stored in the genes in the DNA molecule. A patch of the DNA which includes the information needed to produce a single protein type is copied into a messenger RNA (mRNA) molecule which is translated by the ribosomes into a protein. The ribosomes themselves are composed of many proteins and RNA molecules and are synthesised within the nucleolus, which is a sub-compartment within the nucleus, by a process known as *ribosome biogenesis*. We (Abu-Jamous, et al., 2014a), as demonstrated in Section 5.3, as well as many others (Wade, et al., 2006), have presented evidences showing that ribosome biogenesis significantly increases under growth conditions (e.g. nutrient abundance) while decreases under stress conditions (e.g. lack of nutrients). The cells rather undergo stress-response processes, such as wall maintenance, under stress conditions.

Many aspects of these processes are currently incompletely understood. Nonetheless, this chapter represents a progression towards better understanding of them.

5.2. Analysis of yeast cell-cycle data and the CMR1 gene

Soon after the Bi-CoPaM method had been proposed and validated, we applied it to two filtered yeast cell-cycle datasets leading to biological results elucidating more information regarding the function and regulation of the poorly understood yeast gene CMR1. We published these findings in the *Journal of the Royal Society Interface* (Abu-Jamous, et al., 2013b). We present this study's experiments, results, and conclusions in this section.

5.2.1. Datasets

Two microarray datasets were generated for the yeast *S. cerevisiae* genome using the alpha-30 and alpha-38 synchronisation techniques respectively (Pramila, et al., 2006). Each experiment captures the profiles for the genes over two hours covering two complete cell-cycles. The number of time samples in each is 25 with five-minute intervals between any two consecutive samples.

Pramila and colleagues considered these two datasets in addition to three older ones synchronised by alpha (Spellman, et al., 1998), cdc-15 (Spellman, et al., 1998) and cdc-28 (Cho, et al., 1998) to order the genes according to their periodicity in the cell-cycle (Pramila, et al., 2006). The average time of peak expression for the 1,000 most periodic genes was calculated in that same study as a percentage of time progress in the cell-cycle, that is, peaking at 0% means peaking at the M/G1 transition point, peaking at 50% means peaking in the middle of the cell-cycle, and peaking at 99% means peaking at the very end of the M phase.

The subset of genes which we consider in this study includes the most periodic 500 genes of these 1,000 genes. We consider their profiles from both the alpha-30 and alpha-38 microarray datasets provided in (Pramila, et al., 2006). Supplementary File S1 in our study (Abu-Jamous, et al., 2013b) lists the names of these 500 genes, their peaking times as percentages of the cell-cycle which has been provided by Pramila and colleagues (Pramila, et al., 2006), and their normalised log-ratio expression profiles from both datasets alpha-30 and alpha-38.

5.2.2. Experimental design

The profiles of the selected 500 genes from both alpha-30 and alpha-38 microarray datasets are clustered into four clusters by using the clustering methods: k-means (Pena, et al., 1999), self-organising maps (SOMs) (Kohonen, 1997; Xiao, et al., 2003), hierarchical clustering (HC) (Eisen, et al., 1998), and self-organising oscillator networks (SOON) (Rhouma & Frigui, 2001; Salem, et al., 2008). Both bubble and Gaussian neighbourhood types are used in SOMs; complete, average, and Ward's linkage techniques are used in HC; and varying values of three internal parameters are used in SOON.

The results of these individual clustering experiments are scrutinised to generate one fuzzy consensus partition matrix (CoPaM) which was then binarised by the DTB technique while varying the parameter δ from zero to unity in order to get varying levels of tightness for the clusters. To justify our choice of clustering the 500 genes into four clusters, we have provided more detailed analysis in Supplementary File S2 in (Abu-Jamous, et al., 2013b).

5.2.3. Results

5.2.3.1. Bi-CoPaM results

The numbers of genes (out of a possible 500) included in each of the four clusters C1, C2, C3 and C4 after applying the DTB technique with δ values from 0.0 to 1.0 are listed in Table 5.1. The complete lists of genes included in each of the clusters at all of the considered tightness levels are included in Supplementary File S1 in (Abu-Jamous, et al., 2013b). Note that DTB with $\delta = 0.0$ is equivalent to MVB, and DTB with $\delta = 1.0$ is equivalent to IB. It can be seen that with MVB, the total number of genes assigned to the four clusters is 500 which indicates that complementary clusters are generated where each gene is exclusively assigned to one and only one cluster. While increasing the value of δ to tighten the clusters, fewer genes are included in the clusters and more genes are left unassigned.

It can be seen that the cluster C1 is the tightest cluster as it is the only cluster to survive without being empty until IB. The rest of the clusters ordered by decreasing levels of tightness are C2, C3 and C4. Note that by moving from the absolute tightest case of C1 at IB with 19 genes to the case of DTB with $\delta = 0.95$, which is indeed an extremely tight case, the C1 cluster inflates significantly to include 172 genes while the other three clusters contain few genes if not empty. Less tight clusters derived with DTB and $\delta = 0.95$ do not show big differences in the numbers of genes included in C1.

Table 5.1. Number of yeast genes the clusters C1 to C4 at different δ values

The shaded cases are the ones that are selected to be the clusters' cores.

DTB δ value	C1	C2	C3	C4
0 (MVB)	216	112	90	82
0.1	207	91	85	40
0.2	201	82	83	15
0.3	199	78	81	5
0.4	194	78	76	1
0.5	193	70	60	0
0.6	190	66	21	0
0.7	185	62	2	0
0.8	183	48	1	0
0.9	172	12	0	0
0.95	172	11	0	0
0.98	148	1	0	0
0.99	117	0	0	0
1.0 (IB)	19	0	0	0

To focus on a small subset of genes of potential importance, the smallest reasonable number of genes in each of the four clusters was chosen as the *core* of that cluster. The chosen cores' cases are shaded with grey in Table 5.1. The cores' average peak times as percentages of the cell-cycle as well as the expected corresponding cell-cycle phases from (Pramila, et al., 2006) are listed in Table 5.2. Based on the previous discussion, in the case of C1, although the analysis concentrates on the core at IB, the genes down to DTB with δ = 0.95 are also considered significant and will be referred to as appropriate, see Supplementary File S1 in our study (Abu-Jamous, et al., 2013b) for more details about the profiles of the genes included in C1 at these less tight levels. Revealing the difference in the precision of assignment for these four clusters as well as the ability of choosing different clusters' cores by tuning the level of strictness for different clusters are amongst the useful features provided by the Bi-CoPaM method.

Table 5.2. Average peak time as a percentage of the cell-cycle and the expected cell-cycle phase for the cores of the four yeast clusters

Cluster	C1	C2	C3	C4
Average peak time for core genes	20%	66%	97%	46%
Standard Deviation	3.2%	3.3%	4.9%	6.7%
Min	14%	62%	88%*	40%
Max	27%	75%	6%*	67%
Expected cell-cycle phase	Late G1 / S	G2	M / Early G1	S / G2

* These percentage values are cyclic, that is, after 99%, the cycle goes back to 0%. So the earliest peak in C3 is at 88% of the cycle and the latest is at 6% of the next cycle.

The full lists of the genes in these four cores are listed in Table 5.3 and their expression profiles in both alpha-30 and alpha-38 datasets are plotted in Figure 5.1 (a) and (b) respectively.

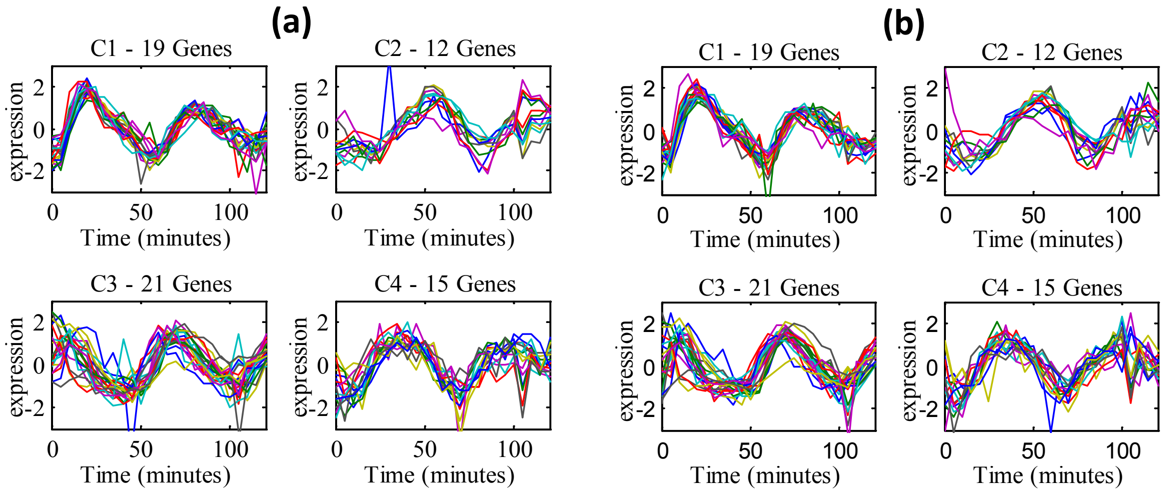

Figure 5.1. The expression profiles of the yeast genes in the four core clusters.

Expression profiles from the (a) alpha-30 dataset and the (b) alpha-38 dataset.

From these two sub-Figures, many observations can be made. First, the alpha-30 and the alpha-38 datasets have very close profiles except for some outlier values; this allows us to use either set for most of the remaining discussions. Second, the profiles of expression over time for the genes that are within each cluster's core are very similar which clearly shows that the Bi-CoPaM approach in increasing strictness to obtain tighter clusters is working as expected. Third, although all of these clusters' cores are tight, the cluster C1 is clearly the tightest, as shown by the δ value at which this core was obtained compared with the others, see Table 5.1. Finally, each set of genes in the four clusters' cores shows periodic peaking at a different stage of the cell cycle, which demonstrates clustering has derived sets of genes with distinct properties (see Table 5.2).

Table 5.3. The four core yeast clusters' genes

C1 core		**C2 core**	**C3 core**		**C4 core**	
at IB (DTB with δ = 1.0)		at DTB with δ = 0.9	at DTB with δ = 0.6		at DTB with δ = 0.2	
(19 genes)		(12 genes)	(21 genes)		(15 genes)	
AXL2	SLK19	BUD20	ASH1	PIG1	ABF1	YGL101W
CDC45	SMC1	CDC5	CHS1	PIL1	CSN9	YJL118W
CHR1	SMC3	CLB1	FAR1	PRY1	FLR1	YLR455W
CMR1	SPC42	CLB2	HSP150	PST1	GDA1	
EXO1	URH1	FET3	HXT2	ROD1	GDT1	
MSH2	YDL163W	FRK1	LSP1	SED1	MBP1	
POL2	YJR030C	PMP3	MCM2	TEC1	MSB1	
POL3		SCW4	MCM3	YLR194C	NDD1	
RAD27		SHE2	MCM4	YNL134C	SSA1	
RFA2		SML1	MCM5		STU2	
RNR1		SRC1	MCM7		TOF2	
RTT107		SWI5	NIS1		VID22	

5.2.3.2. GO term analysis

We have performed GO term analysis (see Section 3.8) for the genes included in the C1 cluster by using the GO Slim tool (SGD, 2014). We have used this tool to search for biological processes, functions, and components GO terms that are enriched in C1 at DTB with all of the values of δ reported in Table 5.1. In summary, the focal cluster in this analysis, C1, is enriched with DNA-binding genes that localise in the nucleus and participate in various cell-cycle and DNA metabolism processes such as DNA repair, recombination, and replication.

5.2.4. Analysis and discussion

In (Gilmore, et al., 2012), a quantitative proteomics approach was adopted to extend the protein network of core histones (H2A, H2B, H3 and H4) in the budding yeast *S. cerevisiae* and identified CMR1 as a member in this network. Some 556 proteins were found binding to one or more histones while only 25 proteins of these were found binding to the four core histone. The 25 proteins include the four histones (H2A, H2B, H3 and H4), two units of the replication factor A (RPA) complex (RFA2 and RFA3), two units of the Ku complex (YKU70 and YKU 80), many units of the RNA polymerase complex (RET1, RPO31, RPC17, RPC37, RPC40 and RPC82), many single-unit proteins (RIM1, YTA7, PSH1, CSE4, ABF2, CKA2, TIF3, DEM1, SUB2 and SMC3), and the previously uncharacterised protein YDL156W / CMR1. Then, associations with the CMR1 protein were investigated and it was found that many proteins showed stable association with it including the six proteins RIM1, RFA2, RFA3, YTA7, YKU70 and YKU80 which are within the 25 proteins found binding to all of the four core histones.

In our Bi-CoPaM gene expression analysis, CMR1 has been found in a small subset of 19 tightly co-expressed genes; Figure 5.2 illustrates the relation between the core histones-associating genes subset and our co-expressed genes subset. It can be seen that three of the 19 co-expressed genes, CMR1, RFA2 and SMC3, in Bi-CoPaM's results are found to be associated with all four core histones. Moreover, RFA2 not only associates with the four histones, it associates with CMR1 itself and is co-expressed with it. Thus Bi-CoPaM provides a strong evidence for the relation between CMR1 and RFA2 in cellular processes.

It is worth mentioning that in our results the histones themselves have been found in the cluster C4 at DTB with $\delta = 0.2$ and not in the cluster C1 which includes CMR1 (see Supplementary File S2 in (Abu-Jamous, et al., 2013b)). This is because the transcription of histones occurs in the S phase in order to synthesise the chromosomes of the forthcoming children cells (Pramila, et al., 2006; Fernandez, et al., 2012); recall from Table 5.2 that the

C4 cluster peaks at the S/G2 phase. Despite that, histone proteins exist within the nucleus, packaging the DNA molecules, at all of the stages of the cell-cycle. Thus, although the CMR1 gene has not been found co-expressed with the histones themselves, it has been found co-expressed with many genes whose products interact with the histones.

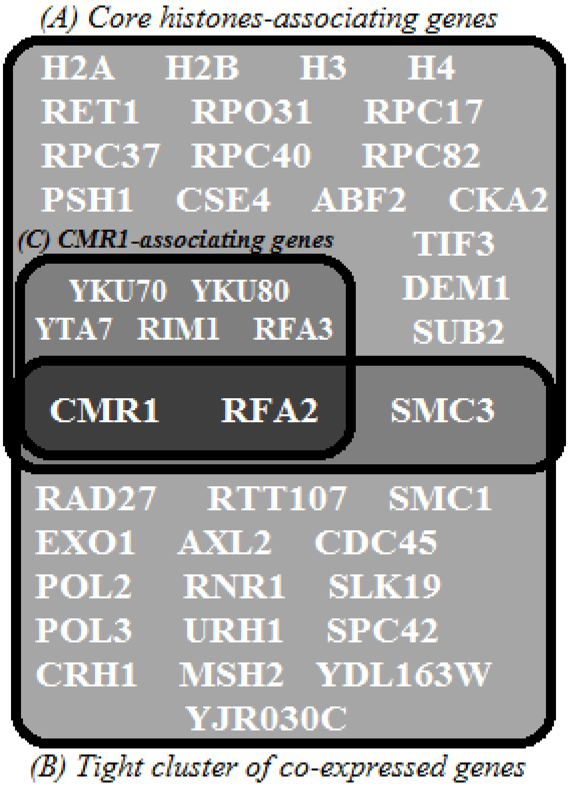

Figure 5.2. Comparison between our 19 tightly co-expressed genes and core histones-associating genes

Venn diagram illustrating relations between the subsets of genes found by using quantitative proteomics to extend the core histone network and the subset of genes found our method of tight gene clustering based on gene expression profiles. The subset (A) represents the 25 genes found to be associated with the four core histones (Gilmore, et al., 2012), the subset (B) represents the seven genes out of those 25 that found associating with CMR1 as well (Gilmore, et al., 2012), and the subset (C) represents the 19 co-expressed genes found in the tightest cluster of genes (C1) by using the Bi-CoPaM method in our study.

Having said that, it can be seen that our computational approach complements the quantitative proteomics approach described by Gilmore and colleagues (Gilmore, et al., 2012) extending the core histone network. The common factor for the genes in the subset provided in (Gilmore, et al., 2012) is the association with the four histones while the common factor for the genes in our results is highly synchronous co-expressions through the cell cycle.

The notable observation in both subsets is the existence of strongly functionally related genes that are often components of the same protein complex or the same pathway. The three components of the replication protein A (RPA); RFA1, RFA2 and RFA3 seem to be the closest to the newly characterised gene CMR1 in that RFA2 appeared in both sets of results associated with the four histones, associated with CMR1 and co-expressed with it, and that RFA1 and RFA3 appeared in the same subset of CMR1 in either results. Gilmore and colleagues explored the relationship between CMR1 and the RNA polymerase complex III (Gilmore, et al., 2012). Although they noticed the possibility that CMR1 would participate in the DNA repair at the G1/S checkpoint, they did not investigate this further. Our results suggest such a relationship may be functionally significant.

We propose that CMR1 may have a functional relationship not only with DNA polymerases but also with the cohesion complex. Most of the components of the DNA polymerases α, δ and ε are found to be tightly co-expressed with CMR1 and suggests a possible role of CMR1 in DNA replication and repair. SMC3, a core component of the cohesion complex, is found in Bi-CoPaM results and by Gilmore and colleagues (Gilmore, et al., 2012) was associated with CMR1, while the other components of the complex were associated with CMR1 in our analysis. The strong association of CMR1 with the known targets of the MBF complex even in the extreme tightest cases clearly suggests the hypothesis that CMR1 expression is controlled by the MBF complex, the hypothesis which can be tested in future experimental work.

5.2.5. Conclusions

Our results have highlighted important subsets of genes based on the computational analysis of high-throughput data from different experiments instead of traditional biological or biochemical experiments. They not only add stronger evidence for the main findings of the study of Gilmore and colleagues (Gilmore, et al., 2012), but they also strongly highlight areas of less previous attention about the function of the CMR1 gene. CMR1 has been postulated to have functions in DNA processing. We have shown its expression through the cell cycle would support a relation between CMR1 with the RPA complex, DNA polymerases and the cohesion complex in addition to its role at the G1/S transition.

Finally, we also provide novel clusters with co-expressed genes under tuneable tightness levels. The evidence for the validity of these clusters' tight cores comes from the fact that they include many genes that are strongly related by being in the same complex or pathway. These novel clusters can serve as an important resource for further focussed gene discovery studies.

5.3. APha-RiB: a novel cluster of poorly understood genes discovered in the analysis of forty datasets

Following its successful application to filtered datasets, we have applied the Bi-CoPaM to unfiltered genome-wide datasets in a novel way demonstrating the ability of the Bi-CoPaM to embed filtering within its course of application. This study, in which the Bi-CoPaM is applied to forty recent budding yeast genome-wide datasets, has led to unveiling a novel cluster of genes which is consistently oppositely co-expressed with a well-known and previously defined cluster of genes. The novel conclusions of this comprehensive study have been published in the journal *BMC Bioinformatics* (Abu-Jamous, et al., 2014a).

5.3.1. Datasets and experimental design

In this set of analysis, we consider forty recent *Saccharomyces cerevisiae* microarray datasets which were generated by using the Affymetrix yeast genome 2.0 array in the last six years, and include at least four different conditions or time-points. Although choosing datasets generated by using the same array is not a condition for Bi-CoPaM analysis, it allows for more genes to be included in the analysis as some genes might not be represented by probes in all types of arrays, and therefore have to be discarded from the analysis in such a case. Each of these datasets measures the genetic expression of the entire yeast genome (5,667 genes) over multiple time-points or conditions. The details of the datasets are listed in Table 5.4. The datasets span a wide range of biological conditions such as cell-cycle, stress response, mutated strains growth, treatment with various types of agents, and others. The 5,667 genes are listed in Supplementary Table 1 in (Abu-Jamous, et al., 2014a).

Table 5.4. Budding yeast forty microarray datasets

ID	GEO accession	Year	N	Description	Ref.
D01	GSE8799	2008	15	Two mitotic cell-cycles (w/t).	(Orlando, et al., 2008)
D02	GSE8799	2008	15	Two mitotic cell-cycles (mutated cyclins).	(Orlando, et al., 2008)
D03	E-MTAB-643*	2011	15	Response to an impulse of glucose.	(Dikicioglu, et al., 2011)
D04	E-MTAB-643*	2011	15	Response to an impulse of ammonium.	(Dikicioglu, et al., 2011)
D05	GSE54951	2014	6	Response of *dal80Δ* mutant yeast to oxidative stress induced by linoleic acid hydroperoxide.	-
D06	GSE25002	2014	9	Osmotic stress response and treatment of transformants expressing the C. albicans Nik1 gene.	-
D07	GSE36298	2013	6	Mutations of OPI1, INO2, and INO4 under carbon-limited growth conditions.	(Chumnanpuen, et al., 2013)
D08	GSE50728	2013	8	120-hour time-course during fermentation.	-
D09	GSE36599	2013	5	Stress adaptation and recovery.	(Xue-Franzén, et al., 2013)
D10	GSE47712	2013	6	Combinations of the yeast mediator complex's tail subunits mutations.	(Larsson, et al., 2013)
D11	GSE21870	2013	4	Combinations of mutations in DNUP60 and DADA2.	-
D12	GSE38848	2013	6	Various strains under aerobic or anaerobic growth.	(Liu, et al., 2013)
D13	GSE36954	2012	6	Response to mycotoxic type B trichothecenes.	(Suzuki & Iwahashi, 2012)
D14	GSE33276	2012	6	Response to heat stress for three different strains.	-
D15	GSE40399	2012	7	Response to various perturbations (heat, myriocin treatment, and lipid supplement).	-

D16	GSE31176	2012	6	W/t, rlm1Δ, and swi3Δ cells with or without Congo Red exposure.	(Sanz, et al., 2012)
D17	GSE26923	2012	5	Varying levels of GCN5 F221A mutant expression.	(Lanza, et al., 2012)
D18	GSE30054	2012	31	CEN.PK122 oscillating for two hours.	-
D19	GSE30051	2012	32	CEN.PL113-7D oscillating for two hours.	(Chin, et al., 2012)
D20	GSE30052	2012	49	CEN.PL113-7D oscillating for four hours.	(Chin, et al., 2012)
D21	GSE32974	2012	15	About 5 hours of cell-cycle (w/t).	(Kovacs, et al., 2012)
D22	GSE32974	2012	15	About 4 hours of cell-cycle (mutant lacking Cdk1 activity).	(Kovacs, et al., 2012)
D23	GSE24888	2011	5	Untreated yeast versus yeasts treated with E. arvense herbs from the USE, China, Europe, or India.	-
D24	GSE19302	2011	6	Response to degron induction for w/t and nab2-td mutant.	(González-Aguilera, et al., 2011)
D25	GSE33427	2011	5	Untreated w/t, and wt/t, yap1Δ, yap8Δ, and double mutant treated with AsV.	(Ferreira, et al., 2012)
D26	GSE17716	2011	7	Effect of overexpression and deletion of MSS11 and FLO8.	(Bester, et al., 2012)
D27	GSE31366	2011	4	Presence and absence of mutli-inhibitors for parental and tolerant strains.	-
D28	GSE26171	2011	4	Response to patulin and/or ascorbic acid.	(Suzuki & Iwahashi, 2011)
D29	GSE22270	2011	4	PY1 and Met30 strains in room temperature or 35 C.	-
D30	GSE29273	2011	4	Time-series during yeast second fermentation.	-
D31	GSE29353	2011	5	Different haploid strains growing in low glucose medium.	(Parreiras, et al., 2011)
D32	GSE21571	2011	8	Different combinations of mutations in HTZ1, SWR1, SWC2, and SWC5.	(Morillo-Huesca, et al., 2010)
D33	GSE17364	2010	4	Untreated w/t and Slt2-deficient yeasts, or treated with sodium arsenate for two hours.	(Matia-González & Rodríguez-Gabriel, 2011)
D34	GSE15352	2010	8	24-hour time-course of yeast grown under a low temperature (10 C).	(Strassburg, et al., 2010)
D35	GSE15352	2010	8	24-hour time-course of yeast grown under a normal temperature (28 C).	(Strassburg, et al., 2010)
D36	GSE15352	2010	8	24-hour time-course of yeast grown under a high temperature (37 C).	(Strassburg, et al., 2010)
D37	GSE16799	2009	21	UC-V irradiation of w/t, mig3Δ, SNF1Δ, RAD23Δ, RAD4Δ, and snf1Δrad23Δ.	(Wade, et al., 2009)
D38	GSE16346	2009	4	BY474 cells grown to mid-log under presence versus absence of L-carnitine and/or H2O2.	-
D39	GSE14227	2009	10	Two hours of wild-type yeast growth.	(Ge, et al., 2010)
D40	GSE14227	2009	9	Two hours of sch9Δ mutant yeast growth.	(Ge, et al., 2010)

The first column shows the unique identifier which is used hereinafter to refer to each of these datasets. The second to the sixth columns respectively show the Gene Expression Omnibus (GEO) accession number, the year in which the dataset was published, number of time-points or conditions after replicate summarisation, dataset description, and reference.

* D03 and D04 have accession numbers in the European Bioinformatics Institute (EBI) repository rather than GEO accession numbers.

Those 5,667 genes were clustered into sixteen clusters by k-means with Kauffman's initialisation (KA) (Pena, et al., 1999), self-organising maps (SOMs) with bubble neighbourhood and four-by-four grid (Xiao, et al., 2003), and hierarchical clustering (HC) with Ward's linkage (Eisen, et al., 1998). This was applied to their profiles from all of the forty datasets. The generated partitions were combined into a single consensus partition matrix (CoPaM) as explained in Section 3.2.3 where a min-min approach was adopted for relabelling at the CoPaM generation step. The final CoPaM was binarised by the DTB technique with δ values ranging from zero to unity and then analysed by the MSE metric described in Section 3.4. Prior to clustering, the datasets were normalized by quantile normalisation (Bolstad, et al., 2003). Then each gene's expression profile was shifted and scaled to be zero-mean and unity standard deviation. Also, when many replicates exist for the same time-point or condition, they are summarised by considering their median value.

5.3.2. Results and analysis

The numbers of genes in the sixteen clusters at all of the varying δ values are shown in Table 5.5. Clusters were ordered based on their tightness such that those clusters that preserve at least seven genes up to higher values of δ are considered tighter. When many clusters preserve at least seven genes up to the same value of δ, they are ordered based on the number of genes they include at that level. The number 'seven' is just used for ordering and is not a critical parameter; if it had been set to 'ten' instead for example, no significant change in cluster ordering would have be observed. The complete lists of gene names included in each of these clusters at all of the δ values are provided in Supplementary Table 1 in (Abu-Jamous, et al., 2014a).

Table 5.5. Numbers of genes included in each of the 16 clusters at all of the considered δ values

Tightness	δ	**Cluster**															
		C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**	**C12**	**C13**	**C14**	**C15**	**C16**
Complementary	0.0	1085	1457	610	655	592	268	303	175	175	154	143	92	51	49	29	10
	0.1	516	394	84	105	79	12	9	3	1	2	2	0	0	0	0	0
	0.2	344	47	17	14	2	0	0	0	0	0	0	0	0	0	0	0
	0.3	257	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
	0.4	164	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.5	79	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.6	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tightest	1.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

5.3.2.1. MSE analysis

The MSE values for each of the tightest six clusters were calculated at all of the DTB δ values as explained in Section 3.4. Each of these values was calculated based on the forty datasets and then averaged and plotted in Figure 5.3 (A). Figure 5.3 (B) shows the numbers of genes included in each of these six clusters at all of the δ values. Missing points in both plots represent empty clusters.

We have considered the MSE metric in tandem with the number of genes included in the clusters to choose a few clusters for further analysis and discard the other ones. The objective here is to minimise the MSE values while maximising the number of genes included in the clusters. This approach overcomes the dependency of MSE values on the numbers of genes included in the clusters. As can be seen in Figure 5.3 (A) and (B), the cluster C1 shows significantly lower (better) values of MSE while including significantly higher numbers of genes. The cluster C2 comes next to C1 in terms of having lower MSE values with more genes.

On the other hand, while the clusters C3 and C4 have comparative MSE values at δ = 0.2 with C2 (Figure 5.3 (A)), they have significantly lower numbers of genes (17 and 14 genes respectively for C3 and C4 in comparison with 47 in C2; see Table 5.5). Furthermore, the clusters C5 and C6 are significantly worse (higher MSE values with fewer genes) than the first four clusters (Figure 5.3). While the average MSE values for the seventh to the sixteenth clusters have not been included in this Figure, the numbers of genes included in these clusters at relatively lower levels of tightness, as shown in Table 5.5, are sufficient to filter them out. Therefore, we have considered the clusters C1 and C2 for further analysis in this study.

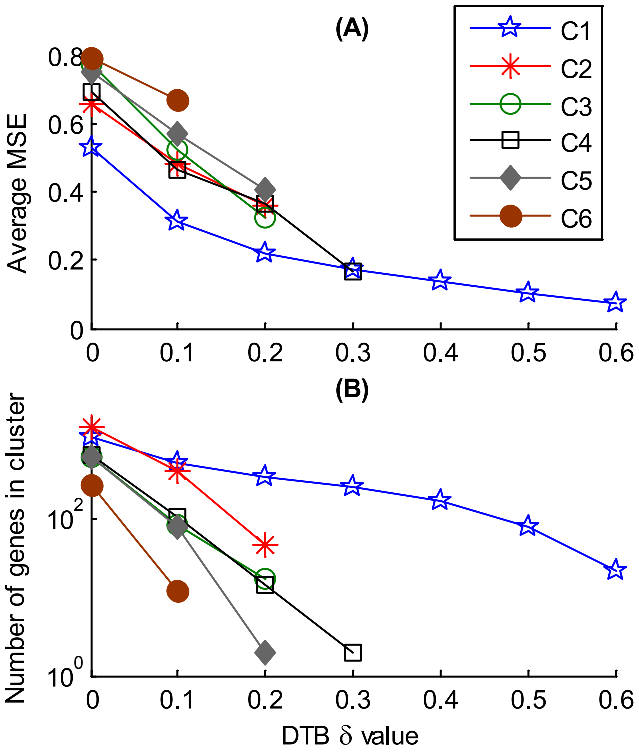

Figure 5.3. MSE and cluster size analysis of yeast clusters.

(A) Average MSE values and (B) number of genes included in the tightest six clusters over all of the adopted δ values.

5.3.2.2. Average expression profiles

The average expression profiles for the clusters C1 and C2 at DTB with δ = 0.3 and 0.2 respectively, in each of the forty datasets are plotted in Figure 5.4. For clarity, error bars have been suppressed as the information, which they provide can be obtained from the MSE analysis in Figure 5.3 and the plots in Supplementary Figure 1 of the study (Abu-Jamous, et al., 2014a), which shows the expression profiles of all of the genes in these two clusters at various δ values.

Detailed scrutiny of Figure 5.4 leads to the general observations that the first cluster, C1, is up-regulated when cells are released from stress conditions such as nutrient limitation; they are down-regulated when stress conditions are re-imposed. Most interestingly, the cluster C2 shows opposite average expression profiles in almost all of the forty datasets to the average profiles of cluster C1 with no phase shift, that is, with neither profile leading or lagging the other; its genes are up-regulated under stress conditions and down-regulated under growth conditions. It is interesting, but had not been anticipated at the time of experimental design before obtaining the results, that the two most consistently co-expressed clusters of genes in budding yeast show such clear opposite expression profiles across large number of datasets.

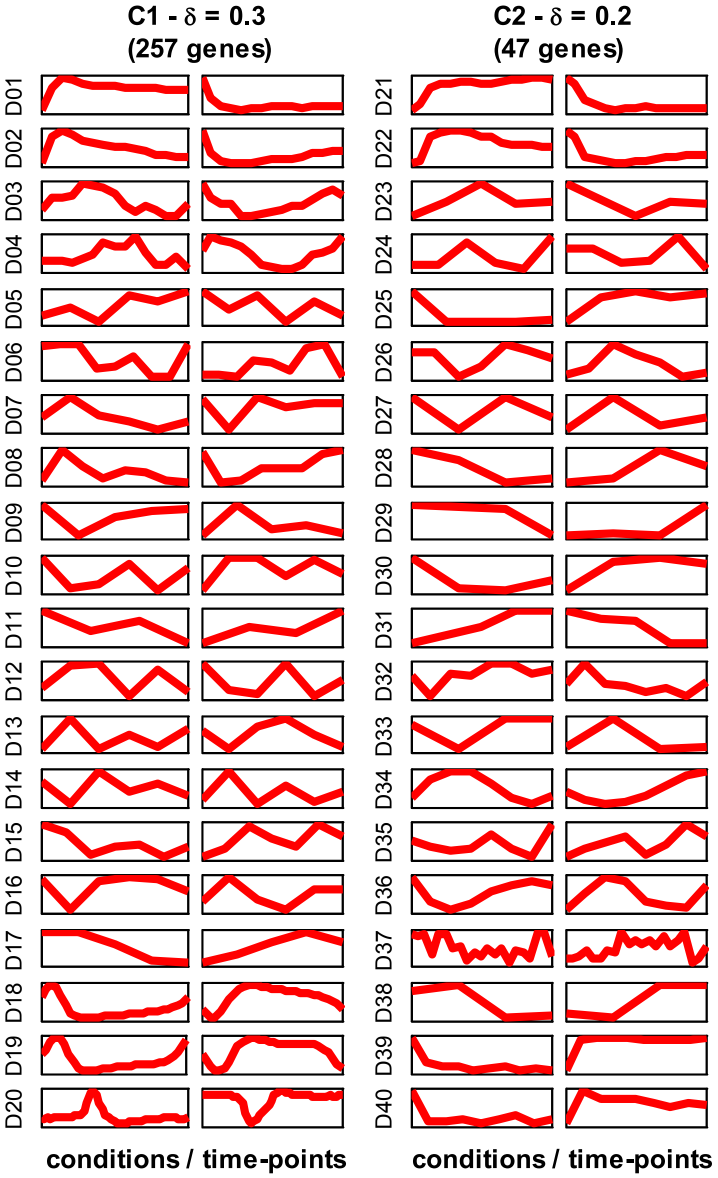

Figure 5.4. Average expression profiles for the clusters C1 and C2

This is at DTB with the respective δ values of 0.3 and 0.2, based on all of the forty datasets. Each column of plots represents a cluster and each row represents a dataset.

To assess that observed opposite co-expression quantitatively, we have calculated the Pearson's correlation values between the average expression profiles of C1 at $\delta = 0.3$ and C2 at $\delta = 0.3$ from each of the forty datasets. A very strong negative correlation has been found, that is lower than the value of –0.75 at 37 out of 40 datasets and never exceeds the value of –0.6 except at a single outlier dataset, D35. This strong negative correlation is consistent even when the δ values are varied. For instance, when considering C1 at the δ values of 0.2 and 0.4, the calculated correlation values are lower than -0.75 at 38 and 36 out of 40 datasets, respectively. Even when considering C2 at $\delta = 0.1$, the case at which its size is many folds larger than at $\delta = 0.2$ (394 genes versus 84), 35 out of 40 datasets show strong negative correlation with values lower than -0.75, and only couple of datasets exceed the value of -0.7. The single outlier dataset D35 has consistently shown notably weaker negative correlation at all of the aforementioned δ values. These experiments demonstrate the robustness of our observation that C1 and C2 are consistently negatively correlated.

5.3.2.3. Upstream sequence analysis

Because co-expression over large number of different microarray datasets strongly indicates co-regulation, we have analysed the upstream DNA sequences for the genes in the clusters C1 and C2 to explore potential common transcription factors' binding sites. We have used the MEME tool (Bailey & Elkan, 1994; MEME, 2014) to search for the most enriched DNA sequence motifs within the 300 upstream base-pairs of the 164 genes included in C1 at DTB with $\delta = 0.4$. The three discovered motifs, which we label as C1-1, C1-2, and C1-3 respectively, were then fed to the TOMTOM tool (Gupta, et al., 2007; TOMTOM, 2014) to mine for previously known motifs with high similarity. The first motif, with an E-value of 3.3×10^{-333}, was found to be the PAC motif, which is the binding site of the two paralogous transcription factors Dot6p and Tod6p with p-values of 2.1×10^{-5} and 1.4×10^{-4}, respectively, and it significantly matches the binding site of the transcription factor Sfl1p with a p-value of 1.3×10^{-4} (Figure 5.5 (A)). The E-value estimates the expected number of motifs with the given probability or higher, and with the same width and site count, that would be found in a set of random sequences of a similar size. The second motif, with an E-value of 2.2×10^{-115}, was found to be the RRPE motif, which is the binding site of the transcription factor Stb3p with a p-value of 8.9×10^{-7} (Figure 5.5 (B)); it also significantly matches the binding sites of the transcription factors Sum1p and Sfp1p with p-values of 2.7×10^{-5} and 3.2×10^{-5}, respectively (Figure 5.5 (B)). The third motif, with an E-value of 3.2×10^{-63}, was found to match the binding sites of the transcription factors Azf1 and Sfl1p with p-values of 1.3×10^{-4} and 2.0×10^{-4}, respectively (Figure 5.5 (C)). The three motifs were respectively found in the upstream sequences of 148, 119, and 56 genes out of 164 possible ones. Figure 5.5 (D) is a Venn diagram, which shows the numbers of genes the upstream DNA sequences of which contain each of these three motifs.

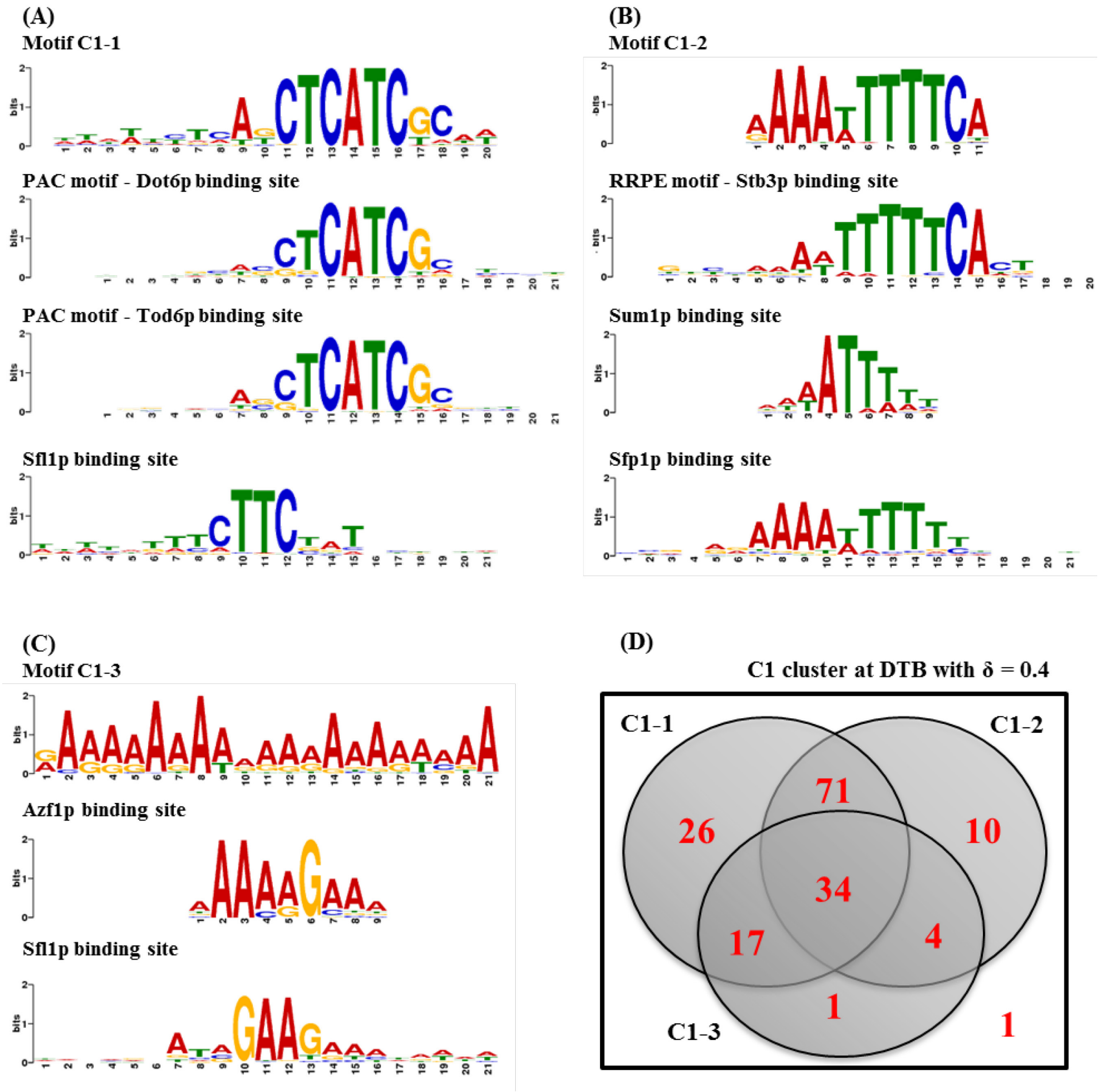

Figure 5.5. Upstream sequence motifs for genes in the cluster C1

(A), (B), and (C) show the motifs C1-1, C1-2, and C1-3 respectively and their highly matched known transcription factors' binding sites. (D) is a Venn diagram that shows the numbers of genes' upstream sequences in C1 that contain each of these three motifs.

Similarly, the MEME tool was used over the 47 genes included in the cluster C2 at DTB with $\delta = 0.2$. The logos of the two discovered motifs, which we label as C2-1 and C2-2, are shown in Figure 5.6 (A) and (B), respectively. The E-values for the two motifs are 1.6×10^{-23} and 5.3×10^{-4} respectively, and they were found in the upstream sequences of 31 genes and 21 genes, out of 47 genes in C2 at DTB with $\delta = 0.2$ (Figure 5.6 (C)). A third motif was found by the MEME tool in this cluster but with the high E-value of $2.8\times10^{+1}$ and in the upstream sequences of 13 genes only; therefore it has been discarded from further analysis. The motifs C2-1 and C2-2 were then fed to the TOMTOM tool (Gupta, et al., 2007; TOMTOM, 2014) to mine for previously known motifs that have high similarity to them.

The motif C2-1 was found to match the binding site of the transcription factor Azf1p (p-value 5.4×10^{-6}), while C2-2 was found to match the STRE element which is the binding site of the transcription factor Msn4p (p-value 5.4×10^{-4}) and its paralogue Msn2p (p-value 6.2×10^{-4}). The logos of the binding sites of these transcription factors aligned with the discovered motifs are shown in Figure 5.6 (A) and (B), respectively.

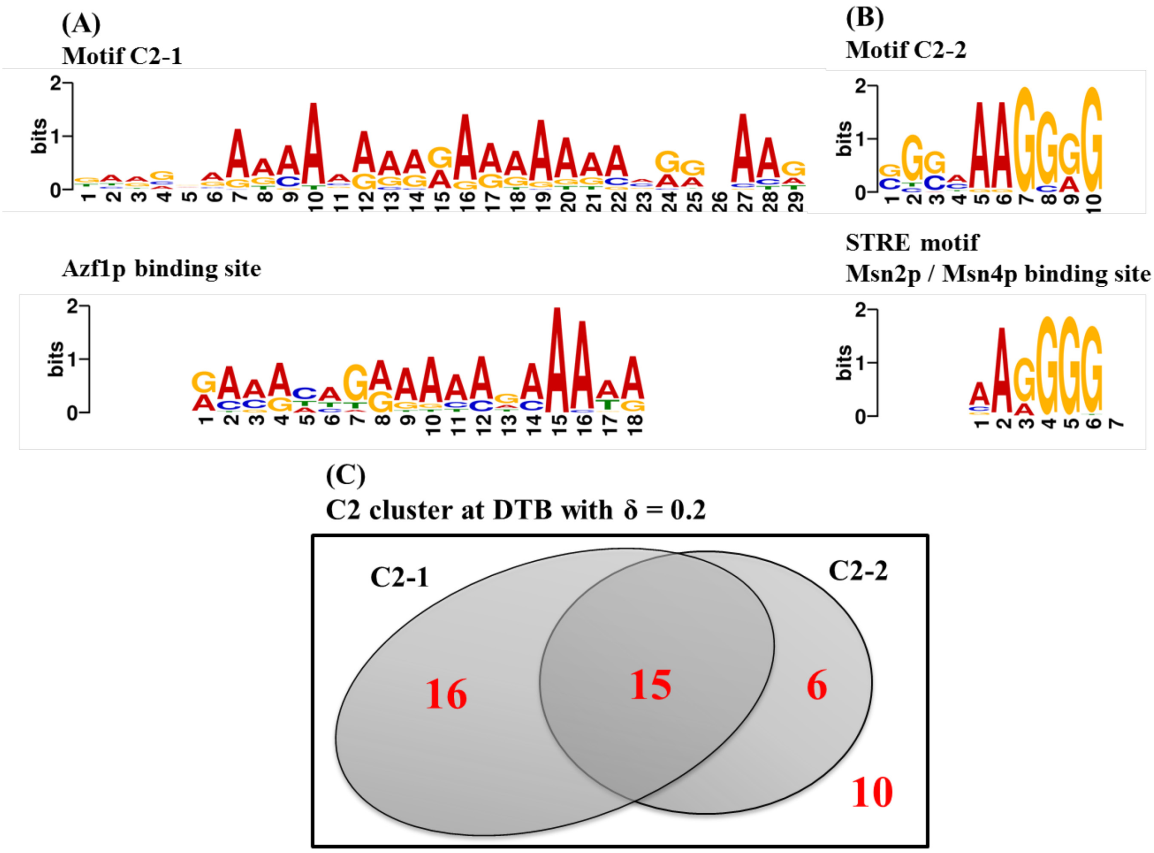

Figure 5.6. Upstream sequence motifs for genes in the cluster C2

(A) and (B) show the motifs C2-1 and C2-2 respectively and their highly matched known transcription factors' binding sites. (C) is a Venn diagram that shows the numbers of genes' upstream sequences in C2 that contain each of these two motifs.

5.3.2.4. GO term analysis

To link our observations over the clusters' expression profiles with biological terms, we have performed Gene Ontology (GO) analysis (Peng, et al., 2013) over the clusters C1 and C2 at different tightness levels by using the GO Term Finder tool (SGD, 2014), and the GO Slim Mapper tool (SGD, 2014). The most enriched GO process terms in these clusters, as well as the numbers of genes annotated with the GO term "biological process unknown", are listed in Table 5.6. Supplementary Tables 2 and 3 in (Abu-Jamous, et al., 2014a) include the complete GO term results, for the clusters C1 and C2 at all of the values of δ at which they are not empty.

The cluster C1 is extraordinarily highly enriched with genes that participate in ribosome biogenesis and rRNA processing (RRB), and it includes a small number of genes of unknown biological process.

Table 5.6. Most enriched GO terms in the clusters C1 and C2 at various levels of tightness

GO process		Back. frequency	δ = 0.1		δ = 0.2		δ = 0.3		δ = 0.4	
			Freq.	P-val.	Freq.	P-val.	Freq.	P-val.	Freq.	P-val.
C1	Ribosome biogenesis	411/7167	210/516	E-140	183/344	E-146	153/257	E-129	124/164	E-123
	Biological process unknown*	1189/6334	46/516		26/344		17/257		9/164	
C2	Response to oxidative stress	101/7167	23/394	E-6	6/47	E-3				
	Oxidation-reduction process	174/7167	33/394	E-7	3/47	>E-1				
	Biological process unknown*	1189/6334	114/394		12/47					

* The enrichment of the "biological process unknown" term has been found by the GO Slim Mapper tool rather than the GO Term Finder tool. Note that the p-value is only provided by the GO Term Finder tool.

In contrast, the genes included in the cluster C2 include a large group of unknowns (12 genes, 25.5%, with unknown biological process out of 47 in C2 at $\delta = 0.2$, and 114 out of 394, 28.9% at $\delta = 0.1$), and even the genes with currently known processes do not show dominant enrichment for any single process. Relatively, the most enriched known biological processes within the 47 genes included in this cluster at $\delta = 0.2$ are response to oxidative stress (six genes, 12.8%) and oxidation-reduction (three genes, 6.4%); no genes are shared between these two processes. Other processes with which some genes in this cluster have been associated are lipid metabolic process (four genes, 8.5%), carbohydrate metabolic process (four genes, two of which has also been associated with oxidation-reduction, and one with response to oxidative stress), cellular amino acid metabolic process (four genes, one of which has also been associated with response to oxidative stress), protein phosphorylation (three genes, one of which has also been associated with oxidation-reduction), mitochondrial organisation (two genes), cofactor metabolic process (two genes), regulation of cell cycle (two genes, one of which has also been associated with oxidation-reduction), endocytosis (two genes, one of which has also been associated with protein phosphorylation), and response to heat (two genes, one of which has also been associated with protein phosphorylation).

We have also searched for the enrichment of the cellular components in which the genes included in C2 at DTB with $\delta = 0.2$ localise. The complete lists of results are provided in Supplementary Table 4 of the study (Abu-Jamous, et al., 2014a). Figure 5.7 shows the distribution of the genes included in C2 at that tightness level over main cellular components while marked based on their biological processes. It can be seen that there is a large distribution of processes as well as components with no single process or component dominating.

In conclusion, we name the subset of genes in C2 as “anti-phase with ribosome biogenesis regulon”, or the *APha-RiB regulon*. This is because its main characterising feature is its consistently opposite expression with the RRB regulon (C1).

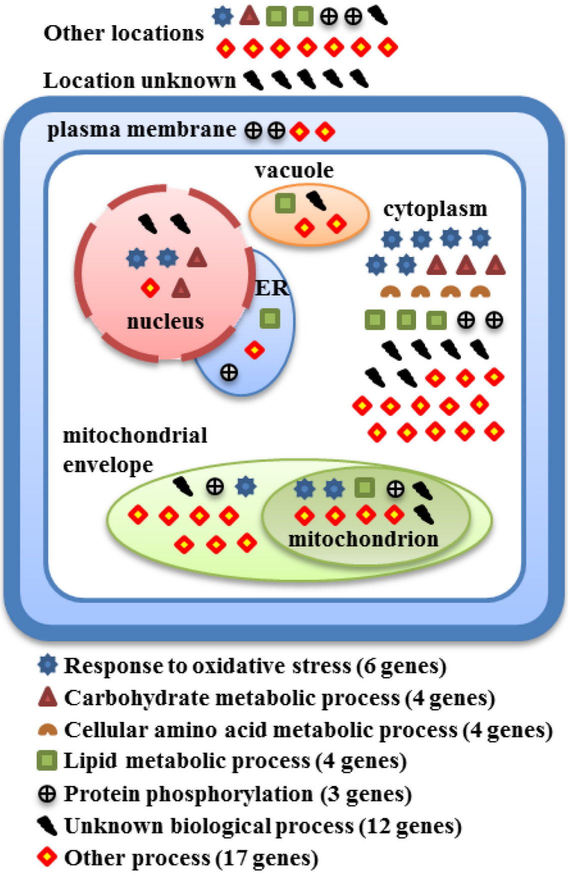
Other locations
Location unknown
plasma membrane
vacuole
cytoplasm
ER
nucleus
mitochondrial envelope
mitochondrion
Response to oxidative stress (6 genes)
Carbohydrate metabolic process (4 genes)
Cellular amino acid metabolic process (4 genes)
Lipid metabolic process (4 genes)
Protein phosphorylation (3 genes)
Unknown biological process (12 genes)
Other process (17 genes)

Figure 5.7. Distribution of C2 genes over cellular components and biological processes

The 47 genes included in C2 at DTB with $\delta = 0.2$ are distributed based on the biological processes with which they have been associated over the major cellular components. Note that any single gene might be found in multiple cellular components, and thus the total number of gene markers in the Figure does not directly correspond to the total number of genes considered.

5.3.2.5. Gene network analysis

GeneMANIA is a tool which mines a database of various types of interactions identified by high-throughput studies in the literature to draw networks of interactions for a subset of query genes (GeneMANIA, 2014). By using this tool, we have obtained networks of genetic interactions (Figure 5.8) and protein-protein physical interactions (Figure 5.9) between the 47 genes included in the APha-RiB regulon (cluster C2 at $\delta = 0.2$).

We have also used GeneMANIA to find the network of genetic co-expression between the 47 APha-RiB genes in order to validate their consistent co-expression. The produced network contains 962 co-expression links out of 1,081 possible ones (89%) in this undirected graph of 47 nodes. To test the statistical significance of these figures, we randomly generated ten different groups of genes, each of which has 47 genes, and fed them to the GeneMANIA tool. The average number of co-expression links was 380 links with a standard deviation of 32. Therefore, by assuming a normal distribution, the p-value of having 962 links between 47 nodes is 6.7×10^{-73}, which proves the validity of including those 47 genes in a single cluster.

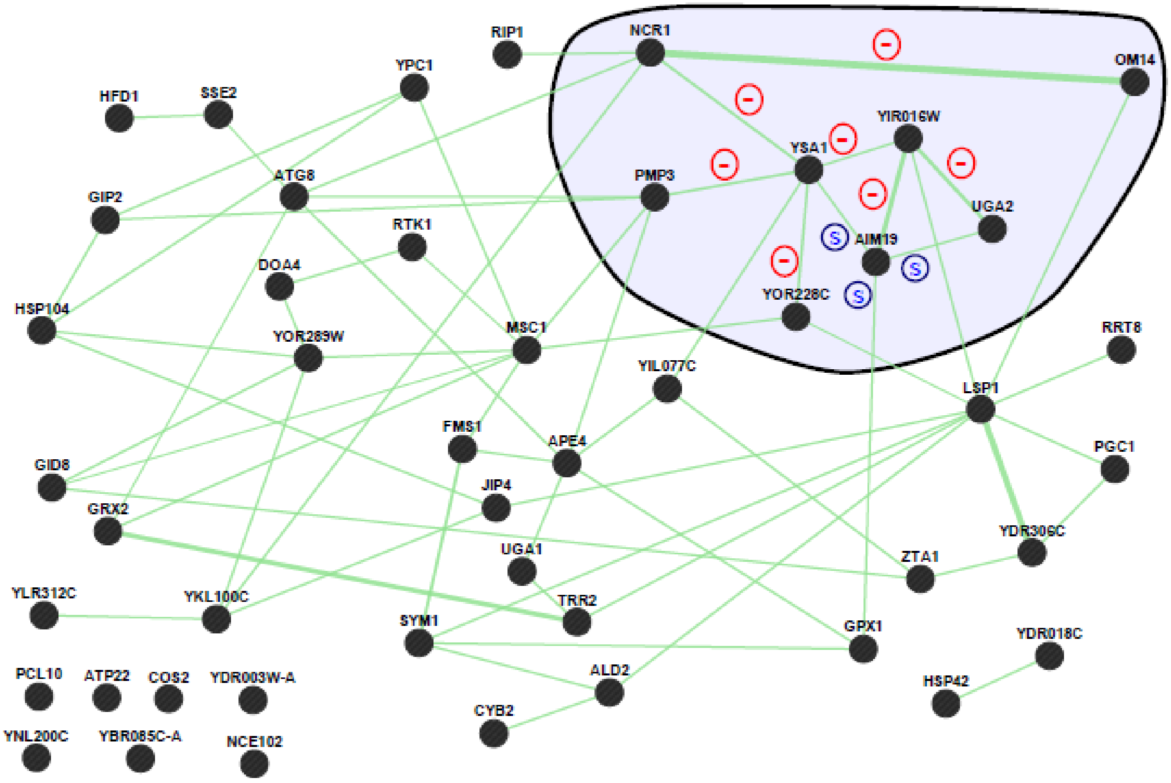

Figure 5.8. Genetic interaction network between the genes in the APha-RiB regulon

The APha-RiB regulon is the cluster C2 at DTB with δ = 0.2. A sub-network of eight genes is highlighted and the types of genetic interactions between its genes are labelled. This is the same sub-network which is highlighted in Figure 5.9. A genetic interaction exists between two genes if the impact of perturbing both genes is different from the additive impact of perturbing each gene individually. A positive genetic interaction is that in which perturbing both genes results in a higher fitness, that is a weaker defect, than the additive defect of perturbing each one individually. On the other hand, a negative genetic interaction exists when the defect caused by perturbing both genes is stronger than the additive defect caused by perturbing each gene individually. A similar profile (S) genetic interaction indicates high correlation between both genes' genetic interaction profiles with the rest of the genes.

A sub-network of eight genes is highlighted in Figure 5.8 and Figure 5.9 because they have high connectivity in both genetic and protein-protein physical interactions networks. The types of the genetic interactions between those eight genes are also labelled in Figure 5.8. Based on the high-throughput study by Costanzo and colleagues (Costanzo, et al., 2010), two genes have positive genetic interaction between them if the effect of perturbing both genes is higher than the additive effect of perturbing each gene individually. Similarly, they have negative genetic interaction if the effect of perturbing both of them is less than the additive effect of perturbing each one of them individually. If the effect of perturbing both of them is similar to the additive effect of perturbing each of them individually, they do not have genetic interaction. The interactions labelled with (S) in

Figure 5.8 indicate that there is high correlation between the genetic interaction profiles of those two genes with the other genes in the yeast genome.

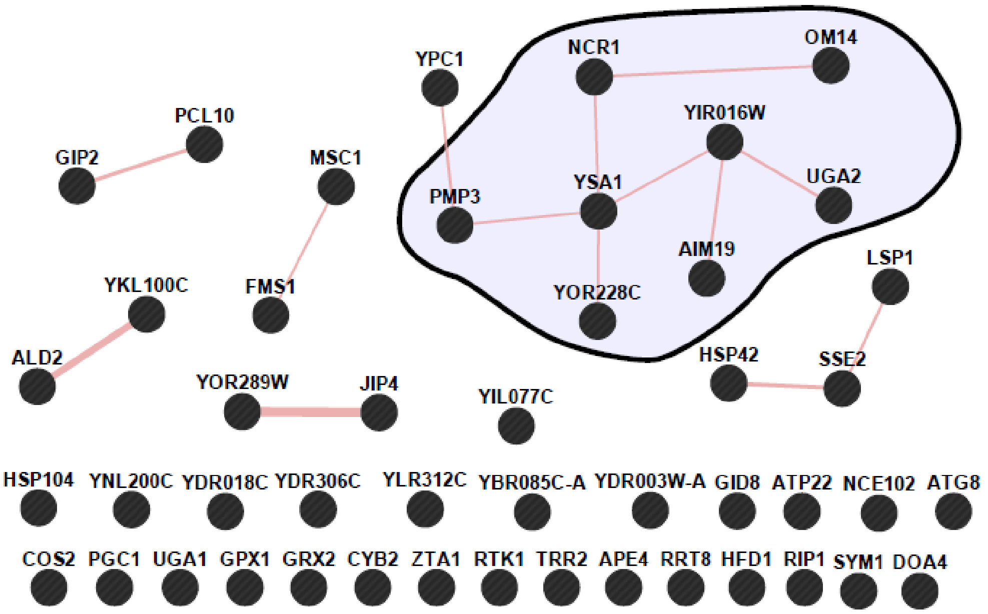

Figure 5.9. Protein-protein physical interaction network between the products of the genes in the APha-RiB regulon

The APha-RiB regulon is the cluster C2 at DTB with $\delta = 0.2$. Each node represents a gene, and a link between any two nodes represents the existence of a physical interaction between the products of those genes, i.e. between the proteins which are encoded by those genes. A relatively highly connected sub-network of eight genes is highlighted for more discussion in the main text; this is the same sub-network highlighted in Figure 5.8.

It is interesting that, within the selected sub-network, there is a perfect one-to-one correspondence between protein-protein physical interactions and negative genetic interactions (Figure 5.8 and Figure 5.9). When this is added to their consistent co-expression over forty different and recent datasets, it can be hypothesised that they are related functionally, which can be tested in future biological studies.

5.3.2.6. APha-RiB comparison with the literature

The phenomenon of opposite co-expression of RRB and stress response genes in budding yeast was reported by various studies (Gasch, et al., 2000; Brauer, et al., 2008; Tsankov, et al., 2010; Roy, et al., 2013). As shown in Figure 5.10, the subsets of genes identified by the studies of Gasch (2000) (Gasch, et al., 2000), Brauer (2008) (Brauer, et al., 2008), and Roy (2013) (Roy, et al., 2013), and their collaborators are much larger than the APha-RiB regulon defined in our study (hundreds of genes versus 47 genes). Moreover, the largest overlap between any of those subsets of genes and APha-RiB does not reach half of the genes in APha-RiB, where the largest overlap, which is between APha-RiB and the subset identified by Gasch and colleagues (Gasch, et al., 2000), includes 22 genes. Furthermore, none of those previously reported, relatively large, subsets includes more than two of the eight genes highlighted for their importance in Figure 5.8 and Figure 5.9, and discussed below. This illustrates the novelty of this focused and specific cluster which has been found by our large scale genome-wide analysis of forty different and recent datasets.

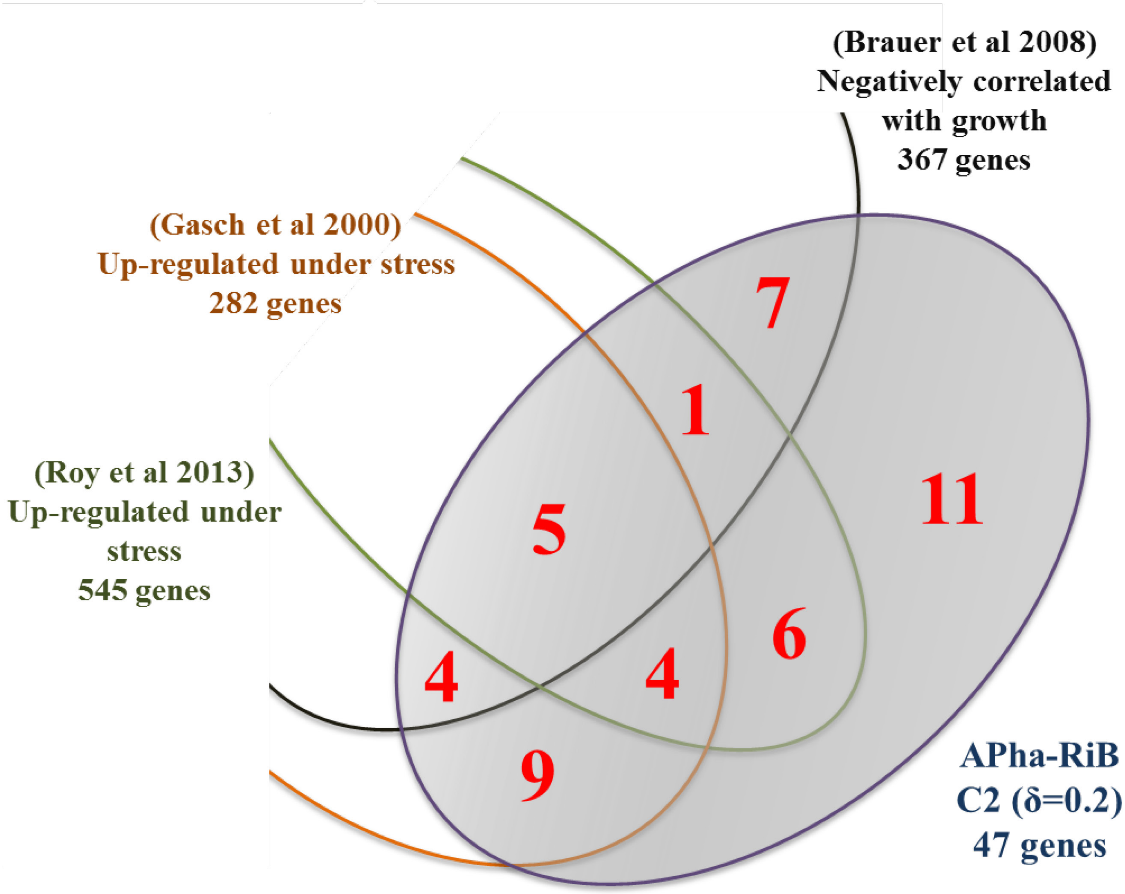

Figure 5.10. Comparison between the APha-RiB cluster and related clusters from the literature

Venn diagram showing the size of overlap between our novel APha-RiB cluster (C2 at DTB with δ = 0.2) and the subsets of genes with expression reported to be positively correlated with stress and negatively correlated with growth in three previous studies (Gasch, et al., 2000; Brauer, et al., 2008; Roy, et al., 2013).

Taken together, firstly, we have observed and reconfirmed the reciprocal behaviour of RRB and some genes participating in stress response over datasets which cover much wider conditions including ones that are not directly related to stress changes, e.g. cell-cycle datasets. Secondly, our APha-RiB subset of genes consistently reciprocally expressed with RRB largely includes genes with unknown or apparently unrelated biological processes, in addition to few genes known to participate in stress response. Thirdly, our method does not require that the microarray samples are combined into a single dataset, in contrast to the studies by Gasch (Gasch, et al., 2000) and Brauer (Brauer, et al., 2008) and their colleagues. It is therefore now possible to analyse large number of datasets in the literature in a single experiment, even if the datasets are diverse in time, location, condition, and use different microarray platforms. Finally, although a proportion of the APha-RiB genes has been explicitly associated with response to oxidative stress processes (six out of 47 genes), the processes in which the rest of the genes in APha-RiB participate are either unknown or apparently unrelated. Additionally, the forty datasets considered in this study cover a much wider range of stress and growth conditions than oxidative stress. Given that, most of the genes in APha-RiB are yet to be associated with biological processes and/or their function to be understood within the context of generic, not specific, stress response; our results suggest these areas would be the subject for fertile future investigation.

5.3.2.7. Proposed model for transcriptional regulation of RRB and APha-RiB

The temporal expression of the cluster APha-RiB (C2) in opposite direction of regulation to the RRB genes (C1), as well as the high enrichment of common motifs in the upstream DNA sequences of genes in APha-RiB (Figure 5.6), strongly support the hypothesis that genes in the subsets RRB and APha-RiB are regulated by the same biological machinery, or possibly that the transcriptional regulators for both clusters are regulated by a common regulator. Therefore, we propose an outline model of regulation for the genes included in RRB and APha-RiB clusters (Figure 5.11).

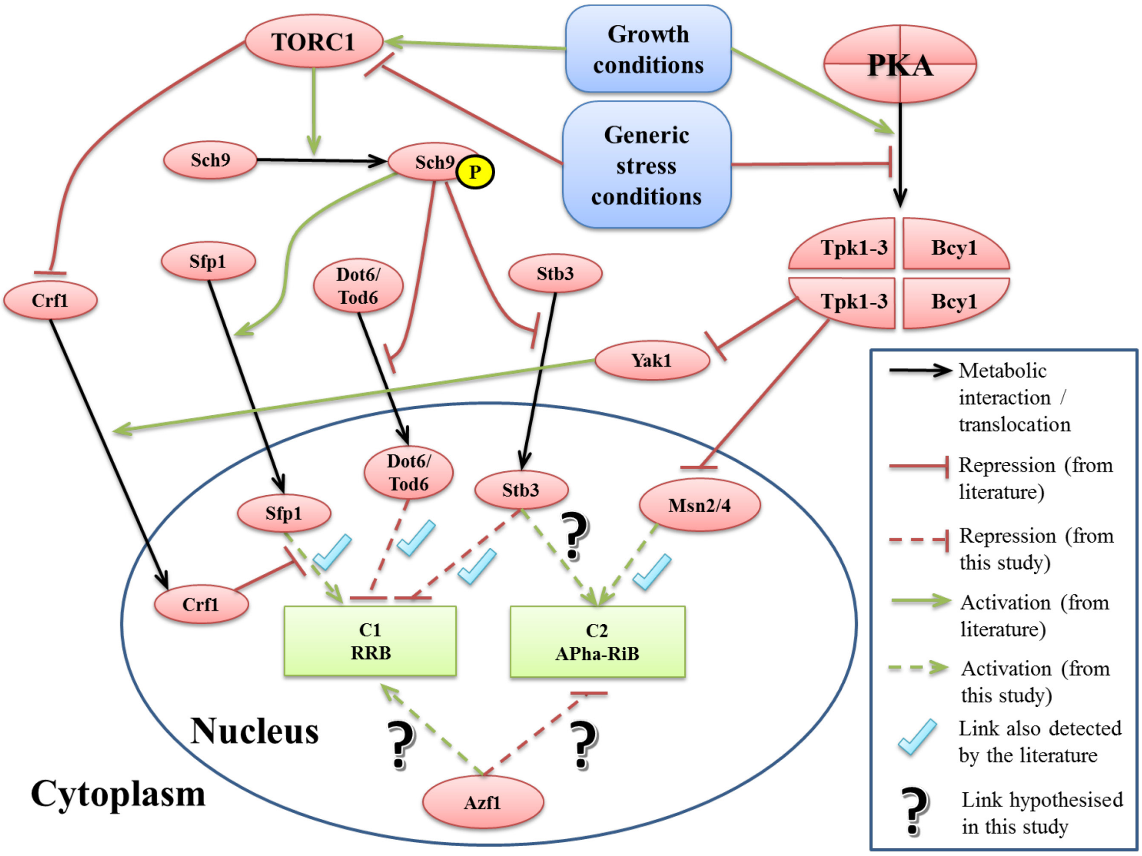
TORC1
Growth conditions
PKA
Generic stress conditions
Sch9
Sch9
P
Sfp1
Dot6/ Tod6
Stb3
Tpk1-3
Bcy1
Tpk1-3
Bcy1
Crf1
Yak1
Dot6/ Tod6
Stb3
Msn2/4
Sfp1
Crf1
C1 RRB
C2 APha-RiB
?
Nucleus
?
?
Azf1
Cytoplasm
Metabolic interaction / translocation
Repression (from literature)
Repression (from this study)
Activation (from literature)
Activation (from this study)
Link also detected by the literature
?
Link hypothesised in this study

Figure 5.11. Regulation of the RRB and the APha-RiB clusters

Ticked dashed links have been detected in this study and were also previously identified in the literature while dashed links with question marks have been only detected in this study. However, most of the previous studies consider one or few stress conditions in contrast to “generic stress conditions”. Notice that the cluster “C2 APha-RiB” is novel and that the links from the literature that point at it are based on the assumption that it is a stress response module.

The model in Figure 5.11 shows parts of the TOR and the PKA signalling pathways which are regulated by the presence of some growth factors (e.g. glucose) or the presence of some stress conditions, and then they regulate RRB and stress response modules of genes. Although we use the general terms "growth conditions" and "generic stress conditions" instead of more specific terms such as "glucose abundance", "oxidative stress", most of the previously discovered links of regulation were in the context of one or few growth conditions such as the presence of glucose (Liko, et al., 2007; Liko, et al., 2010; Dikicioglu, et al., 2011), ammonium (Dikicioglu, et al., 2011), or other specific nutrients, or to types of stress such as oxidative stress (Drobna, et al., 2012) or methyl methanesulfonate (MMS) DNA-damage stress (Gasch, et al., 2001). However, using such general terms here reflects the comprehensive nature of the data analysed by the Bi-CoPaM approach as we have been able to consider and analyse a wide range of different growth and stress conditions in a comprehensive and systematic way. Indeed, we can now reach a consensus conclusion, that up- and down-regulation of the RRB and APha-RiB clusters are influenced by a wide range of growth and stress conditions (Table 5.4).

Many of the direct regulators detected in this study by upstream sequence analysis of the RRB and the APha-RiB subsets of genes (dashed links in Figure 5.11) were also previously identified in the literature (ticked dashed links). Indeed, the regulatory links from the literature to the novel APha-RiB cluster are based on the assumption that it is a stress response subset of genes.

It could be argued that one of the two clusters actually negatively regulates the other. This seems unlikely for several reasons. First, the synchronisation between both clusters is very high such that there is insufficient phase shift between them for one to regulate the other. Second, the functionality of a transcription factor is likely to be regulated post-translationally in many ways, such as the existence of another metabolite or signal, localisation changes, or others (Liko, et al., 2010; Tkach, et al., 2012). It is doubtful that many regulators could be functionally active in a consistently similar profile for a very large number of target genes. Therefore, we would suggest that these two clusters of genes are transcriptionally regulated by common machinery rather than one of the clusters transcriptionally regulates the other.

It could also be hypothesised that the two clusters are regulated by two separate pathways that are oppositely activated in synchrony with growth and stress conditions. Though, this hypothesis necessitates that those two transcriptional regulation pathways are consistently and synchronically regulated by various types of growth and stress signals, or that those signals regulate a single signalling pathway which regulates both transcriptional regulatory machineries. In this case, the common upstream regulator of the two clusters would be a signalling pathway or the signals themselves. Although this is a possible proposal, the fact that the signals that consistently and synchronically regulate both groups are largely variant, we focus on the hypothesis that both groups are regulated by a common machinery, or that their regulatory machineries regulated by a common regulator. Indeed, the latter proposal conforms to the more general statement of Brauer and colleagues that such consistent positive or negative correlation reflects system-level regulatory mechanisms (Brauer, et al., 2008).

5.3.2.8. Potential regulators for APha-RiB and common regulators for RRB and APha-RiB

Gasch and Roy and their collaborators commonly identified the Msn2p and its paralogue Msn4p as regulators for the subsets of genes which they identified as negatively correlated with growth (Gasch, et al., 2000; Roy, et al., 2013). Gasch and colleagues also identified Yap1p as a regulator for their group (Gasch, et al., 2000) while Roy and colleagues identified Rtg1p and Adr1p (Roy, et al., 2013). Interestingly, upstream analysis for our novel cluster APha-RiB (C2) has identified Azf1p and the paralogous pair Msn2p and Msn4p as potential regulators (Figure 5.6). It is worth noticing that the three studies mutually identify Msn2p and Msn4p, which are well known for their role in stress response regulation through binding to the STRE motif (Figure 5.11) (Martínez-Pastor, et al., 1996; Schmitt & McEntee, 1996).

More interestingly, Azf1p has been identified by our results as a potential regulator for in both clusters RRB (C1) and APha-RiB (C2) (Figure 5.5, Figure 5.6, and Figure 5.11). Azf1p is a zinc-finger transcription factor, which has been predicted to have role in one of the putative stress response regulatory modules (Segal, et al., 2003; SGD, 2014). Moreover, it is exclusively localised in the nucleus and it was found to be synthesised in higher amounts under non-fermentable growth conditions (Stein, et al., 1998). By monitoring differentially expressed genes when AZF1 was knocked down, Slattery and colleagues showed that this gene's product participates in the transcription of two non-overlapping subsets of genes under two different conditions. The common aspect between these non-overlapping subsets of genes is having the motif AAAAGAAA in their promoters (Slattery, et al., 2006). Although our C2 genes at $\delta = 0.2$ are not included in any of these two subsets, the existence of the AZF1 binding site in their promoters indicates that AZF1 may regulate expression of genes in this cluster under other conditions.

Another candidate common regulator is Stb3p (Figure 5.11), which binds to the consensus motif TGAAAAA (Liko, et al., 2010; Liko, et al., 2007; Zhu, et al., 2009). This motif largely overlaps with the RRPE motif found in the upstream sequences of the RRB genes in our results, as identified by the TOMTOM tool (Figure 5.5 (B)). Although not identified by the TOMTOM tool as a potential binding transcription factor, its binding motif TGAAAA largely overlaps with the part of the motif C2-1 (Figure 5.6). Moreover, Stb3p overexpression was shown to increase resistance to oxidative stress (Drobna, et al., 2012) and to result in down-regulation of ribosome biogenesis genes (Liko, et al., 2010; Liko, et al., 2007; Zhu, et al., 2009), and Liko and colleagues also predicted that Stb3p would be expected to regulate transcription of other unknown sets of genes positively (Liko, et al., 2010; Liko, et al., 2007).

The evidence for Azf1p or Stb3p acting as a transcription activator and/or repressor with relation to both groups of genes – RRB genes (C1), and APha-RiB genes (C2) is unclear. Nevertheless, there are enough observations to speculate that one of them or both of them may play a role in the mutual transcriptional regulation of both RRB and APha-RiB. The molecular mechanism(s) and significance of those transcription factors in this context remain to be established.

5.3.2.9. Experiments with different numbers of clusters (K values)

We have repeated the Bi-CoPaM experiment over the same datasets but with different *K* values other than sixteen, that is, with different numbers of clusters. We tried the *K* values of 8, 9, 10, 18, 24, 30, and 40. At all of the given *K* values, the cluster RRB was found as the absolutely tightest cluster with very high similarity in its gene content to the cluster found at *K* = 16. At the *K* values of 8, 9, and 10, the results have shown that the second tightest cluster is similar to the APha-RiB regulon found in this study, while at the *K* values of 18 and 24, it was split into two smaller clusters. Moreover, at the *K* values of 30 and 40, many other small tight clusters appeared but many of them are redundant in terms of their expression profiles and should be rather combined. Interestingly, no other significant cluster found in any of those results. This experiment shows that our proposed approach of applying the Bi-CoPaM method to genome-wide datasets is robust over a wide range of *K* values.

5.3.3. Discussion and conclusions

We have applied the Bi-CoPaM method over genome-wide data from forty microarray datasets with wide range of different biological contexts and experimental conditions in order to identify the subsets of budding yeast genes that are most consistently co-expressed. We found two clusters of genes that have significant consistency of co-expressions, which we have labelled as RRB (C1) and APha-RiB (C2). These two clusters preserved their status as the tightest two clusters at varying values of *K*, which shows their importance as well as the robustness of the proposed Bi-CoPaM approach. By GO term analysis, C1 has been found to be highly enriched with ribosome biogenesis and rRNA processing (RRB) genes. On the other hand, most of the genes included in C2 have unknown or apparently unrelated functions.

Finding RRB genes (C1) in the tightest cluster by this completely unsupervised approach, confirms not only that these genes are consistently co-expressed under various conditions (Wade, et al., 2006), but also that they are the most consistently co-expressed genes across the whole genome. Additionally, our C1 cluster includes few genes with unknown processes that may be worthy of biological investigation.

The most interesting cluster of genes in our results appears to be C2, and this is for three main reasons – first, these genes are mostly unknown or apparently unrelated to each other, despite the fact that they are the second most consistently co-expressed subset of genes in budding yeast; second, their average expression profiles show consistently anti-phase (opposite) expression to the average expression profiles of RRB genes (C1) across all of the forty datasets; and third, significant genetic and protein-protein physical interactions have been reported between them by high-throughput studies in the literature. These observations lead us to label C2 as the subset of genes in *anti-phase with ribosome biogenesis (APha-RiB)*, to suggest that many of the unknown genes in APha-RiB (C2), such as YIR016W, may participate in different generic, in contrast to specific, stress response mechanisms, and to suggest that RRB genes (C1) and the APha-RiB genes (C2) may be transcriptionally regulated by common machinery or that their regulation machineries may be controlled by common post-translational regulators. We have identified potential factors that might be involved in such reciprocal regulation, for example Azf1p and Stb3p.

This study has yielded globally consistent co-expression in budding yeast and produced new, focused insights for future work to elucidate and confirm the components of the common regulatory machinery for RRB and APha-RiB, and to define the function of poorly characterised genes in both clusters. The results from the application of the Bi-CoPaM method to yeast datasets strongly suggests that it may be helpful for the analysis of other groups of microarray datasets from other species and systems for the exploration of global genetic co-expression.

Chapter 6

Red Blood Cells Production (Erythropoiesis) Data Analysis

6.1. Introduction to erythropoiesis

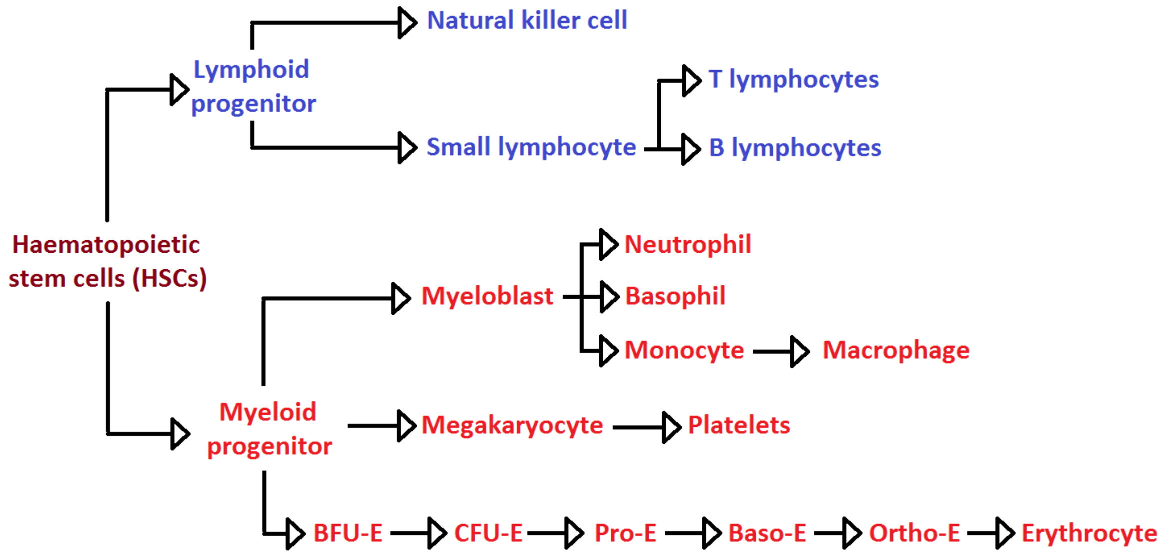

Figure 6.1. Tree of haematopoietic stem, progenitor, and mature cells in mammals

Erythropoiesis, that is, the production of red blood cells (RBCs), is a key molecular biological process in human bodies. The process starts from one type of stem cells known as *haematopoietic stem cells (HSCs)*. Those are cells which can be developed to be specialised, eventually, to be one of the different types of blood cells such as white blood cells (e.g. killer cells, and T and B lymphocytes), red blood cells (erythrocytes), platelets, or others (Figure 6.1). Many intermediate cell types, known as *progenitors*, are produced in the way from the HSCs to the final mature cells. When a stem cell or a progenitor produces a daughter cell of a downstream cell type, it is said that it has *differentiated* (e.g. myeloid progenitors producing BFU-E cells in Figure 6.1). However, some stem cells and progenitors may produce daughter cells of their same type in order to increase the number of cells of that intermediate stage (e.g. a BFU-E cell producing daughter BFU-E cells); this is known as *self-renewal*, or *proliferation*. Some progenitors, like the myeloid progenitors, have potential to differentiate to different types of final mature cells, while some other progenitors, like the BFU-E cells, are committed to differentiate to single final mature cell type. Erythropoiesis in Figure 6.1 is the branch that starts at HSCs and ends at the erythrocytes (RBCs).

Various groups of genes and sub-processes are involved in erythropoiesis at different stages. For example, the setup of the genome expression at the early stages should be in a way that pushes the cells to differentiate and commit to the erythropoietic branch rather than to any other potential branch of cells. Then, in the burst-forming unit erythroids (BFU-E) stage and the colony-forming unit erythroids (CFU-E), a lot of proliferation takes place in order to increase the number of produced cells to the required ranges. This is because the early stage cell types, the HSCs and the myeloid progenitors, are usually of small quantities. Therefore, cell-cycle and proliferation genes should be very active in these proliferative stages. Towards the end of erythropoiesis, haem, which is a key component of the oxygen carrying molecule haemoglobin, should be synthesised in large quantities by the products of the haem biosynthesis genes. Redness of those cells appears only at the last stages due to the accumulation of the red iron-containing haemoglobin molecules. Moreover, and while heading towards the late stages, the cell-cycle should be arrested, transcription and translation processes should be shut down gradually, the genetic material and the nucleus should be extremely condensed, and finally the nucleus should be expelled from the cell. Expelling the nucleus from the cell at the end is known as *enucleation*, and produces *reticulocytes*, which represent the last stage of differentiation just before the final mature erythrocytes (RBCs). Mammal RBCs are enucleated, that is, are nucleus-free, while other blood-containing animals, like birds, are not. Enucleation is important because it allows the RBCs to be thinner and flexible for bending, which in its turn allows them to flow through the thin blood vessels, the capillaries, whose diameters are smaller than the diameter of the RBCs.

The regulatory machineries leading to the strong impulse of proliferation at the BFU-E and the CFU-E stages, as well as enucleation towards the end, are poorly understood. Additionally, various aspects related to haem biosynthesis such as the signalling pathways through which haem is imported to the mitochondrion are yet to be elucidated. Moreover, not all of the molecular factors that cause blood disorders like anaemia, myelodysplasia, or myeloproliferative diseases have well been described (Abu-Jamous, et al., 2015c). Having said that, better understanding of erythropoiesis is indeed a hot area of research.

This chapter describes the collective analysis of eight human and murine erythropoietic datasets by the Bi-CoPaM method. Mice are mammals whose RBCs are enucleated and whose erythropoiesis has a lot of similarity to the human one. This justifies this collective analysis despite the few differences between the two erythropoietic systems.

6.2. Datasets and experimental design

Eight human and murine erythropoiesis datasets were considered in our comprehensive study and are listed in Table 6.1. The symbols A to H in the first column of the Table will hereinafter be used as unique identifiers for these eight datasets. The second column shows the Gene Expression Omnibus (GEO) accession numbers which can be used to freely access the datasets on the online databases of the U.S. National Centre for Biotechnology Information (NCBI). In some cases, the samples included under the same GEO accession number can be split into more than one independent series of samples. In these cases, we consider each such subset of samples as a separate dataset; these cases are denoted by an asterisk symbol at the end of the GEO accession number. The third and the fourth columns of the Table respectively show the species (human or mouse) and the year in which the dataset was published. The fifth column shows the number of time points or stages represented by the samples in the dataset. Biological and technical duplicates for the same time point or stage are considered as one time point or stage. The sixth and the seventh columns respectively show a brief description of the dataset and a reference to the study in which it was published.

Table 6.1. Summary of eight human and murine erythropoiesis datasets

ID	GEO Accession	Species	Year	N	Description	Reference
A	GSE22552	Human	2011	4	Human maturing erythroblasts (Oxford)	(Merryweather-Clarke, et al., 2011)
B	GSE35292*	Mouse	2012	3	B6 mouse hematopoietic development	(Walasek, et al., 2012)
C	GSE35292*	Mouse	2012	3	D2 mouse hematopoietic development	(Walasek, et al., 2012)
D	GSE20391	Mouse	2010	5	Mouse primary fetal liver terminal erythroid differentiation	(Hattangadi, et al., 2010)
E	GSE18042	Mouse	2009	6	Erythroid differentiation: G1E model	(Cheng, et al., 2009)
F	GSE4655	Human	2006	6	In vitro human adult erythroid differentiation (Keller)	(Keller, et al., 2006)
G	GSE36994*	Human	2012	4	Human fetal erythropoiesis	(Xu, et al., 2012)
H	GSE36994*	Human	2012	4	Human adult erythropoiesis	(Xu, et al., 2012)

* These datasets include more than one time series samples and thus they have been considered separately as multiple datasets; see the description of each of them in the Description column.

We identified the genes commonly included in all of the eight datasets; genes from different species and / or different microarray platforms are considered similar if they are mapped to the same NCBI homologous group identifiers. If multiple probes from the same dataset were found to be mapped to the same homologous group, the one with the highest mean expression values is considered while the others are filtered out. The result of this filtering is the inclusion of 13,269 genes with unique homologous group identifiers common to human and mice, and found represented by probe-sets in all of datasets. The homologous group identifiers for these genes, as well as their probe-set identifiers and gene names in each of the eight datasets, are provided in the Supplementary File 'Erythropoiesis/S1'.

We applied the Bi-CoPaM over those 13,269 genes from the eight datasets (Abu-Jamous, et al., 2013a). This has been done by firstly applying k-means, self-organising maps (SOMs), and hierarchical clustering (HC) with Ward's linkage to each one of these eight datasets with $K = 9$, then combining all of the clustering results into a single fuzzy consensus partition matrix (CoPaM), and finally binarising the CoPaM by the difference threshold binarisation (DTB) technique with δ values varying from zero to unity. Prior to clustering analysis, we normalised the datasets by quantile normalisation (Bolstad, et al., 2003), then each gene's expression profile was shifted and scaled to have a zero-mean and a unity standard deviation (Quackenbush, 2002).

6.3. Results and discussion

6.3.1. Bi-CoPaM clustering results

Table 6.2 shows the numbers of genes included in each of the nine clusters at each of the considered DTB δ values. The clusters were ordered from the *tightest* to the *widest* such that the cluster which preserves at least 15 genes up to a higher value of δ is considered tighter, and if two clusters do so up to the same value of δ then the cluster which includes more genes at that δ value is considered tighter. After ordering, the clusters were labelled as C1 to C9. Supplementary File 'Erythropoiesis/S1' shows the lists of genes included in each of these nine clusters at all of those δ values.

Table 6.2. Numbers of genes included in each of the nine erythropoietic clusters at varying δ values

Tightness	δ	Cluster								
		C1	C2	C3	C4	C5	C6	C7	C8	C9
Complementary	0.0	4642	2257	2988	1683	913	937	891	643	201
	0.1	2650	1142	1555	513	225	220	177	97	11
	0.2	1674	702	952	196	74	62	55	20	1
	0.3	780	358	425	38	9	8	12	4	0
	0.4	466	192	222	8	0	2	3	0	0
	0.5	271	100	98	0	0	0	1	0	0
	0.6	123	33	21	0	0	0	0	0	0
	0.7	66	12	7	0	0	0	0	0	0
	0.8	27	3	1	0	0	0	0	0	0
	0.9	5	1	0	0	0	0	0	0	0
Tightest	1.0	1	0	0	0	0	0	0	0	0

6.3.2. MSE analysis

MSE values were calculated for each of the clusters C1 to C7 at each of the DTB δ values based on their genes' profiles in each of the eight datasets independently. The average MSE values over the eight datasets for each of these seven clusters are plotted in Figure 6.2 (a) versus the DTB δ values. Figure 6.2 (b) shows the number of genes included in each of these clusters at each of the δ values; the logarithmic scale at the vertical axis is for the clarity of presentation. Missing values in both sub-plots represent empty clusters.

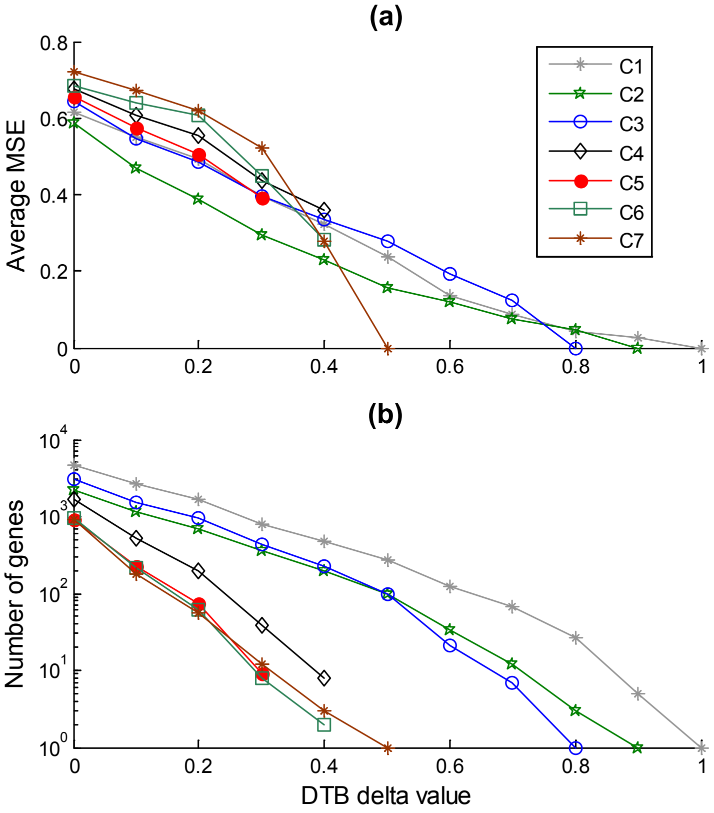

Figure 6.2. MSE and cluster size analysis for erythropoiesis clusters

(a) Average MSE values for the first seven clusters C1 to C7 over the eight datasets and (b) clusters' sizes plotted versus all of the considered DTB δ values.

Both Figure 6.2 (a) and Figure 6.2 (b) have been analysed in tandem in the view of minimising the average MSE values while maximising the number of genes included. We have followed a systematic approach to select one 'representative' instance for each of the first five clusters, that is, one δ value for each. The approach is to select the δ value below which a significant increase in the average MSE value occurs with no parallel increase in the number of genes included; indeed the cases in which the clusters include very few genes are not considered.

Based on this approach, the cluster C1 shows a significant increase in average MSE value when the δ value is decreased from 0.6 to 0.5 (from 0.14 to 0.24), and it has a sufficient number of genes at 0.6 (123 genes); therefore, the representative instance of C1 is at δ =

0.6. The same observation can be seen for the cluster C2 at $\delta = 0.5$. When the tightness of the cluster C3 is decreased from $\delta = 0.5$ to 0.4, the number of genes included in it increases significantly from 21 to 98 genes with no large difference in MSE values; thus, the representative of C3 has been chosen at 0.5. The cluster C4 has higher MSE values than the first three and lower numbers of genes than them. Though, we consider its representative at $\delta = 0.3$ because at the tighter level of $\delta = 0.4$ it only includes eight genes, and at the wider level of $\delta = 0.2$, its MSE value is significantly higher (0.55). C5 might be considered insignificant enough to be filtered out, but its case at $\delta = 0.2$ is not very different from the considered C4 at $\delta = 0.3$; it has a slightly higher MSE value (0.51 compared to 0.44) but with significantly more genes (74 compared to 38), and therefore we have considered the representative case of C5 to be at $\delta = 0.2$.

In contrast to the first five clusters, the clusters C6 and C7 show significantly higher average MSE value and lower numbers of genes. Moreover, the clusters C8 and C9, which have not been included in this Figure for clarity of demonstration purposes, include significantly lower numbers of genes than the first five clusters and become empty at relatively low δ values (Table 6.2). Therefore, we further focus our analysis on the clusters C1 to C5. The selected representative instances for these five clusters are provided in Table 6.3, and will hereinafter be labelled as C1* to C5*.

Table 6.3. Selected erythropoiesis clusters' representatives

Cluster	**C1***	**C2***	**C3***	**C4***	**C5***
DTB δ value	0.6	0.5	0.5	0.3	0.2
Average MSE	0.14	0.16	0.28	0.44	0.51
Number of genes	123	100	98	38	74

Considering the mean MSE value of the genes in a cluster over their profiles in all of the eight datasets has been useful for coarse-grained filtering. We further investigate the quality of the representative instances of the first five clusters C1* to C5* in each of the eight datasets individually. Figure 6.3 shows the MSE values for the clusters C1* to C5* versus the eight datasets A to H (see Table 6.1 for datasets details). It is very clear in this Figure that the MSE value of the clusters C3*, C4*, and C5* in the dataset (E) are many folds worse (higher) than the average of MSE values shown in this Figure. These clusters also show high MSE values in the dataset (A) but less extreme than in the dataset (E). These observations indicate that the level of co-expression of these clusters in these specific datasets is not as tight as in the others, which should be taken into consideration while analysing their expression profiles.

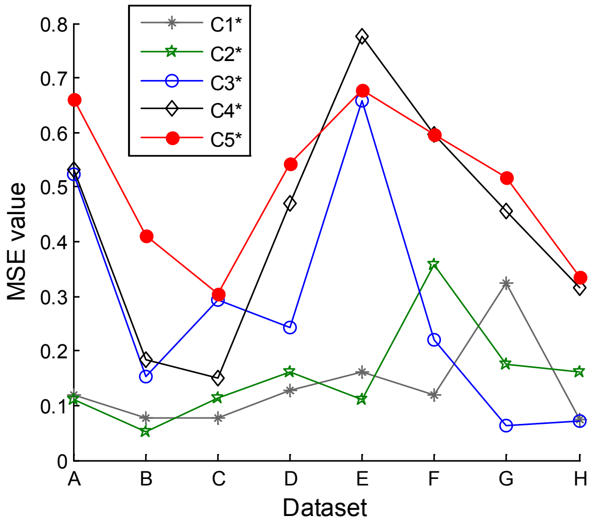

Figure 6.3. MSE values for the cores of the clusters C1* to C5* plotted versus the eight erythropoiesis datasets A to H

6.3.3. Comparison with datasets from wider conditions

It is useful to distinguish between the clusters of genes that are specifically consistently co-expressed under erythropoiesis and the clusters that show such consistency in co-expression over wider range of conditions. Therefore, the MSE values were calculated for the representative instances of the first five clusters C1* to C5* as well as C6 at $\delta = 0.2$ (which was labelled as C6* for this section's purposes) over 90 randomly selected human and murine microarray datasets. The 90 datasets were randomly selected from the thousands of datasets available at the GEO repository based on the three microarray platforms with the GEO accession numbers GPL570, GPL6887, and GPL1261.

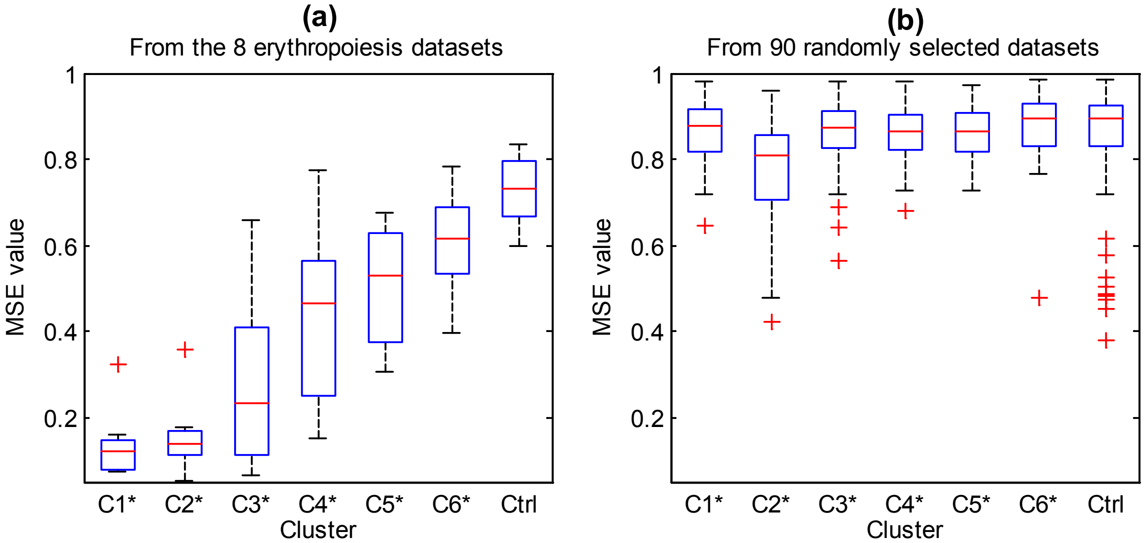

Figure 6.4. Box plots comparing the core clusters C1* to C6* in (a) the eight erythropoiesis datasets and (b) 90 randomly selected datasets

The rightmost box in both sub-plots is a control. The control box in (a) includes the MSE values for 100 randomly generated clusters with average number of genes of 70 based on the eight considered datasets. The control box in (b) includes the MSE values for ten randomly generated clusters with average number of genes of 70 in each of the 90 randomly selected microarray datasets; for each of the 90 datasets, ten different randomly generated clusters were considered. Thus, the control in (a) has 800 MSE values and the control in (b) has 900 MSE values.

Figure 6.4 shows box plots for the MSE values of each of the six clusters in each of the eight datasets (a) as well as in each of the 90 randomly selected datasets (b) respectively. An additional control has been added to the box plots showing the MSE values of randomly generated clusters of an average of 70 genes per each. It can be seen in this Figure that the MSE values of the clusters C1* to C6* show a significant difference from the control in the eight considered datasets when compared to the 90 randomly selected datasets with obvious distinction of C1* and C2*. This indicates that they are specifically highly co-expressed under erythropoiesis and not under a wider range of more general conditions. Despite that, the cluster C2* can be seen to have slightly lower MSE values than the control in the 90 randomly selected datasets, which indicates that although it is significantly lower in erythropoiesis datasets, it still preserves some co-expression in other conditions.

6.3.4. Average expression profiles

Figure 6.5 shows the average normalised expression profiles of the genes included in each of the five representative clusters C1* to C5* from each of the eight datasets (A) to (H). This Figure shows these average profiles in a grid of plots where each column of the grid represents a cluster and each row represents a dataset. In order to take the analysis of these profiles further, the erythropoietic stages to which the samples in each of the eight datasets belong were investigated, and our estimations for these are provided in Figure 6.6. This Figure shows a chronological order of the erythropoietic stages on the horizontal axis and the datasets (A) to (H) on the vertical axis. Different symbols are used for different datasets.

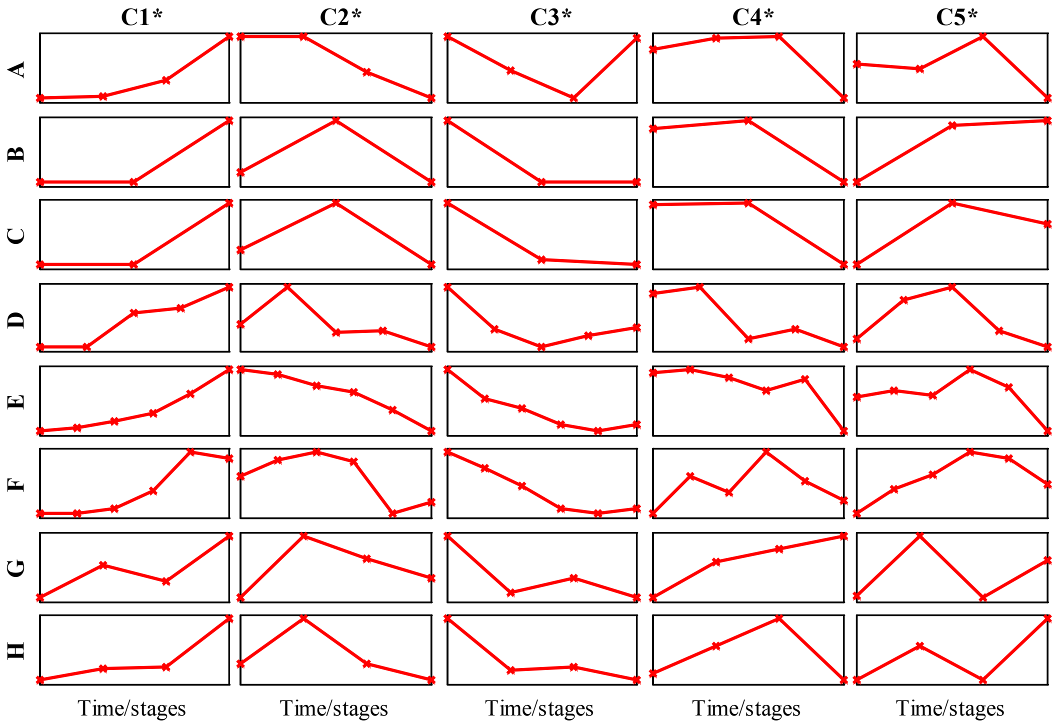

Figure 6.5. Average expression profiles for the core clusters C1* to C5* in each of the eight erythropoiesis datasets

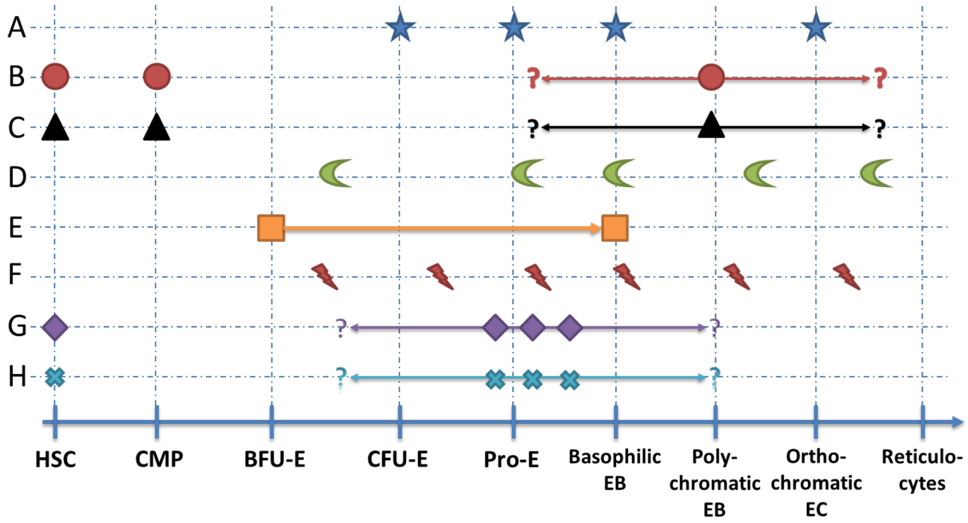

Figure 6.6. Estimation of the erythropoietic stages to which the samples of the eight datasets belong

By examining the profiles in Figure 6.5 in the light of the information in Figure 6.6, we can extract the generic behaviour of each of the five clusters C1* to C5* over the erythropoietic stages, which is demonstrated in Figure 6.7.

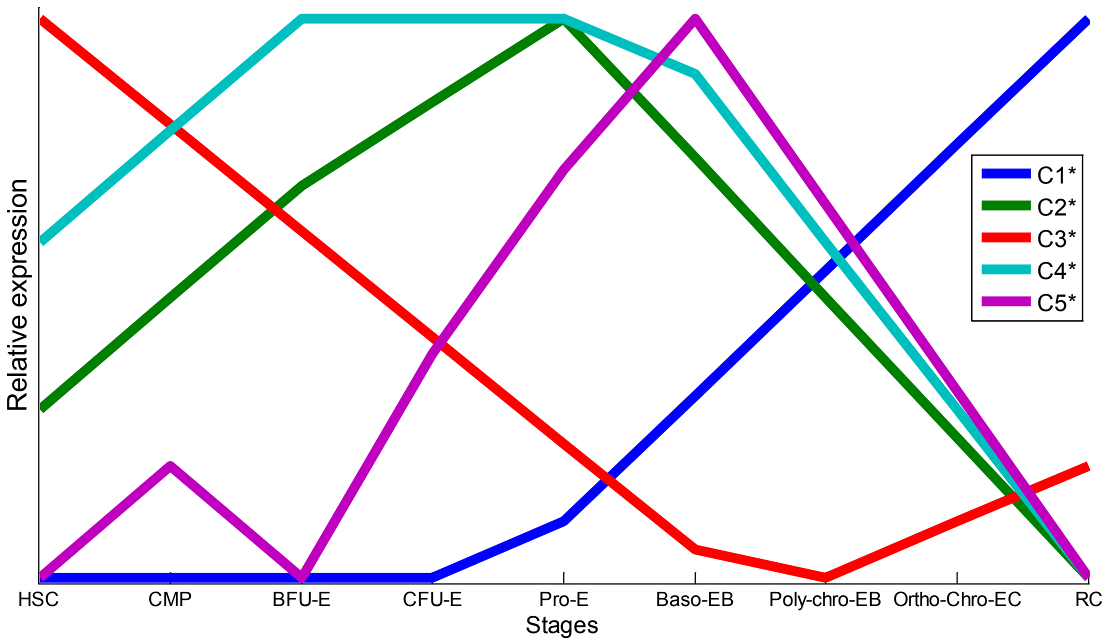

Figure 6.7. Estimated summarisation of the average profiles of the core clusters C1* to C5* over erythropoietic stages

C1* preserves very low expression up to colony forming-unit erythroids (CFU-E) and is gradually up-regulated thereinafter towards the latest stages of erythropoiesis. C2* starts with a moderate expression at the early stage of haematopoietic stem cells (HSCs) and is up-regulated to peak at pro-erythroblasts (Pro-E); then, it is down-regulated to reach its minimum expression at the latest stages. C3* starts with an expression-peak at HSCs and is constantly down-regulated up to the stages of basophilic erythroblasts and polychromatic erythroblasts where it starts a slight up-regulation until the latest stages. C4* starts with moderate to high expression values at HSCs and is then up-regulated to plateau from the early committed progenitors BFU-E and CFU-E to the mid-stages of pro-erythroblasts and basophilic erythroblasts; after that, it is down-regulated to reach its minimum expression at the final stages of erythropoiesis. C5* starts from a very low expression at HSCs and increases gradually to peak at the stage of basophilic erythroblasts; it subsequently drops to low values at the terminal stages. It can be seen from Figure 6.5 and Figure 6.7 that these observations are consistent across almost all of the eight datasets with minor exceptions.

The summary of this profile-analysis is that these five clusters show their major peak expression values at five different stages of development, namely and chronologically at the stages of HSCs, pro-erythroblasts, approximately BFU erythroblasts to basophilic erythroblasts, basophilic erythroblasts, and orthochromatic erythrocytes for the clusters C3*, C2*, C4*, C5*, and C1*, respectively. Another notable difference is in their relative expression at HSCs where some clusters show moderate expression values while others show very low ones.

6.3.5. GO term analysis

Table 6.4 and Table 6.5 respectively show the most enriched GO biological processes and cellular components in the first five clusters C1 to C5 at varying δ values. Complete GO analysis results are provided in the Supplementary Tables Erythropoiesis/S2 to S11.

Table 6.4. Most enriched GO process terms in the erythropoietic clusters C1 to C4 at various levels of tightness

	GO process	Back. frequency	$\delta = 0.3$		$\delta = 0.4$		$\delta = 0.5$		$\delta = 0.6$	
			Freq.	P-val.	Freq.	P-val.	Freq.	P-val.	Freq.	P-val.
C1	Heme biosynthetic proc.	16/13269	8/780	E-6	5/466	E-4	5/271	E-5	2/123	E-3
	Autophagy	50/13269	14/780	E-7	11/466	E-7	8/271	E-6	3/123	E-2
	protein K63-linked deubiquitination	15/13269	6/780	E-4	4/466	E-3	4/271	E-4	4/123	E-6
	protein K48-linked deubiquitination	11/13269	5/780	E-4	3/466	E-3	3/271	E-3	3/123	E-4
	protein ubiquitination	213/13269	29/780	E-5	21/466	E-5	11/271	E-3	8/123	E-4
	protein phosphorylation	298/13269	35/780	E-5	24/466	E-4	14/271	E-3	5/123	0.14
	cell cycle arrest	113/13269	19/780	E-5	15/466	E-6	11/271	E-5	8/123	E-6
	negative regulation of cell proliferation	282/13269	28/780	E-3	21/466	E-4	13/271	E-3	8/123	E-3
	unknown process	2326/13269	128/780	0.81	85/466	0.36	49/271	0.43	23/123	0.40
C2	ribosome biogenesis	26/13269	10/358	E-10	9/192	E-11	6/100	E-8	0/33	1.0
	Gene expression	543/13269	54/358	E-17	27/192	E-8	9/100	E-2	2/33	0.39
	RNA splicing	193/13269	29/358	E-14	11/192	E-4	4/100	E-2	2/33	E-2
	rRNA processing	71/13269	16/358	E-11	13/192	E-11	7/100	E-6	4/33	E-5
	tRNA processing	39/13269	9/358	E-7	6/192	E-5	5/100	E-5	2/33	E-3
	mRNA processing	162/13269	22/358	E-10	13/192	E-7	6/100	E-3	2/33	E-2
	translation	167/13269	22/358	E-10	14/192	E-7	7/100	E-4	1/33	0.34
	unknown process	2326/13269	59/358	0.72	34/192	0.50	19/100	0.39	6/33	0.53
C3	signal transduction	840/13269	50/425	E-5	32/222	E-5	15/98	E-3	3/21	0.14

	small GTPase mediated signal transduction	269/13269	29/425	E-8	20/222	E-8	7/98	E-3	3/21	E-3
	apoptotic process	570/13269	38/425	E-5	25/222	E-5	11/98	E-3	4/21	E-2
	blood coagulation	387/13269	28/425	E-5	18/222	E-5	6/98	E-2	2/21	0.12
	immune response	242/13269	19/425	E-4	13/222	E-4	4/98	0.10	1/21	0.32
	cell proliferation	288/13269	20/425	E-4	9/222	E-2	2/98	0.63	1/21	0.57
	unknown process	2326/13269	70/425	0.74	30/222	0.96	19/98	0.35	5/21	0.30
C4	viral transcription	31/13269	2/38	E-3	0/8	1.0				
	translation	167/13269	3/38	E-2	0/8	1.0				
	glycolysis	32/13269	1/38	E-2	0/8	1.0				
	unknown process	2326/13269	10/38	0.12	3/8	0.15				
C5	G1/S transition of mitotic cell cycle	117/13269	2/9	E-3						
	DNA replication	120/13269	2/9	E-3						
	mitotic cell cycle	271/13269	3/9	E-4						
	DNA repair	231/13269	2/9	E-2						
	unknown process	2326/13269	1/9	0.82						

Table 6.5. Most enriched GO component terms in the erythropoietic clusters C1 to C4 at various levels of tightness

	GO component	Back. frequency	$\delta = 0.3$		$\delta = 0.4$		$\delta = 0.5$		$\delta = 0.6$	
			Freq.	P-val.	Freq.	P-val.	Freq.	P-val.	Freq.	P-val.
C1	autophagic vacuole	23/13269	10/780	E-7	7/466	E-6	5/271	E-5	2/123	E-2
	cortical cytoskeleton	17/13269	7/780	E-5	6/466	E-5	6/271	E-7	3/123	E-4
	endosome membrane	125/13269	13/780	E-2	8/466	E-2	5/271	0.11	4/123	E-2
	early endosome	109/13269	12/780	E-2	8/466	E-2	7/271	E-3	5/123	E-3
	late endosome	67/13269	7/780	E-2	6/466	E-2	5/271	E-2	4/123	E-3
	Golgi apparatus	567/13269	41/780	E-2	27/466	E-2	21/271	E-3	8/123	0.16
	unknown component	1258/13269	71/780	0.66	49/466	0.24	31/271	0.16	17/123	E-2
C2	mitochondrion	984/13269	82/358	E-21	41/192	E-10	22/100	E-6	10/33	E-5
	nucleolus	1249/13269	102/358	E-25	64/192	E-20	40/100	E-16	11/33	E-4
	nucleoplasm	777/13269	58/358	E-12	25/192	E-4	7/100	0.37	1/33	0.86
	spliceosomal complex	67/13269	18/358	E-13	10/192	E-8	5/100	E-4	1/33	0.15
	Many other mitochondrial components (membrane, matrix, nucleoid, etc.)									
	unknown component	1258/13269	24/358	0.98	12/192	0.96	6/100	0.92	2/33	0.83
C3	plasma membrane	2332/13269	111/425	E-6	71/222	E-7	28/98	E-3	5/21	0.30
	lysosome	156/13269	13/425	E-3	9/222	E-3	3/98	0.11	1/21	0.22
	cytosol	1923/13269	91/425	E-5	52/222	E-4	22/98	E-2	7/21	E-2
	unknown component	1258/13269	30/425	0.97	15/222	0.94	7/98	0.83	4/21	0.13
C4	mitochondrion	984/13269	7/38	E-2	1/8	0.46				
	ribosome	94/13269	2/38	E-2	0/8	1.0				
	unknown component	1258/13269	3/38	0.71	1/8	0.55				
C5	nucleoplasm	777/13269	4/9	E-3						
	unknown component	1258/13269	0/9	1.0						

As can be seen in Table 6.4, the cluster C1 is highly enriched with processes related to haem biosynthesis, protein ubiquitination, deubiquitination, phosphorylation, cell-cycle arrest, and negative regulation of cell proliferation. It can also be seen in Table 6.5 that it is enriched with various cellular components including autophagic vacuoles, cortical cytoskeleton, endosome membrane, early and late endosomes, and Golgi apparatus. The second cluster, C2, is more focused as it is mainly enriched with RNA- and ribosome-related processes with high enrichment in mitochondrial and nuclear components. The third cluster, C3, is enriched with processes related to signal transduction, apoptosis, blood coagulation, immune response, and cell proliferation with high component enrichment in the plasma membrane, lysosomes, and cytosol. C4 is highly enriched with genes participating in viral transcription, translation, and glycolysis with component enrichment in the mitochondrion and ribosomes. C5 is highly enriched with cell-cycle and DNA metabolism related processes whose component enrichment is focused in the nucleoplasm.

6.3.6. Upstream sequence analysis

We have used the Promoter Analysis and Interaction Network Toolset (PAINT) (Vadigepalli, et al., 2003; PAINT, 2013) to mine the 2,000 upstream DNA sequence base-pairs of the genes in the five representative clusters C1* to C5* for enriched transcription factors-binding sites. We chose the complete list of human promoter sequences available in PAINT's database as our option for the reference list, and we considered FDR-adjusted p-values as the enrichment metric.

Table 6.6. TF binding sites enriched in the five erythropoietic clusters C1* to C5*

Included binding sites are those with positive FDR-adjusted p-values ≤ 0.1.

TF	TF-binding site	C1*	C2*	C3*	C4*	C5*
AHR, ARNT, HIF1A	AhR, Arnt, HIF-1/V$AHRHIF_Q6		×			
ATF6	ATF6/V$ATF6_01	×				
CACD	CACD/V$CACD_01					
CREB1	CREB/V$CREB_02	×				
CREB1	CREB/V$CREB_Q4_01	×				
EGR/KROX family	KROX/V$KROX_Q6	×		×		
ETS1	c-Ets-1 p54/V$CETS1P54_03	×	×			
ETS1	c-Ets-1(p54)/V$CETS1P54_02		×		×	
GABPA, GABPB2	GABP/V$GABP_B	×	×	×		
GTF3A	AP-2/V$AP2_Q6	×				
GTF3A	AP-2/V$AP2_Q6_01	×		×		
HIC1	HIC1/V$HIC1_02					
HIF1A	HIF1/V$HIF1_Q3					
MAZ	MAZ/V$MAZ_Q6			×		
MYB	v-Myb/V$VMYB_02		×			
PAX3	Pax-3/V$PAX3_B	×		×		
PAX5	Pax-5/V$PAX5_02		×			
PAX8	Pax-8/V$PAX8_01					
RFX1	RFX/V$RFX_Q6		×	×		
SP1	Sp1/V$SP1_Q2_01	×	×	×		
STRA13	Stra13/V$STRA13_01	×				
TFAP2A	AP-2alpha/V$AP2ALPHA_01	×		×		
TFCP2	CP2/LBP-1c/LSF/V$CP2_02	×				
TFDP1	E2F/V$E2F_Q6_01	×	×			
Unknown TF	E2F/V$E2F_03	×				
Unknown TF	ETF/V$ETF_Q6	×	×	×		
Unknown TF	Tax/CREB/V$TAXCREB_01	×				
Unknown TF	Tax/CREB/V$TAXCREB_02		×			×
WT1	WT1/V$WT1_Q6	×		×		
ZBTB14 (ZFP161)	ZF5/V$ZF5_B	×	×	×		×
ZBTB7A	LRF/V$LRF_Q2	×				
ZBTB7B	CKROX/V$CKROX_Q2	×				

Table 6.6 shows the most enriched transcription factors-binding sites in C1* to C5*. The first and the second columns show the names of the transcription factors and their binding sites. The third to the seventh columns represent the clusters C1* to C5* where a cross (×) sign indicates the enrichment of the corresponding binding site (row) in the upstream sequences of the genes in the corresponding cluster (column) with a positive FDR-adjusted p-value less than or equal to 0.1.

The correlation was then investigated between the average expression profiles for each of the clusters and the expression profiles for the transcription factors (TFs) whose binding sites are enriched in these clusters' genes' upstream sequences. This was done for these profiles based on each of the eight datasets. Few TFs in Table 6.6 are not represented by any probe-set in some (up to two) of the eight datasets; in these cases, these TFs' profiles in those datasets in which they are represented were considered. Figure 6.8 shows the results of this investigation in the form of box plots.

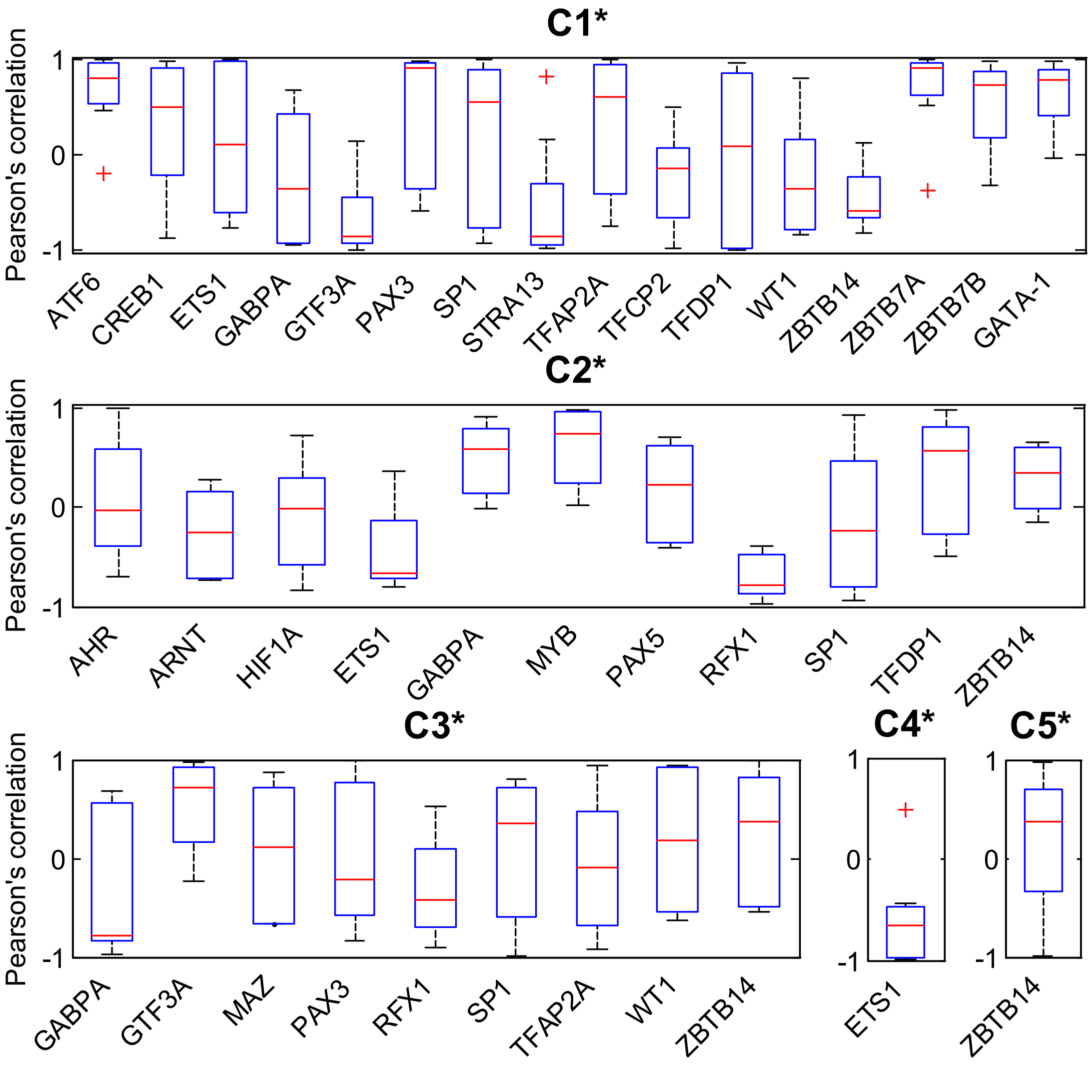

Figure 6.8. Correlation between the erythropoietic clusters C1* to C2* and the TFs whose binding sites found highly enriched in their upstream sequences

Five sub-plots are shown in Figure 6.8 for the five clusters C1* to C5*. Each single box in any of these box plots represents the Pearson's correlation values between the expression profile of the corresponding TF and the average expression profile of the corresponding cluster based on each of the eight clusters (or less than eight if not represented in some of the datasets).

It can be seen in this Figure that some TFs have consistent positive or negative correlation with the clusters of genes which represent their candidate transcriptional targets by binding sites' enrichment. On the other hand, most of the transcription factors' correlation values span a wide range of correlation values. The positively correlated transcription factors are ATF6, ZBTB7A, and ZBTB7B with C1*, GABPA and MYB with C2*, and GTF3A with C3*. The negatively correlated transcription factors are GTF3A, STRA13, and ZBTB14 with C1*, RFX1 and maybe ETS1 with C2*, and ETS1 with C4*.

6.3.7. The transcription factor ZBTB7A (LRF)

The gene ZBTB7A encodes the transcription factor LRF (also known as FBI-1 and Pokemon), which was shown to play roles in breast cancer induction as an oncogene (Zu, et al., 2011), the regulation of T-cells differentiation (Carpenter, et al., 2012), and the regulation of erythropoiesis (Maeda, et al., 2009). LRF is directly transcriptionally activated by the well-known erythropoietic master regulator GATA-1 (Maeda, et al., 2009). It was also shown to highly co-occupy, with GATA-1, the loci whose genes are up-regulated after the addition of GATA-1 to GATA-1-null erythropoietic cells (Yu, et al., 2009); thus, there is a positive feedback loop in which GATA-1 mediates LRF activation, and this could be critical to erythropoiesis (Hattangadi, et al., 2011). Another important regulatory loop in erythropoiesis was observed in which LRF is activated by EKLF (KLF1) (Yu, et al., 2009; Doré & Crispino, 2011), and EKLF's promoter itself is occupied by both LRF and GATA-1 (Doré & Crispino, 2011).

It is notable that LRF plays the general role of proliferation induction and/through apoptosis repression. It was shown to induce the apoptosis repressor BCL-2 through the activating NF-κB in hepatocellular carcinoma (liver cancer) (Zhao, et al., 2011; Kong, et al., 2012), to repress the tumour suppressor ARF in breast cancer (Maeda, et al., 2005a), and to repress the pro-apoptotic factor Bim (BCL2L11) during terminal erythroid differentiation (Maeda, et al., 2009). The latter two cases were confirmed as direct repression activities by promoter binding analysis while the first one was inferred by analysing the effects of siRNA LRF knockdown on its targets expression.

LRF loss in mice fetal liver resulted in normal erythropoiesis until the pro-erythroblasts and basophilic erythroblasts stages, and then blocked development at the polychromatic and orthochromatic erythroblasts with poor condensed chromatin pattern and very few enucleated cells. This ultimately resulted in lethality due to severe anaemia and profoundly impaired cellular differentiation (Maeda, et al., 2009). Loss of LRF in adult mice was also shown to result in defects erythropoietic development specifically in the transition from R II to R III/IV which correspond to basophilic erythroblasts and poly/orthochromatic erythroblasts respectively (Maeda, et al., 2009). Double mutants which lost both LRF and Bim partially recovered the effects of LRF loss. This fact elucidated one side of the role of LRF in erythropoiesis, which is inducing cell growth and repressing apoptosis through repressing the pro-apoptosis factor Bim (Maeda, et al., 2009).

6.3.8. The transcription factor GATA-1

Although GATA-1 is a well-known master regulator in erythropoiesis (Keller, et al., 2006; Welch, et al., 2004; Yu, et al., 2009), its binding site has not been shown enriched in any of the 2kb upstream sequences of the genes in the five clusters C1* to C5* (Table 6.6). Even though, this can be justified by what many studies observed in that the annotation of the binding site of GATA-1 is a poor predictor for in vivo GATA-1-dependent regulation because it tends to bind to distal rather than proximal sites to the start of transcription sites of the genes which it regulates (Yu, et al., 2009; Hattangadi, et al., 2011). To investigate this issue further, we have projected the results of the genome-wide GATA-1 chromatin occupancy analysis by Yu and collaborators unto our results.

Yu and colleagues identified 1834 genes in whose loci the transcription factor GATA-1 shows peak occupancy in vivo. A gene's locus has been defined as the DNA sequence starting 10kb upstream the start transcription site (STT) and ending 3kb downstream the 3' end of the gene (Yu, et al., 2009). Yu et al then used microarray analysis to identify those genes that differentially expressed when GATA-1 was provided to arrested cells in erythropoiesis by the GATA-1-estrogen receptor ligand binding domain fusion molecule (G1-ER4) murine cell system (Yu, et al., 2009). Out of the 1834 genes occupied by GATA-1, 1328 genes are within the 13269 genes included in our study, and 300 of these 1328 were also differentially expressed when GATA-1 was provided (Yu, et al., 2009).

Figure 6.9 shows the percentage of genes included in each of the five clusters C1* to C5* whose loci were found occupied by GATA-1 in vivo according to Yu and colleagues' study (Yu, et al., 2009). It also shows the percentage of genes which in addition to being occupied by GATA-1, were differentially expressed when GATA-1 was provided to

GATA-1-null cells in that same study (Yu, et al., 2009). 41 genes' loci out of 123 genes in C1* were found occupied by GATA-1 (33%, p-value 1.3×10^{-12}) and 30 of them also differentially expressed (24%, p-value 3.4×10^{-12}). This is significantly higher than the clusters C2* to C5* whose GATA-1 occupancy showed the percentages of 4%, 18%, 5%, and 7%, respectively with the respective p-values of 0.99, 8.0×10^{-3}, 0.91, and 0.88 (Figure 6.9). The enrichment of genes that were both occupied by GATA-1 and differentially expressed in these four clusters C2* to C5* is even lower with the percentages of 0%, 2%, 0%, and 0% respectively, and the respective p-values of 1.0, 0.94, 1.0, and 1.0.

Out of the 41 genes in C1* whose loci were found occupied by GATA-1, only 13 of them had their occupancy site in their upstream 10kb sequences while most of the others were occupied within the introns or the exons, or in less often cases in their downstream sequences. Moreover, only five of these 13 had the GATA-1 occupancy in their 2kb upstream sequences while the others were more distal. We conclude that C1* is indeed enriched with GATA-1 targets, and that the adoption of genome-wide results for GATA-1 occupancy such as Yu and colleagues' can be a better predictor for such enrichment than direct mining in the proximal upstream sequences of the genes.

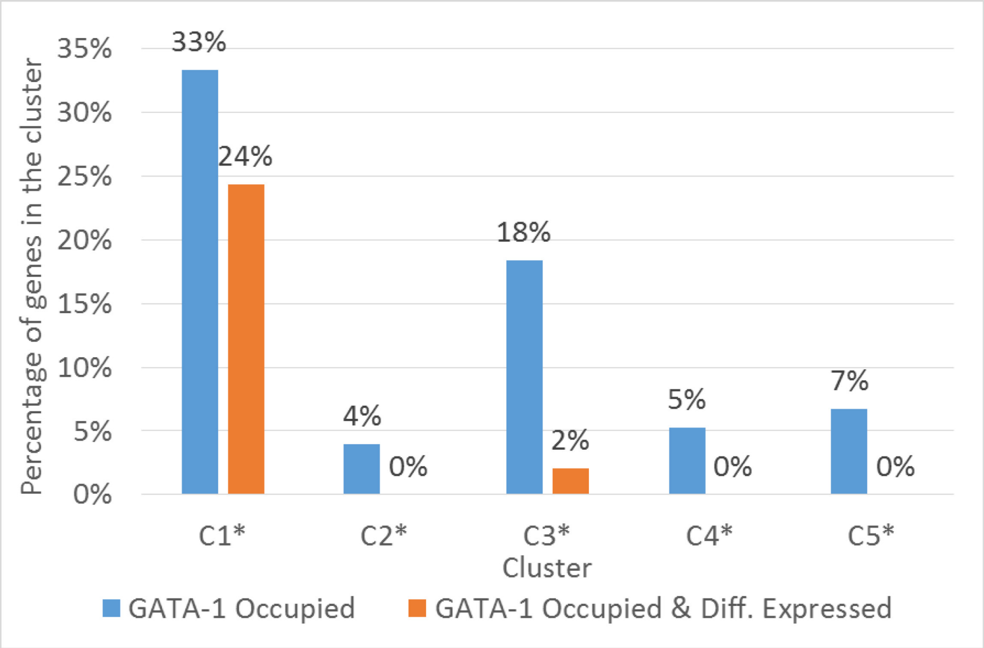

Figure 6.9. Percentage of GATA-1 potential targets in the erythropoietic clusters C1* to C5* based on (Yu, et al., 2009)

6.4. Summary and conclusions

Four human and four murine erythropoietic gene expression datasets were analysed collectively by the Bi-CoPaM to identify the subsets of genes that are consistently co-expressed. Five significant clusters were found and labelled as C1* to C5*. Interestingly, when the average profiles of any of those five clusters from all of the datasets are projected to a common horizontal axis representing the erythropoietic stages from haematopoietic stem cells (HSCs) to reticulocytes (RCs) (Figure 6.6), they show high consistency in terms of the erythropoietic stages in which their expression peaks (Figure 6.5 and Figure 6.7).

C1*, which is highly enriched with processes related to haem biosynthesis and cell-cycle arrest, preserves very low expression up to CFU-E and is then gradually up-regulated until the latest stages of erythropoiesis. C2*, which is enriched with RNA- and ribosome-related processes, starts with a moderate expression at the early HSCs and is thereafter up-regulated to peak at Pro-E, then down-regulated to reach its minimum expression at the terminal stages. C3*, which is enriched with signal transduction, apoptosis, blood coagulation, immune response, and cell proliferation, peaks at HSCs and is then continuously down-regulated towards the stages of basophilic and polychromatic erythroblasts where it starts gaining some up-regulation until the latest stages. C4*, which is highly enriched with viral transcription, translation, and glycolysis, starts with moderate to high expression values at HSCs and is then up-regulated to plateau from BFU-E to basophilic erythroblasts; after that, it is down-regulated to reach its minimum expression at the final stages of erythropoiesis. C5*, which is highly enriched with cell-cycle and DNA metabolism-related processes, has very low expression at HSCs and consequently increases gradually to peak at basophilic erythroblasts; it then drops to low values at the final stages.

Upstream DNA sequence analysis has identified some potential positive and negative transcriptional regulators for the clusters. The candidate positive regulators are ATF6, ZBTB7A (LRF), and ZBTB7B for C1*, GABPA and MYB for C2*, and GTF3A for C3*, while the candidate negative regulators are GTF3A, STRA13, and ZBTB14 for C1*, RFX1 and potentially ETS1 for C2*, and ETS1 for C4*.

The transcription factor GATA-1, which is a well-known positive regulator for genes needed in the late stages of erythropoiesis, was not identified by upstream sequence analysis because it tends to bind to distal rather than proximal sites to the start of transcription sites of its target genes (Yu, et al., 2009; Hattangadi, et al., 2011). However, based on the data provided by the in vivo analysis of Yu and colleagues (Yu, et al., 2009), C1* is indeed highly enriched with potential targets of GATA-1 (Figure 6.9).

Taken together, this pipeline of in silico analysis of erythropoietic datasets has established a focused set of results regarding the expression profiles, functions, and regulators of five subsets of genes which are consistently co-expressed over eight different human and murine gene expression datasets. Biological functional experiments can take place to follow up these findings.

Chapter 7
E. Coli Bacterial Data Analysis

7.1. Introduction to *E. coli* bacteria

Escherichia coli, which is considered as a model prokaryotic organism, is a rod-shaped bacterium that commonly inhabits the intestine of warm-blooded animals, including humans. Similar to the model eukaryotic organism *Saccharomyces cerivisiea* (budding yeast), *E. coli* is extensively studied due to its relative accessibility and ease in culturing and manipulation, and due to the general knowledge that can be gained by projecting the findings back to the general level of bacteria, prokaryotes, or living species.

After the success of applying the Bi-CoPaM method to budding yeast, a similar approach has been carried out to analyse *E. coli* bacterial datasets. The findings, which are detailed in this chapter, have been published in an invited journal paper in the *Journal of Signal Processing Systems* (Abu-Jamous, et al., 2015b).

Here, we mine five different *E. coli* microarray datasets to identify the subsets of genes that are consistently co-expressed over such wide range of biological conditions. We also aim at scrutinising the results of the Bi-CoPaM analysis to build biological hypotheses which relate some genes with previously unknown biological functions to the potential biological processes in which they may participate. These hypotheses serve as pilots for future more focused biological gene discovery studies.

7.2. Datasets and experimental design

Five *E. coli* microarray datasets have been considered in this study and are listed in Table 7.1. The first column of this Table shows the letter identifier that we shall use hereinafter to refer to each of these datasets. The five datasets were generated from a range of different biological conditions like different temperatures (Lee, et al., 2008), treatment with cefsulodin (Laubacher & Ades, 2008), mecillinam (Laubacher & Ades, 2008), and colicin M (Kamenšek & Žgur-Bertok, 2013), cofactor perturbations (Holm, et al., 2010), and growth under different glycerol conditions (Arunasri, et al., 2013). We have applied the Bi-CoPaM method with DTB binarisation to these five datasets to obtain the subsets of genes which are consistently co-expressed across all of them. The individual clustering methods used are k-means with the deterministic Kauffman's initialisation (Pena, et al., 1999), self-organising maps (SOMs) (Xiao, et al., 2003), and hierarchical clustering (HC) with Ward's linkage (Eisen, et al., 1998). The DTB δ value ranged from zero to unity with a step size of 0.1. The chosen number of clusters is three.

Table 7.1. Five *E. coli* microarray datasets

ID	Acc. No.*	N	Description	Ref.
A	GSE9923	10	Indole signalling at low temperatures	(Lee, et al., 2008)
B	GSE10159	9	Treatment with cefsulodin and mecillinam	(Laubacher & Ades, 2008)
C	GSE20374	3	Response to cofactor perturbations	(Holm, et al., 2010)
D	GSE34275	6	Growth in presence and absence of glycerol	(Arunasri, et al., 2013)
E	GSE37026	4	Treatment with colicin	(Kamenšek & Žgur-Bertok, 2013)

* The accession numbers represent the NCBI GEO database's identifiers.

7.3. Results and discussion

7.3.1. Clusters average expression profiles

The numbers of genes included in each of the three clusters at each of the δ values are listed in Table 7.2. The three clusters were ordered based on the number of genes kept in them at the tightest δ value of 1.0, and they were labelled as C1, C2, and C3, respectively. The profiles of the genes included in C1, C2, and C3 from each of the five datasets at four different δ values are respectively shown in Figure 7.1, Figure 7.2, and Figure 7.3. It can be seen in these Figures that while moving from high δ values to lower ones, the clusters are widened with more genes included. At low δ values, the clusters become relatively noisy.

Table 7.2. Numbers of genes included in each of the three *E. coli* clusters C1, C2, and C3 at all δ values

δ	C1	C2	C3
0.0	2076	1735	460
0.1	1520	1209	193
0.2	1208	864	97
0.3	885	599	33
0.4	565	377	11
0.5	378	234	2
0.6	283	149	1
0.7	120	57	1
0.8	61	20	0
0.9	21	3	0
1.0	21	3	0

It is worth mentioning while analysing these clusters that the general pattern of the profiles of the genes included in any single cluster at any given δ value is very different between the different datasets. For example, the 21 genes included in C1 at $\delta = 0.9$ show generally down-regulated profiles in the datasets C and D, while their profiles in the datasets

A, B, and E are very different from that. This is because the criterion upon which the Bi-CoPaM stands is that those genes are consistently well correlated with each other across different datasets even if their average profiles differs from one dataset to another.

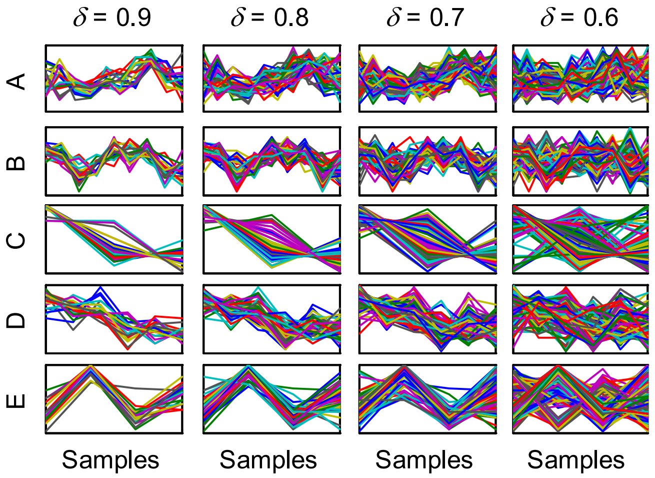

Figure 7.1. Profiles of genes in C1 from each of the five *E. coli* datasets at different *δ* values

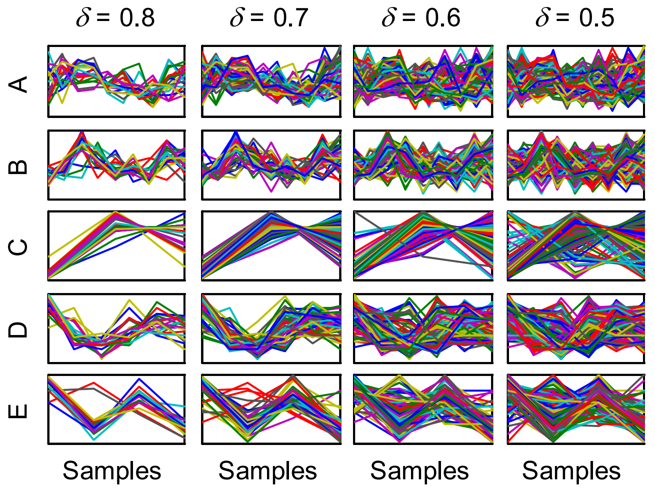

Figure 7.2. Profiles of genes in C2 from each of the five *E. coli* datasets at different *δ* values

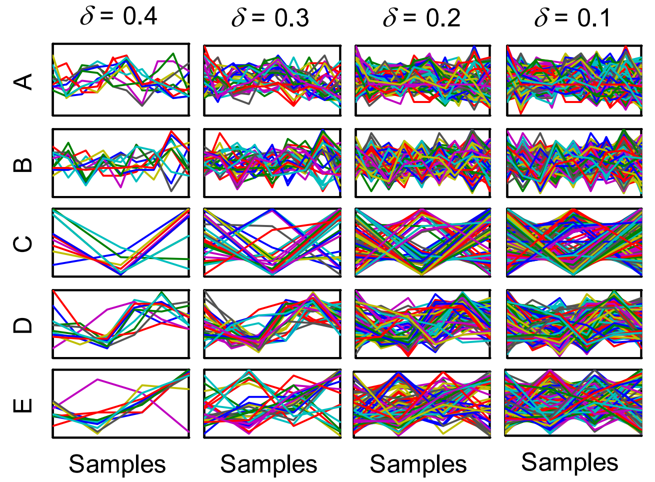

Figure 7.3. Profiles of genes in C3 from each of the five *E. coli* datasets at different *δ* values

The horizontal axis of each of the sub-plots represents samples while the vertical axis represent normalised expression values. The *δ* values, reflecting the tightness of the clusters, decrease from the left to the right resulting in more genes being included in the clusters are lower tightness levels. Each row of sub-plots represents one of the five datasets A to E described in Table 7.2.

Another key observation in Figure 7.1 is that, considering the profiles of C1 at $\delta = 0.9$ in the dataset C, there are few genes that show very different profiles to the majority of the genes in the cluster. The same observation applies to a single gene from that same cluster in the dataset E, yet they are included within the same cluster. This is because those few genes which lose their co-expression with the rest of the genes in the cluster in one dataset, are still well co-expressed with them in the other four datasets. Hence their inclusion within the same cluster. The same applies to some genes in Figure 7.2 and Figure 7.3.

Interestingly, the profiles of the clusters C1 and C2 consistently show reciprocal profiles over the five datasets A to E. We have quantified this observation by calculating the Pearson's correlation values (ρ) between the average profiles of each pair of the three clusters based on each one of the five datasets (Figure 7.4). Figure 7.4 shows that at four out of five datasets, namely all but D, the average profiles of the clusters C1 and C2 show strong negative correlation (ρ < -0.75). On the other hand, the correlation values for the cluster pairs (C1, C3) and (C2, C3) show low absolute correlation values at all datasets with the exception of the pair (C1, C3) at the single dataset D. A consistently negatively correlated pair of subsets (clusters) of genes indicates that they may be regulated by a common genetic regulator which when activates one subset of genes deactivates the other subset.

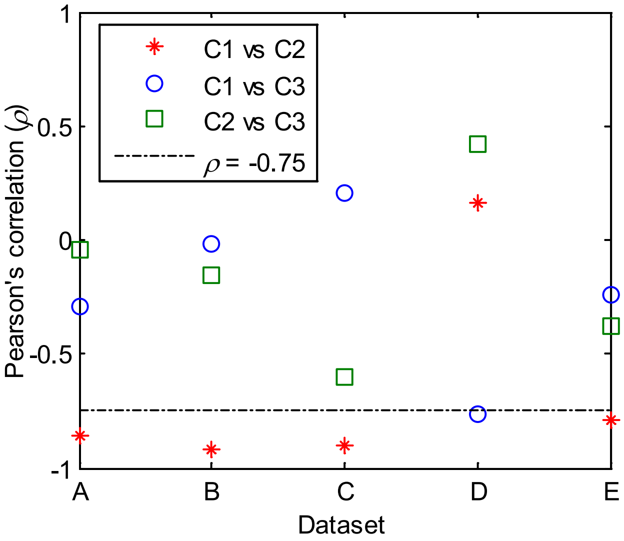

Figure 7.4. Pairwise Pearson's correlation values (ρ) between the average profiles of the pairs of clusters C1-C2, C1-C3, and C2-C3

The dashed black line marks the value ρ = -0.75. It can be seen that the clusters C1 and C2 have very strong negative correlation values across the datasets in contrast to the other pairs of clusters, namely (C1, C3) and (C2, C3), which do not show such pattern.

7.3.2. Biological relevance

To investigate the biological relevance of the genes included in these clusters, we have conducted Gene Ontology (GO) term enrichment analysis to the constituent genes of these clusters. The Gene Ontology Consortium is a major bioinformatics initiative which assigns the relevant terms out of a list of defined GO terms to the genes of different species based on the evidence existing in the published literature (The Gene Ontology Consortium, 2013). Three types of GO terms were defined by that project, namely biological process terms, molecular function terms, and cellular component terms. The assignment of GO terms to the genes is regularly updated as new research studies are published.

We have analysed the subsets of genes included in the clusters C1 and C2 at different tightness levels (δ values) to identify the GO terms that are highly enriched in those clusters. Because C1 and C2 are strongly negatively correlated, C3 loses all of its genes at a relatively lower δ value (Table 7.2), and C3 is noisier than C1 and C2 (Figure 7.3), we have excluded C3 from further biological analysis.

The most enriched biological processes in C1 and C2 at different δ values are shown in Table 7.3 and Table 7.4 respectively. It can be seen that C1 is highly enriched with translation, which is the processes of translation and tRNA processing, which are involved in producing new proteins. When this is observed while considering the dataset C for example, it can be seen that C1 genes have high expression values at the first point, which represents reference cells, while having down-regulated (low) values at the second and the third points, which represent cells transformed with plasmids containing with NADH oxidase and soluble ATPase respectively. This type of transformation lowers the levels of the two important metabolic cofactors NADH and ATP respectively, and therefore lowers the growth of the cells. It is well known that protein synthesis is repressed under poor growth conditions in species ranging from bacteria (Barria, et al., 2013; Orelle, et al., 2013), to fungi (Wade, et al., 2006), and even mice and humans (Shalgi, et al., 2013). Therefore, these results resonate well with the existing literature. Genes involved in methylation, which is a common process in living cells that adds a methyl group to a molecule, are also enriched in this cluster; it is interesting to investigate the reason and consequences of this consistent co-expression between significant numbers of genes involved in translation as well as methylation. Although large numbers of genes involved in transport processes are included in C1, they are not significantly enriched. This is because the number of genes known to participate in this process in the background, i.e. the entire *E. coli* genome, is large (Table 7.3).

Table 7.3. Most enriched biological processes in C1 at different δ values

Process	**$\delta = 0.9^*$**	**$\delta = 0.8^*$**	**$\delta = 0.7^*$**	**$\delta = 0.6^*$**	**Back$^\#$**
Translation	3 (1.4×10^{-2})	7 (6.7×10^{-4})	10 (6.7×10^{-4})	19 (4.6×10^{-5})	98
DNA repair	3 (5.8×10^{-3})	3 (9.5×10^{-2})	8 (1.2×10^{-3})	13 (1.3×10^{-3})	71
tRNA processing	1 (2.1×10^{-1})	3 (2.8×10^{-2})	8 (3.4×10^{-5})	11 (1.5×10^{-4})	43
Transport	3 (6.5×10^{-1})	6 (9.3×10^{-1})	13 (9.4×10^{-1})	46 (3.7×10^{-1})	611
Methylation	2 (4.0×10^{-2})	3 (6.4×10^{-2})	7 (2.0×10^{-3})	11 (3.0×10^{-3})	60
Unknown process	5 (5.2×10^{-1})	16 (2.7×10^{-1})	28 (4.2×10^{-1})	65 (4.0×10^{-1})	880
All genes in the subset	21	61	120	283	3956

* The contents of the cells in these columns are in the format [number of genes (p-value)], where the p-value is based on the hypergeometric distribution.

# Number of genes from the entire *E. coli* genome, which are associated with the corresponding process

In contrast to C1, the cluster C2 is highly enriched with transport genes, and more specifically with the sub-process of carbohydrate transport, and even more specifically with the transport of the carbohydrate maltose. It is also highly enriched with carbohydrate metabolic processes, especially with the L-ascorbic acid catabolic process (Table 7.4). This high consistency in co-expression, over multiple *E. coli* datasets from various conditions, between the genes in this subset which is highly enriched with carbohydrate transport and metabolism is an important observation. This indicates that the regulation machinery of the processes dealing with the different carbohydrate nutrients may be global at the level of the species regardless of the specific biological context.

Table 7.4. Most enriched biological processes in C2 at different δ values

Process	**$\delta = 0.8^*$**	**$\delta = 0.7^*$**	**$\delta = 0.6^*$**	**$\delta = 0.5^*$**	**Back$^\#$**
Transport	6 (7.5×10^{-2})	18 (1.6×10^{-3})	32 (2.9×10^{-3})	50 (8.0×10^{-3})	611
Carbohydrate transport	3 (9.5×10^{-3})	11 (3.2×10^{-8})	15 (7.4×10^{-7})	21 (2.4×10^{-8})	89
Maltose transport	0 (1.0)	3 (2.8×10^{-5})	3 (5.0×10^{-4})	4 (5.7×10^{-5})	5
Carbohydrate metabolic process	2 (1.4×10^{-1})	4 (1.2×10^{-1})	15 (1.1×10^{-4})	22 (7.5×10^{-6})	133
L-ascorbic acid catabolic process	1 (2.0×10^{-2})	2 (1.2×10^{-3})	3 (2.0×10^{-4})	4 (1.2×10^{-5})	4
Unknown process	3 (8.6×10^{-1})	9 (9.1×10^{-1})	30 (7.7×10^{-1})	51 (5.9×10^{-1})	880
All genes in the subset	20	57	149	234	3956

* The contents of the cells in these columns are in the format [number of genes (p-value)], where the p-value is based on the hypergeometric distribution.

# Number of genes from the entire *E. coli* genome, which are associated with the corresponding process

7.3.3. Hypothesis for genes with previously unknown biological processes

Many genes included in the clusters C1 and C2 have not been associated with any known biological process yet. We have investigated those genes to draw hypotheses that associate some of them with potential processes. Although such hypotheses are speculative and require biological functional experiments for them to be confirmed, their proposal based on bioinformatics represents a focused starting point for guided biological studies.

The molecular function and the cellular component for most of the genes of unknown biological processes in C1 and C2 are also unknown. Despite that, few genes of unknown processes have been associated with known functions or components, which, when scrutinised in tandem with our results, leads to very interesting hypotheses. The gene yegD, which is included in C1 at the very tight case of $\delta = 0.9$, was associated with the molecular functions ATP binding and nucleotide binding, which makes it a candidate gene participating in the DNA repair process with which C1 is enriched (Anon., 2014).

In the cluster C2 at δ = 0.8, the gene ydhY, whose biological process and cellular components are unknown, is associated with many molecular function terms including electron carrier activity and iron-sulphur cluster binding. Such associations make ydhY a candidate gene which participates in transport processes, with which C2 is enriched (Anon., 2014; Partridge, et al., 2008). In C2 at δ = 0.7, there are two other genes of interesting observations, yhdU and aphA. The gene yhdU is integral to a membrane, and the gene aphA is localised in the outer membrane-bounded periplasmic space, and is associated with the molecular functions metal ion binding, cofactor binding, and hydrolase activity. Those observations lend support to the idea of both genes being candidate transport genes.

These hypotheses, which relate some genes whose biological process terms are unknown to their potential processes, serve as pilots for directed future biological functional studies.

7.4. Conclusions

The *Bi-CoPaM* method is a recently proposed ensemble clustering method which allows analysis of multiple datasets collectively, and generation of clusters that vary in tightness. While clustering a set of genes, the Bi-CoPaM allows any single gene to be exclusively assigned to a single cluster, which generates complementary clusters, or to be simultaneously assigned to multiple clusters, which generate wide and overlapping clusters, or not to be assigned to any of the clusters, which generates tight and focused clusters. By applying the tight-clusters approach of the Bi-CoPaM to the genetic expression profiles of a defined set of genes from multiple datasets, the subsets of genes consistently co-expressed over these datasets are identified. In this paper, we have identified two main clusters with that attribute, which have consistently negatively correlated expression profiles. The first cluster is highly enriched with genes that participate in the biological processes of translation and DNA repair while the second cluster is highly enriched with genes that participate in transport processes. Based on biological analysis of our results, we have drawn some hypotheses, relating some of the genes whose biological processes are unknown with their potential processes. Biological researchers can use these hypotheses as bases for focussed future studies.

Chapter 8
Malarial Data Analysis

8.1. Introduction to malaria

Malaria is an infectious disease that is caused by parasites of species belonging to the genus *Plasmodium* and carried by female mosquitoes from the genus *Anopheles*. Five *Plasmodium* species have been reported as infectious to humans while other species infect other animals such as rodents. Most human deaths are caused by the *Plasmodium falciparum* species.

As malaria is responsible for up to one million deaths annually, the medical relevance of malarial research needs little further elaboration (World Health Organization (WHO), 2013). Furthermore, 40% of the world population are at risk of malarial infection in endemic countries distributed over the Sub-Saharan Africa, the Amazon Basin, and South and South East Asia (Hay, et al., 2009).

Plasmodium species are unicellular eukaryotes. However, their genomes and functions of cellular organelles differ greatly from other known eukaryotes. Due to this, a lot of the aspects of the molecular biology of them are poorly understood. The genome of *Plasmodium falciparum*, which includes more than 5,300 genes, was completely sequenced in 2002 (Gardner, et al., 2002).

Plasmodium species have a very special and complex cycle; while being in the saliva of the host mosquito, they are known as *sporozoites*. Sporozoites are transmitted to the blood stream of the target human (or any other relevant animal) by a mosquito bite. Sporozoites travel through the blood system until they invade the liver infecting the hepatocytes (liver cells). Some *Plasmodium* species' cells may enter a dormant stage in the liver, in which they are known as *hypnozoites*, a stage which may last for up to 30 years. Whether they turn into hypnozoites for a period of time or not, ultimately they divide and form a large number of *merozoites*, which burst from the liver cells and flee into the blood stream.

Merozoites are infectious to erythrocytes (red blood cells (RBCs)) as they invade them, reproduce asexually therein producing more merozoites, explode the erythrocytes, and flee again into the blood stream to invade more erythrocytes. This sub-cycle of erythrocyte invasion and merozoites' asexual reproduction is known as the *intra-erythropoietic developmental cycle (IDC)*, or the *erythropoietic stages*, or simply the *blood stages*.

Some merozoites decide not to invade more erythrocytes; they rather differentiate to the sexual form of male or female gametocytes. Those gametocytes are ingested by the mosquito through blood feeding. In the mosquito, gametocytes develop further to male or female gametes which form zygotes through sexual fusion. Zygotes develop to ookinetes which, in their turn, produce sporozoites. This lands the cycle at the first stage which we discussed, allowing for a new cycle to start thereinafter.

In order to commence my research in malaria, I have performed a preliminary analysis of two well-known blood-stage malarial datasets by using the Bi-CoPaM method (UNCLES type A) and the M-N scatter plots. The objective is to evaluate the ability of the method to obtain results which resonate with the literature as well as to experience real analysis in this field by actual practice. Each one of the two datasets, which were produced by Boztech and colleagues (Bozdech, et al., 2003) and Le Roch and colleagues (Le Roch, et al., 2003), measures the expression of the *Plasmodium falciparum* parasite's genome over a single complete intra-erythropoietic developmental cycle (IDC).

8.2. Experimental design

The Bi-CoPaM method was applied to the two datasets while adopting the initial clustering methods *k*-means with the Kauffman's deterministic initialisation (Pena, et al., 1999), hierarchical clustering (HC) with Ward's linkage (Eisen, et al., 1998), and self-organising maps (SOMs) (Xiao, et al., 2003), and while considering the K values of 8, 9, 10, 16, 18, 24, 30, and 40. The datasets were normalised by quantile normalisation and then by making each gene's profile with a zero-mean and a unity standard deviation. DTB binarisation was employed with δ values ranging from zero to unity with steps of 0.1. Finally, the resulting clusters were exposed to M-N scatter plots for cluster selection.

8.3. Results

The distances from the top-left corner of the M-N plot for the first 30 clusters are plotted in Figure 8.1. It is clear in here that the first nine clusters have significantly lower distances, and therefore better quality, compared to the rest of the clusters.

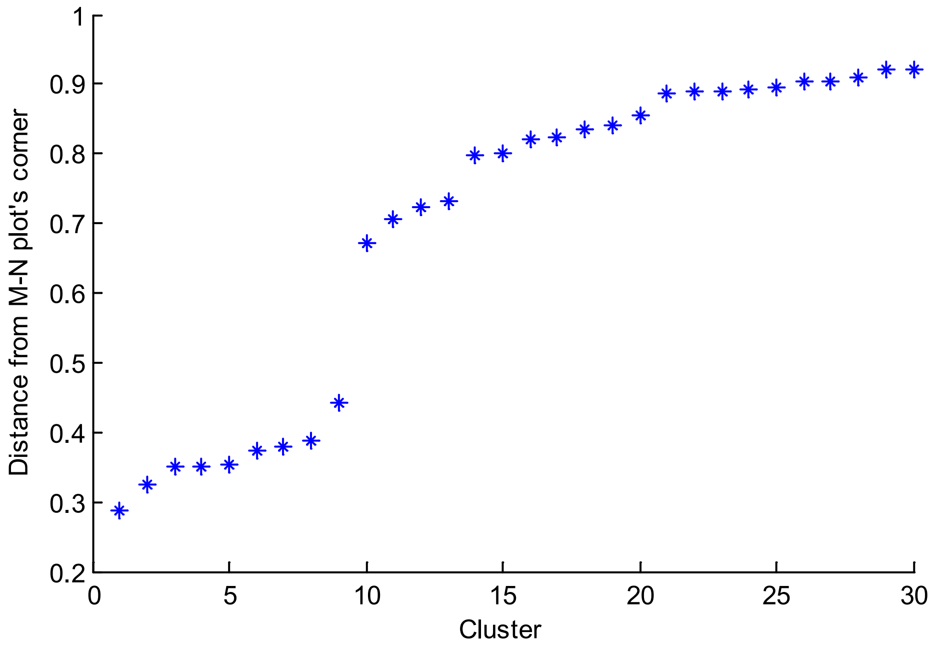

Figure 8.1. M-N scatter plot distances corner for the first 30 malarial clusters

Given that the M-N plots suggest that there are nine key clusters of consistently co-expressed genes in those two datasets, a closer look at the selected nine clusters has been considered. The average expression profiles of those clusters in each of the two datasets as well as the most enriched GO process terms in them are presented in Figure 8.2. It is clearly observed in this Figure that the nine clusters show a cascade of periodic profiles each of which has a single peak at one of the IDC cycle's stages, with general agreement on that between both datasets. The nine clusters were labelled as C1 to C9 after re-ordering in accordance to their peak times in the IDC cycle.

8.4. Discussion and conclusions

The findings of this preliminary analysis highly agree with previous findings regarding the periodicity in the *Plasmodium*'s gene transcription over blood-stage cycles (Bozdech, et al., 2003; Le Roch, et al., 2003). Despite the intensive analysis of these parasites in the blood stages, expression in other stages is still not as clearly understood (Kooij, et al., 2006). Nonetheless, large-scale datasets are becoming more available with an increasing pace of generation for multiple human and rodent *Plasmodium* species and for blood- and non-blood-stages (Kooij, et al., 2006; Otto, et al., 2014).

This preliminary analysis demonstrates the adopted computational framework's applicability to malarial datasets and is an evidence that there is a great potential in applying this framework to more existing and emerging datasets from different stages and different *Plasmodium* species collectively to advance our understanding of the molecular biology of this parasite.

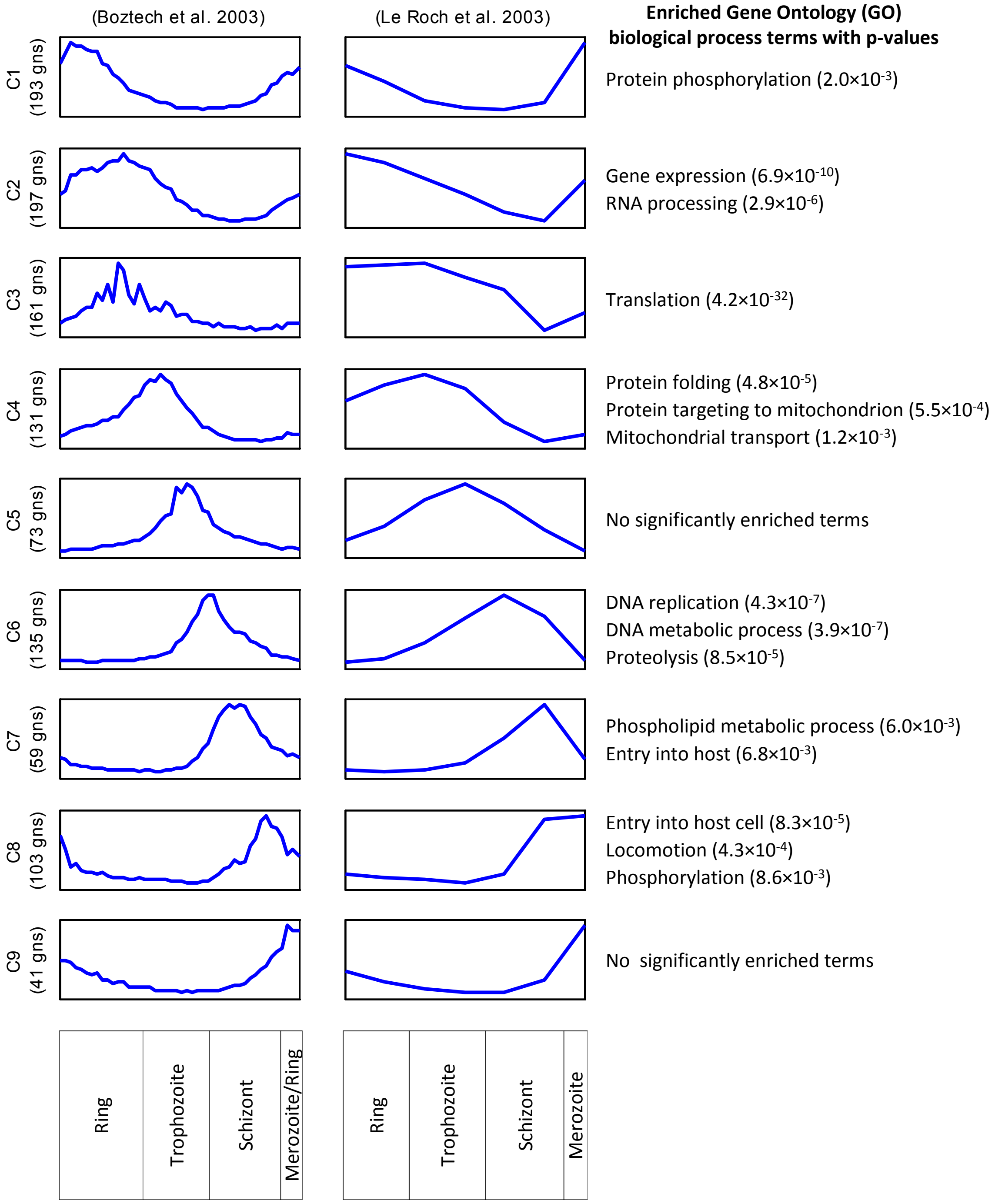

Figure 8.2. Expression profiles and GO terms of the nine malarial clusters found by preliminary Bi-CoPaM analysis

The average expression profiles of the nine clusters in each one of the two datasets are shown in nine rows of sub-plots, while the most enriched biological process GO terms with their p-values (p-value < 0.001) are shown in the third column of this grid of sub-plots. The horizontal axis of the sub-plots represent time while the vertical axis represents normalised expression values. The ranges of time points which represent the different IDC developmental stages (ring, trophozoite, schizont, and merozoite) are illustrated with labels below the grid of sub-plots.

Chapter 9
Summary, Conclusions, and Future Work

This thesis comprises advancements in the computational analysis of multiple high-throughput biological, mainly gene expression, datasets collectively. These advancements cover both the methodological and the application sides by the proposal of a novel suite of computational methods as well as elucidating important insights into various biological aspects by the application of such methods to real datasets.

The focal method in the proposed suite of methods is the ***UN**ification of **CL**ustering results from multiple datasets by using **E**xternal **S**pecifications (UNCLES)* method (Abu-Jamous, et al., 2015c). This method mines multiple gene expression datasets collectively in order to identify the subsets of co-expressed genes (genes with high correlation between their genetic expression profiles) consistently over the subject datasets while adhering to some external specifications. Two types of external specifications have been proposed here; type A mines for the genes that are consistently co-expressed in all of the given datasets while type B mines for the genes that are consistently co-expressed in one subset of datasets while being poorly co-expressed in another subset of datasets. An earlier development of UNCLES is the ***Bi**narisation of **Co**nsensus **Pa**rtition **M**atrices (Bi-CoPaM)* method (Abu-Jamous, et al., 2013a), which is equivalent to UNCLES' type A.

Amongst the key aspects of the Bi-CoPaM and the UNCLES methods is that they have tuning parameters which allow for unconventional clustering results to be formed. For instance, while clustering a set of genes, any gene may have one of three eventualities; it may be exclusively assigned to a single cluster, as conventional clustering methods do, or it may be simultaneously assigned to multiple clusters, or it may not be assigned to any of the clusters at all. As for the clusters, they may be conventional complementary clusters, or tight and focused clusters which leave many genes unassigned to any cluster, or wide and overlapping clusters. Amongst the benefits of such capability is the ability to inject genome-wide datasets (datasets including the entire unfiltered set of genes) into the method without filtering, and then to tighten the resulting clusters to be focused while expelling many genes outside all of the clusters. By this, the method applies the filtering step implicitly while clustering, and eventually meets the biological fact that most of the genes in an organism's genome are expected to be irrelevant to any single given biological context. Most of the experiments detailed in this thesis have such setup and demonstrate its applicability.

UNCLES and the Bi-CoPaM require various parameters to be set such as the number of clusters (*K*) and the tuning parameters δ and (δ^+, δ^-). Also, the results of these methods need to be validated. In order to address these aspects, a cluster validation and selection technique is proposed in this thesis based on M-N scatter plots (Abu-Jamous, et al., 2014b; Abu-Jamous, et al., 2015c). This technique favours those clusters which include higher numbers of genes (*N*) while maintaining lower levels of dispersion as measured by a mean-squared error-based (MSE-based) metric (*M*). The UNCLES method assisted by the M-N scatter plots technique represents a complete framework of consensus clustering for multiple datasets without the need to set any of the key parameters manually; in other words, it is a parameter-free framework.

In order to test and validate this framework, artificial datasets which meet relevant properties were synthesised by adopting a new approach of expression data synthesis (Abu-Jamous, et al., 2015c). This approach produces datasets with a known-ground truth, which is a desirable feature of artificial datasets rendering them as suitable means to test and validate other methods; yet this is not the unique feature of the proposed approach compared to other approaches of data synthesis; rather, the unique feature is that the values within the artificial datasets are borrowed directly from real datasets overcoming the issue of the faithfulness of the synthetic datasets in representing real data properties.

Another technique of cluster assessment and validation has been proposed in this work, namely the *F-P scatter plots* technique, which validates the results of clustering while taking the known ground-truth as a reference (Abu-Jamous, et al., 2015c). This technique has been employed while testing the UNCLES method and the M-N plots technique over the synthetic datasets for which the ground-truth is known and has shown that the UNCLES method combined with M-N plots can find those clusters which highly match the ground-truth.

The mature suite of methods, or partially developed versions of it, has been applied to various biological contexts revealing several biological findings and insights. Two major applications to yeast datasets were conducted and published; the first of them revealed important insights into the poorly understood yeast gene CMR1 and its relation to cell-cycle and DNA metabolism genes by analysing two yeast cell-cycle datasets (Abu-Jamous, et al., 2013b). On the other hand, the second experiment scrutinised forty yeast gene expression datasets from various contexts concluding that the well-known subset of ribosome biogenesis genes and a novel subset of genes are consistently co-expressed over all of the datasets and, more surprisingly, are consistently oppositely expressed. Hypotheses with respect to the functions and the regulation of both subsets of genes were drawn, mainly regarding the novel subset, which was named as the *anti-phase with ribosome biogenesis (APha-RiB)* subset of genes (Abu-Jamous, et al., 2014a).

Five *Escherichia coli* bacterial datasets from different contexts were mined by the Bi-CoPaM method identifying two subsets of genes as consistently co-expressed over all of the five datasets. Biological hypotheses regarding the function and regulation of those subsets were drawn and published (Abu-Jamous, et al., 2015b).

While collaborating with the group of Professor David Roberts at the University of Oxford, which is a research group focusing on the biomedicine of the human blood, eight human and murine blood gene expression datasets were analysed by the Bi-CoPaM method. Those datasets were all generated in the context of red blood cells production (erythropoiesis). Five focused subsets of genes, out of the entire human or murine genome, were identified as consistently co-expressed over all of the eight datasets. Interestingly, these five clusters show peak expression values at different stages of development throughout the erythropoiesis process. When this observation was added to the analysis of the regulation and functions of the clusters, several hypotheses were drawn. These hypotheses and other related ones are under investigation with our collaborators in order to take this research forward.

Finally, the UNCLES method with the M-N plots were applied to two popular malarial datasets as a preliminary experiment. The discovered nine clusters showed a perfect temporal cascade of peaks of expression throughout the blood stages of the malarial parasites. Alongside the analysis of the functions of the genes in those nine clusters, this preliminary experiment demonstrated the applicability of this suite of methods to malarial datasets, and represents a seed for my fellowship/grant applications as well as my prospective collaborations.

The current suite of methods does not answer all of the possible questions with respect to the collective analysis of multiple high-throughput biological datasets. For instance, other types of datasets, such as proteomic, glycomic, and metabolomic datasets exist abundantly. Moreover, more investigations of the efficiency and reliability of the methods can be conducted. Such concerns constitute subjects for my future work at the side of methods' design and development. As for the applications, many other areas in biology and biomedicine have produced a great deal of datasets, such as cancer research, and can represent targets for future applications of my methods.

As the future of this research is considered, the focus will be on the analysis of the malaria parasite. Malaria causes up to one million deaths annually and about 40% of the population of the earth live in malaria endemic regions.

Taken together, a mature suite of computational methods with the capability to analyse collectively, validate, and simulate multiple high-throughput gene expression datasets have been described in this thesis alongside a set of real applications to yeast, bacterial, human and murine blood, and malarial datasets. Despite filling many gaps and elucidating many poorly understood aspects in research, this work has opened the eyes to more questions and potential future work, which keeps the wheels of bioinformatic research and personal career development turning.

Appendix I

Introduction to the Molecular Biology of the Cell

I.A. The cell

Cells are the building blocks of organisms. A cell is a membrane-bound compartment crowded with different types of large and small molecules performing various series of biochemical interactions in order to grow, reproduce, and maintain its integrity. If the cell contains a subcellular membrane-bounded compartment, known as the nucleus, as well as other subcellular organelles (membrane-bounded compartments), it is a *eukaryotic* cell (Figure I.1 (a)). In contrast, if the cell lacks a real nucleus, it is a *prokaryotic* cell (Figure I.1 (b)). Eukaryotic organisms include animals, plants, fungi, and protists, while prokaryotes include bacteria and archaea.

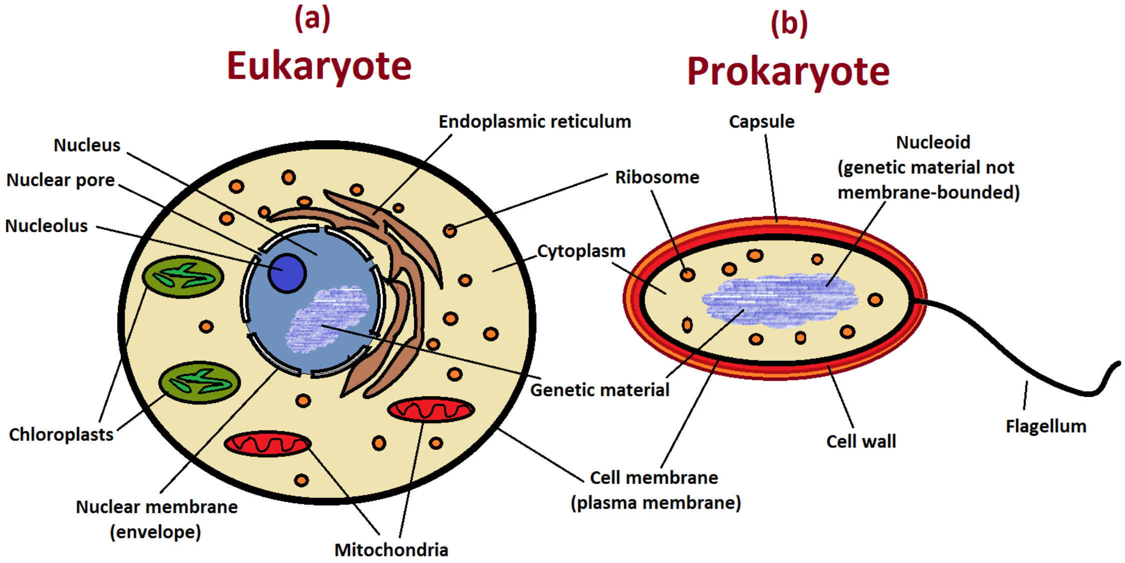

Figure I.1. Illustration of (a) eukaryotic and (b) prokaryotic cells

A more detailed demonstration of a typical eukaryotic cell is shown in Figure I.2. The cellular membrane protects the interior of the cell and allows for controlled exchange of materials with the extracellular space by the membrane transport proteins embedded in it.

Outside the cell, there is the *extracellular matrix*, which is a scaffold on which the cells of a multicellular organism adhere, and inside the cell there is a fluid known as the *cytosol* which forms the environment for the biochemical interactions to take place. The cell also contains solid filaments, known as the *cytoskeleton*, which give the cell its shape and strength, and play roles in component transportation.

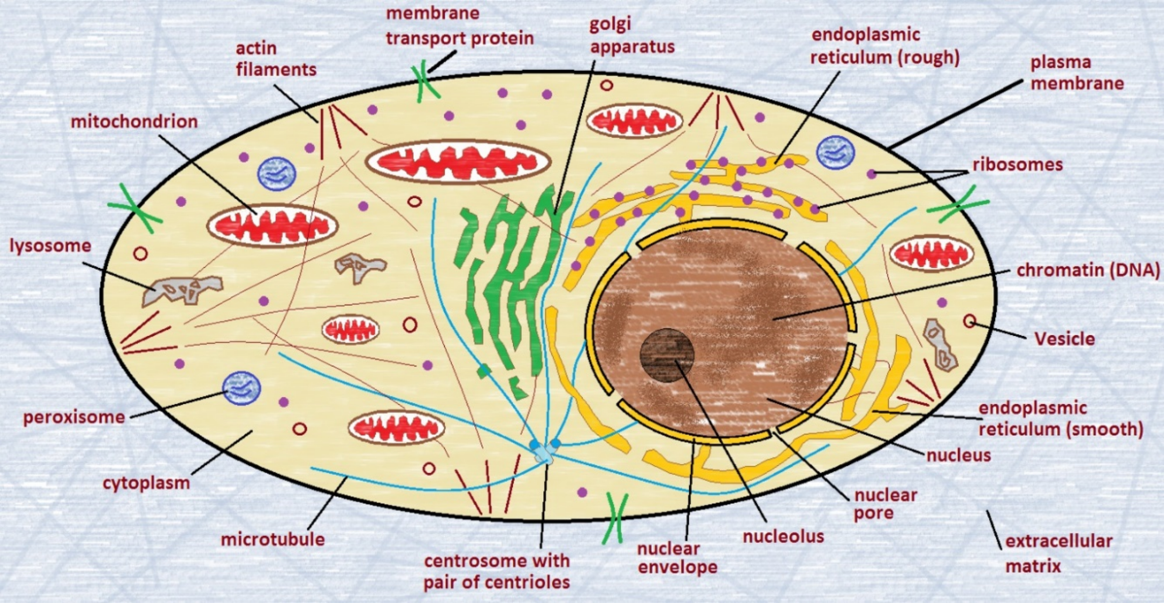

Figure I.2. A typical eukaryotic cell and its main components

The membrane-bounded nucleus in eukaryotes, or the non-membrane bounded nucleoid region in prokaryotes, contains the genetic material, which encodes the complete set of information required by the cell for growth, maintenance, and reproduction. The way in which this information is decoded and exploited will be detailed later in this Appendix.

Many types of subcellular organelles are found in cells. For instance, the mitochondrion is the energy factory for cell where sugars are decomposed to produce energy. Mitochondria also participate in the synthesis of some key molecular such as haem. Other organelles include the vesicles, which are membrane-bounded bubbles that actively transport molecules within the cell and participate in their export and import from and into the cell. Ribosomes, with the assistance of the rough endoplasmic reticulum, produce proteins, the smooth endoplasmic reticulum produces lipids, and Golgi apparatus produces carbohydrates. In plants and fungi, chloroplasts intake carbon dioxide, water, and sun light energy to produce oxygen and sugars. Moreover, plants possess a cell wall, which a tough wall of polysaccharides like cellulose. Chloroplasts and the cell wall are not illustrated in Figure I.2. Other types of subcellular organelles include the lysosomes, peroxisomes, and the centrosome.

Understanding the physiology of the cell is an important aspect, but a similarly important, or even a more important, aspect for bioinformaticians is to understand the different types of large molecules within the cell, their general functions, and the genetic programmes which control the cellular processes by decoding the information encoded in the genetic material; this is the topic of the rest of this Appendix.

I.B. Proteins

Before delving into the details of the genetic information encoding, decoding, and transmission, it is worth being briefed first on the most abundant class of large molecules in the cell, namely, the *proteins*.

Proteins conduct most of the biological processes within the cell. A *protein* is long linear chain of units known as *amino acids* (Figure I.3). Since there are 20 different types of amino acids, a protein can be considered as a linear text written in a 20-character language. In fact, each of the amino acids is denoted by a unique name, or a three-letter symbol, or a single-letter symbol (Table I.1).

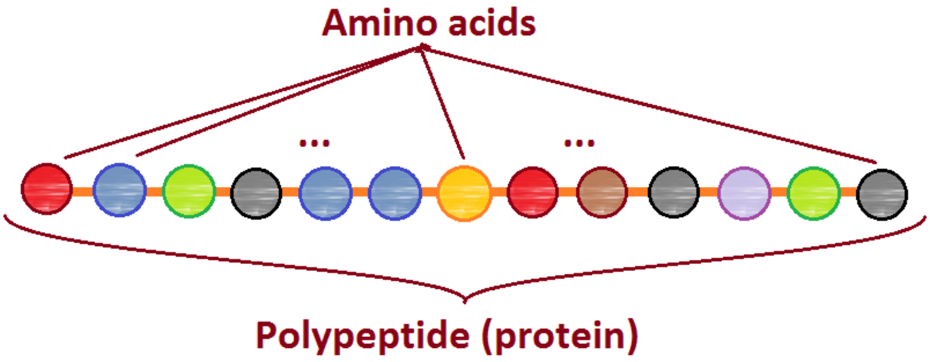

Figure I.3. The protein is a polypeptide, that is, a chain of joint amino acids

Table I.1. The twenty amino acids.

Symbol	Name	Charge / polarity	Symbol	Name	Charge / polarity
K (Lys)	Lysine	Basic	A (Ala)	Alanine	Nonpolar
R (Arg)	Arginine	Basic	V (Val)	Valine	Nonpolar
H (His)	Histidine	Basic	L (Leu)	Leucine	Nonpolar
D (Asp)	Aspartic acid	Acidic	I (Ile)	Isoleucine	Nonpolar
E (Glu)	Glutamic acid	Acidic	P (Pro)	Proline	Nonpolar
N (Asn)	Asparagine	Uncharged polar	F (Phe)	Phenylalanine	Nonpolar
Q (Gln)	Glutamine	Uncharged polar	M (Met)	Methionine	Nonpolar
S (Ser)	Serine	Uncharged polar	W (Trp)	Tryptophan	Nonpolar
T (Thr)	Threonine	Uncharged polar	G (Gly)	Glycine	Nonpolar
Y (Tyr)	Tyrosine	Uncharged polar	C (Cys)	Cysteine	Nonpolar

Owing to the attractive and repulsive forces between differently and similarly charged amino acids in a protein's linear chain, the protein folds upon itself to its most stable structure in the 3-D space. Therefore, the differences in the sequences of amino acids between different types of proteins lead to differences in their final 3-D shapes, and consequently to differences in their physical and chemical properties.

Lengths of proteins vary widely. Neidigh and colleagues were able to design a stable 20-amino acid-length protein-like polypeptide (Neidigh, et al., 2002). On the other hand, the largest known protein, with more than 38,000 amino acids, is the giant protein titin, which functions as a molecular spring contributing to the elasticity of muscles in humans (Bang, et al., 2001; Opitz, et al., 2003). The average length of proteins in the eukaryote *Saccharomyces cerevisiae* (budding yeast) is about 400 to 450 amino acids (Harrison, et al., 2003; Brocchieri & Karlin, 2005). In contrast, bacteria and archaea have average protein lengths of about 250 to 300 amino acids (Brocchieri & Karlin, 2005).

As for the number of different proteins in species, it is between 20,000 and 25,000 in humans (Collins, et al., 2004), ~6,300 in budding yeast, ~26,000 in the plant *Arabidopsis thaliana* (thale cress), ~4,300 in the *Escherichia coli* bacteria, and less than 500 in the *Mycoplasma genitalium* bacteria (Alberts, et al., 2008). These numbers show the large variation between species in terms of the number of proteins as well as their average length.

I.C. Central dogma of molecular biology

The *central dogma of molecular biology* is the set of rules which control the flow of information within the cells. Genetic information is encoded in the large deoxyribonucleic acid (DNA) molecule within the nucleus or the nucleoid of the cell. This information is sufficient to know, amongst other things, how to produce each single type of proteins, when it should be produced, and with which amounts. However, cells do not produce proteins directly by using the DNA molecules; they rather *transcribe* (copy) patches of this information from the DNA to be encoded in the form of ribonucleic acid (RNA) molecules, which are *translated* afterwards into proteins (Figure I.4). RNA molecules in reality are mere copies of patches of the DNA.

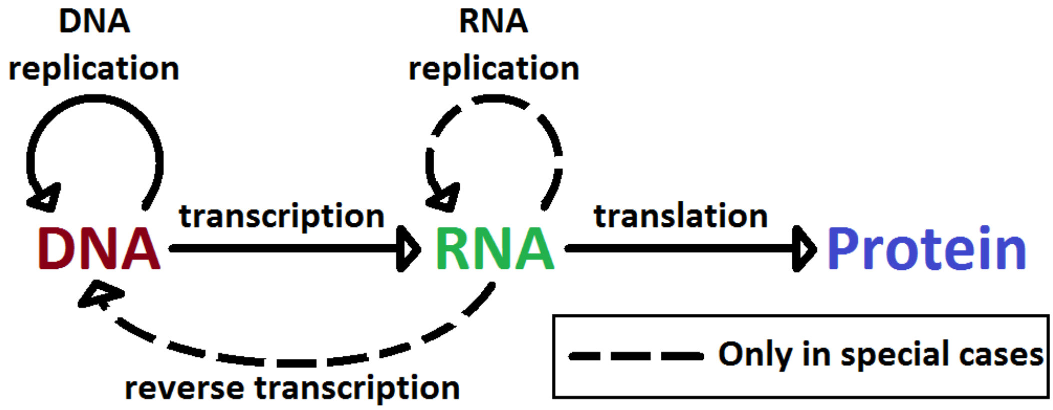

Figure I.4. Overview of the central dogma of molecular biology: information flow in cells

All of the cells within a multicellular organism, like a human individual, hold identical copies of the DNA molecule. This is because of the fact that all of the cells within that organism were in reality produced by series of self-replication of a single initial 'general-purpose' cell known as the *zygote*. However, at certain stages of cells' replication, some cells differentiate to become specialised in specific functions such as skin cells, bone cells, blood cells, retinal cells, and the like. Indeed, the DNA is faithfully replicated while replicating a cell in order to provide each of the daughter cells an identical copy of the mother cell's DNA (Figure I.4; *DNA replication*). Other uncommonly occurring directions of information flow in cells include *reverse transcription*, which produces a DNA molecule based on an RNA molecule, and *RNA replication*, which produces a new identical copy of an RNA molecule (Figure I.4).

I.D. DNA

The DNA encodes information in the form of a linear chain of *nucleobases*, or simply, *bases*. Having four types of bases in the DNA makes it equivalent to a four-letter linear text. The four bases are the *adenine (A)*, *thymine (T)*, *guanine (G)*, and *cytosine (C)*. For example, the human genome, that is the human DNA molecule, consists of more than three billion bases. The sequences of those four bases are chemically formed as bases protruding from a homogenous sugar-phosphate backbone (Figure I.5 (a)).

Because of their physical and chemical properties, the bases A and T can form a weak hydrogen bond when facing each other; the bases G and C similarly do. Consequently, a DNA chain of bases attracts the formation of a complementary chain of bases where each base in the complementary chain forms a hydrogen bond with its corresponding base in the original chain. Thus, the two chains of DNA bases, known as the two *strands*, hold identical information but in a complementary manner (Figure I.5 (b)). Although the hydrogen bond is a weak bond compared to the *phosphodiester* bond which joins any two successive bases in a single strand connecting (the yellow triangles in Figure I.5 (a) and (b)), the large number of hydrogen bonds between the pairs of bases in the two strands form a stable double-stranded DNA molecule. In reality, the two-strands do not take a plain structure as in Figure I.5 (b); they rather twist to form a *double helix* structure (Figure I.5 (c)), at which the DNA is most stable.

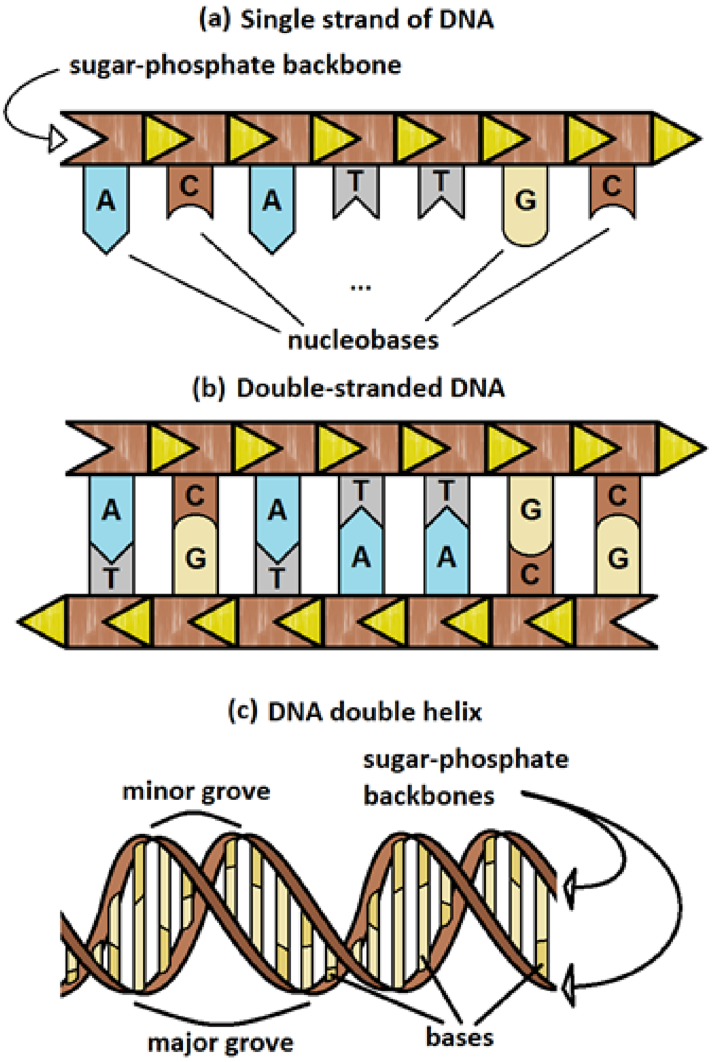

Figure I.5. The DNA molecule

(a) single-stranded DNA (b) double-stranded DNA; (c) the DNA double helix.

I.E. RNA

The RNA molecules are chains of four nucleobases with the same chemical structure of the DNA molecules except for the following differences:

1. The sugar component in the sugar-phosphate backbone is slightly different; it is a ribose instead of a deoxyribose. Hence the different name.
2. The uracil (U) nucleobase is used instead of thymine (T).
3. It is stable in its single-stranded form and therefore does not form a second complementary strand.

Having said that, transcription is the process of producing an RNA molecule with a sequence identical to a patch of the DNA sequence; indeed while replacing T bases with U bases.

Various types of RNA molecules exist. The most notable one is the *messenger RNA (mRNA)*, which represents a mere message carrying the information required to build a single type of proteins. Translation is the process of synthesising a protein molecule based on the information provided by an mRNA molecule.

Other types of RNA molecules are collectively known as *functional RNAs*, and they perform several cellular functions whilst staying in their RNA forms without being translated into proteins. *Ribosomal RNAs (rRNAs)*, *transfer RNAs (tRNAs)*, *microRNAs (miRNAs)*, *small interfering RNAs (siRNAs)*, *small nuclear RNAs (snRNAs)*, and *small nucleolar RNAs (snoRNAs)* are amongst the classes of functional RNAs.

I.F. Genes

The gene is that patch of the DNA molecule which is transcribed into a single RNA molecule. The protein-coding gene is the gene which is transcribed to an mRNA; in other words, the protein-coding gene is that DNA sequence which holds the required information to synthesise a single type of proteins. Non-protein-coding genes are those which are transcribed into functional RNAs. Certainly, most of the genes are protein-coding.

I.G. The genetic code

Again, a protein-coding gene, is a DNA sequence which encodes the sequence of a single type of proteins. In order to encode 20 different types of amino acids by using four different types of nucleobases, triplets of bases are required, and this is how it is in reality. A triplet of bases, that is, three consecutive bases, which is known as a *codon*, can encode for $4^3 = 64$ different symbols. The genetic code is the mapping between the 64 different codons and the 20 different amino acids, and is shown in Table I.2. Note that the code in this Table considers the RNA base U instead of the DNA base T, which are equivalent.

It be clearly seen in this Table that there is redundancy, that is, some different codons are mapped to the same target amino acid. For example, the codons UUA and CUG map to the same amino acid, which is the leucine (L). Some codons encode for punctuation marks, to indicate where translation starts and where it ends.

For example, if an mRNA sequence is ‘AUGUCACAA…’, it will be read in triplets and therefore will be translated to the protein sequence ‘MSQ…’, where ‘AUG’ is the code for ‘M’, ‘UCA’ is the code for ‘S’, and ‘CAA’ is the code for ‘Q’.

Table I.2. The genetic code

Mapping the three-base mRNA codons to their corresponding amino acids

First base (5' end)	**Second base**				**Third base (3' end)**
	U	**C**	**A**	**G**	
U	F	S	Y	C	**U**
	F	S	Y	C	**C**
	L	S	STOP	STOP	**A**
	L	S	STOP	W	**G**
C	L	P	H	R	**U**
	L	P	H	R	**C**
	L	P	Q	R	**A**
	L	P	Q	R	**G**
A	I	T	N	S	**U**
	I	T	N	S	**C**
	I	T	K	R	**A**
	M + START	T	K	R	**G**
G	V	A	D	G	**U**
	V	A	D	G	**C**
	V	A	E	G	**A**
	V	A	E	G	**G**

I.H. Transcription and transcriptional regulation

Transcription, or *gene expression*, is the process of synthesising an RNA molecule by copying a patch of a DNA sequence, that is, a gene. A machinery composed mainly of a large protein complex known as the *RNA polymerase* performs transcription by sliding over the DNA molecule at the required site and synthesising an RNA molecule by copying one base at a time. However, the DNA molecule is normally folded and packed by various proteins and is not straightforwardly accessible by RNA polymerases. Rather access to any specific gene is provided to RNA polymerases only when it is due for this gene to be transcribed and only for the required period of time.

The processes which control the expression, that is, the transcription, of genes is known as *transcriptional regulation* or *gene expression regulation*. This process is mainly conducted by proteins known as *transcription factors (TFs)* whose role is influence the amount of expression of genes by *activation* or *repression*, that is, by *positive regulation* or *negative regulation*, respectively.

Gene-specific TFs are able to detect specific DNA short sequences, known as motifs, which are found in the *upstream* sequences of a specific gene or group of genes. For example, the binding site of the SBF transcription factor is the motif 'CGCGAAAA' (Iyer, et al., 2001) (Figure I.6). By such types of TF-binding, the TF either activates transcription of the corresponding gene by recruiting RNA polymerases to the site and stabilising them, or represses transcription by blocking RNA polymerases from binding to the gene's sequence and consequently transcribing it.

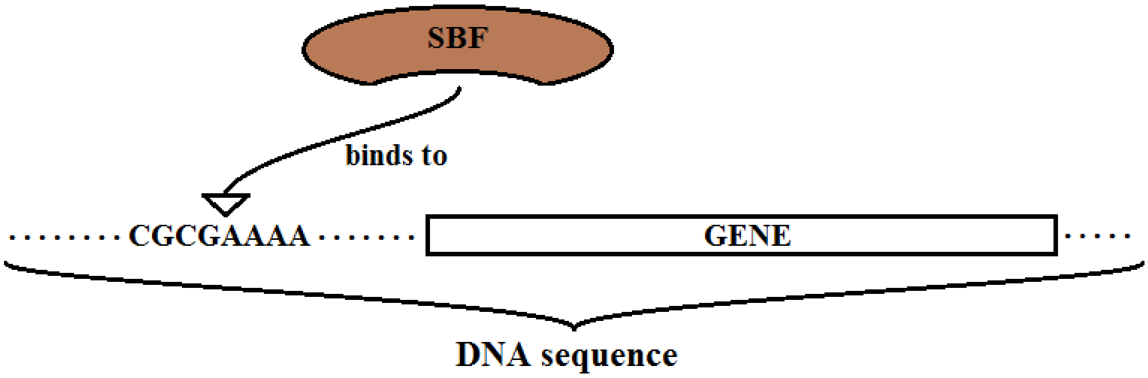

Figure I.6. The transcription factor SBF recognises and binds to its binding site 'CGCGAAAA' in a gene's upstream sequence

Different TFs may affect each while co-existing near the gene. For example, a TF may bind to a motif and blocks another TF from binding to that same motif. If the second TF is an activator of the corresponding gene, the first TF will be in reality nullifying the second TF's function. In this case, the condition for the transcription of that gene will include the existence of the second TF and the absence of the first one.

Moreover, the TFs themselves are proteins and protein complexes (multiple proteins assembled to form a single more complex unit). Therefore, TFs are products of genes; in other words, they are generated by the translation of mRNA molecules transcribed from genes. Thus, there are some TFs which can activate or repress the expression of those genes whose products are other TFs, or even the regulating TFs themselves (self-regulation). This in reality forms networks of transcriptional regulation. One example of such feedback loops is when a gene's product is a TF which negatively regulates itself; as this gene is transcribed and consequently translated, its product starts to increase in numbers; this product thereafter represses its own producing gene in order to halt its production; when this leads to a significant decrease in the numbers of this gene's product, that repression is released and the gene is allowed again to be transcribed to generate, again, more of its products.

Another notable information here is the fact that it is common to find that many genes have in their upstream sequences the same TF's binding site. This results in the phenomenon of *co-regulation*, where multiple genes are regulated similarly by the same TF or transcriptional regulatory machinery leading them to have similar expression profiles, that is, their expression levels go up and down synchronously. This is usually the case when the products of a group of genes work together in the same biological process. Reading this phenomenon reversely, observing that a group of genes are co-expressed, that is, they have similar expression profiles, indicates that they may be co-regulated, that is, regulated by the same machinery. Observing that the upstream sequences of these genes also have some similar motifs strengthen hypothesising that they are actually co-regulated.

I.I. Reference for more details

For more details regarding the physiology of the cell (the roles of the different subcellular organelles), as well as the central dogma of molecular biology (genes, transcription, translation, gene expression regulation, and related aspects), the reader is encouraged to refer to our Chapters 3 and 4 in our book *Integrative Cluster Analysis in Bioinformatics* (Abu-Jamous, et al., 2015a). These chapters were authored to brief non-specialists in essentials of molecular biology.

Appendix II

Bibliography

Abu-Jamous, B., Fa, R. & Nandi, A. K., 2015a. *Integrative cluster analysis in bioinformatics.* 1st ed. s.l.:Wiley.

Abu-Jamous, B., Fa, R., Roberts, D. J. & Nandi, A. K., 2013a. Paradigm of Tunable Clustering using Binarization of Consensus Partition Matrices (Bi-CoPaM) for Gene Discovery. *PLOS ONE,* 8(2), p. e56432.

Abu-Jamous, B., Fa, R., Roberts, D. J. & Nandi, A. K., 2013b. Yeast gene CMR1/YDL156W is consistently co-expressed with genes participating in DNA-metabolic processes in a variety of stringent clustering experiments. *Journal of the Royal Society Interface,* 10(81), p. 20120990.

Abu-Jamous, B., Fa, R., Roberts, D. J. & Nandi, A. K., 2013c. *Method for the identification of the subsets of genes specifically consistently co-expressed in a set of datasets.* Southampton, UK, s.n.

Abu-Jamous, B., Fa, R., Roberts, D. J. & Nandi, A. K., 2013d. *Identification of genes consistently co-expressed in multiple microarray datasets by a genome-wide Bi-CoPaM approach.* Vancouver, Canada, s.n., pp. 1172-1176.

Abu-Jamous, B., Fa, R., Roberts, D. J. & Nandi, A. K., 2014a. Comprehensive Analysis of Forty Yeast Microarray Datasets Reveals a Novel Subset of Genes (APha-RiB) Consistently Negatively Associated with Ribosome Biogenesis. *BMC Bioinformatics,* Volume 15, p. 322.

Abu-Jamous, B., Fa, R., Roberts, D. J. & Nandi, A. K., 2014b. *M-N scatter plots technique for evaluating varying-size clusters and setting the parameters of Bi-CoPaM and UNCLES methods.* Florence, Italy, s.n., pp. 6726-6730.

Abu-Jamous, B., Fa, R., Roberts, D. J. & Nandi, A. K., 2015b. Application of the Bi-CoPaM method to five Escherichia coli datasets generated under various biological conditions. *Journal of Signal Processing Systems,* 79(2), pp. 159-166.

Abu-Jamous, B., Fa, R., Roberts, D. J. & Nandi, A. K., 2015c. UNCLES: method for the identification of genes differentially consistently co-expressed in a specific subset of datasets. *BMC Bioinformatics,* Volume 16, p. 184.

Adriaens, M. E. et al., 2012. An evaluation of two-channel ChIP-on-chip and DNA methylation microarray normalization strategies. *BMC Genomics,* Volume 13, p. e42.

Ailon, N., Charikar, M. & Newman, A., 2008. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM),* 55(5), pp. 23:1-23:27.

Alberts, B. et al., 2008. *Molecular Biology of the Cell.* 5th ed. New York: Garland Science.

Alexe, G. et al., 2004. Consensus algorithms for the generation of all maximal bicliques. *Discrete Applied Mathematics,* 145(1), pp. 11-21.

Alizadeh, A. A. et al., 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature,* 403(6769), pp. 503-511.

Anon., 2014. [Online] Available at: http://amigo.geneontology.org/cgi-bin/amigo/go.cgi

Arunasri, K. et al., 2013. Effect of simulated microgravity on E. coli K12 MG1655 growth and gene expression. *PLOS ONE,* 8(3).

Ayad, H. G. & Kamel, M. S., 2008. Cumulative voting consensus method for partitions with variable number of clusters.. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 30(1), pp. 160-173.

Ayadi, W., Elloumi, M. & Hao, J.-K., 2009. A biclustering algorithm based on a Bicluster Enumeration Tree: application to DNA microarray data. *BioData mining,* 2(1), p. 9.

Bailey, T. L., 2011. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics,* 27(12), pp. 1653-1659.

Bailey, T. L. & Elkan, C., 1994. *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Menlo Park, California, s.n., pp. 28-36.

Bang, M.-L.et al., 2001. The complete gene sequence of titin, expression of an unusual approximately 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. *Circulation Research,* 89(11), pp. 1065-1072.

Barkow, S. et al., 2006. BicAT: a biclustering analysis toolbox. *Bioinformatics,* 22(10), pp. 1282-1283.

Barria, C., Malecki, M. & Arraiano, C. M., 2013. Bacterial adaptation to cold. *Microbiology,* 159(12), pp. 2437-2443.

Baumgartner, R., Windischberger, C. & Moser, E., 1998. Quantification in functional magnetic resonance imaging: fuzzy clustering vs. correlation analysis. *Magnetic Resonance Imaging,* 16(2), pp. 115-125.

Bertolacci, M. & Wirth, A., 2007. *Are approximation algorithms for consensus clustering worthwhile?..* Minneapolis, Minnesota, USA, Proceedings of the Seventh SIAM International Conference on Data Mining.

Bertoli, C., Skotheim, J. M. & de Bruin, R. A. M., 2013. Control of cell cycle transcription during G1 and S phases. *Nature Reviews Molecular Cell Biology,* Volume 14, p. 518–528.

Bester, M. C., Jacobson, D. & Bauer, F. F., 2012. Many Saccharomyces cerevisiae Cell Wall Protein Encoding Genes Are Coregulated by Mss11, but Cellular Adhesion Phenotypes Appear Only Flo Protein Dependent. *G3 (Bethesda),* 2(1), pp. 131-141..

Bhattacharjee, A. et al., 2001. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences,* 98(24), pp. 13790-13795.

Bolstad, B., Irizarry, R., Astrand, M. & Speed, T., 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics,* Volume 19, pp. 185-193.

Bozdech, Z. et al., 2003. The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum. *PLOS Biology,* 1(1).

Brannon, A. R. et al., 2010. Molecular stratification of clear cell renal cell carcinoma by consensus clustering reveals distinct subtypes and survival patterns. *Genes \& cancer,* 1(2), pp. 152-163.

Brauer, M. J. et al., 2008. Coordination of Growth Rate, Cell Cycle, Stress Response, and Metabolic Activity in Yeast. *Molecular Biology of the Cell,* 19(1), pp. 352-367.

Brazma, A. et al., 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics,* Volume 29, pp. 365-371.

Brocchieri, L. & Karlin, S., 2005. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research,* 33(10), p. 3390–3400.

Bryan, K. et al., 2014. Discovery and visualization of miRNA-mRNA functional modules within integrated data using bicluster analysis. *Nucleic Acids Research,* 42(3), p. e17.

Cahan, P. et al., 2007. Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. *Gene,* Volume 401, p. 12–18.

Caldas, J. & Kaski, S., 2008. *Bayesian biclustering with the plaid model.* Cancún, Mexico, IEEE Workshop on Machine Learning for Signal Processing, MLSP 2008 , pp. 291-296.

Calza, S. & Pawitan, Y., 2010. Normalization of gene-expression microarray data. *Methods in Molecular Biology,* Volume 673, pp. 37-52.

Carpenter, A. C. et al., 2012. The transcription factors THPOK and LRF are necessary and partly redundant for T helper cell differentiation. *Immunity,* Volume 37, p. 622–633.

Cheng, Y. & Church, G. M., 2000. *Biclustering of expression data.* Boston, MA, USA, ISMB, pp. 93-103.

Cheng, Y. et al., 2009. Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Research,* Volume 19, pp. 2172-2184.

Chin, S. L., Marcus, I. M., Klevecz, R. R. & Li, C. M., 2012. Dynamics of oscillatory phenotypes in Saccharomyces cerevisiae reveal a network of genome-wide transcriptional oscillators. *FEBS J.,* 279(6), p. 1119–1130.

Cho, R. J. et al., 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell,* Volume 2, p. 65–73.

Cho, R. J. et al., 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular cell,* 2(1), pp. 65-73.

Chumnanpuen, P., Nookaew, I. & Nielsen, J., 2013. Integrated analysis, transcriptome-lipidome, reveals the effects of INO-level (INO2 and INO4) on lipid metabolism in yeast. *BMC Systems Biology,* 7(Suppl 3), p. S7.

Cleveland, W. S., 1979. Robust locally weighted regression and smoothing scatter-plots. *Journal of the American Statistical Association,* Volume 74, p. 829–836.

Cohen, J., 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.. *Psychological bulletin,* 70(4), p. 213.

Collins, F., Lander, E., Rogers, J. & Waterson, R., 2004. Finishing the euchromatic sequence of the human genome. *Nature,* 431(7011), pp. 931-945.

Costanzo, M. et al., 2010. The genetic landscape of a cell. *Science,* 327(5964), pp. 425-431.

Dhillon, I. S., 2001. *Co-clustering documents and words using bipartite spectral graph partitioning.* s.l., s.n., pp. 269-274.

Dikicioglu, D. et al., 2011. How yeast re-programmes its transcriptional profile in response to different nutrient impulses. *BMC Systems Biology,* Volume 5, p. 148:163.

Dimitriadou, E., Weingessel, A. & Hornik, K., 2001. Voting-merging: An ensemble method for clustering. In: *Artificial Neural Networks—ICANN 2001.* New York: Springer, pp. 217-224.

Doré, L. C. & Crispino, J. D., 2011. Transcription factor networks in erythroid cell and megakaryocyte development. *Blood,* Volume 118, pp. 231-239.

Drobna, E. et al., 2012. Overexpression of the YAP1, PDE2, and STB3 genes enhances the tolerance of yeast to oxidative stress induced by 7-chlorotetrazolo[5,1-c]benzo[1,2,4]triazine. *FEMS Yeast Research,* Volume 12, p. 958–968.

Dudoit, S., Yang, Y. H., Callow, M. J. & Speed, T. P., 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica,* Volume 12, p. 111–139.

Duina, A. A., Miller, M. E. & Keeney, J. B., 2014. Budding yeast for budding geneticists: A primer on the Saccharomyces cerevisiae model system. *Genetics,* 197(1), p. 33–48.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci (PNAS),* Volume 95, pp. 14863-14868.

Eren, K., 2012. *Application of biclustering algorithms to biological data,* Columbus, USA: The Ohio State University.

Eren, K., Deveci, M., Küçüktunç, O. & Çatalyürek, Ü. V., 2013. A comparative analysis of biclustering algorithms for gene expression data. *Briefings in bioinformatics,* 14(3), pp. 279-292.

Fernandez, M. A., Rueda, C. & Peddada, S. D., 2012. Identification of a core set of signature cell cycle genes whose relative order of time to peak expression is conserved across species. *Nucleic Acids Research,* Volume 40, p. 2823–2832.

Ferreira, R. T. et al., 2012. Arsenic stress elicits cytosolic Ca(2+) bursts and Crz1 activation in Saccharomyces cerevisiae. *Microbiology,* 158(Pt 9), pp. 2293-2302.

Filkov, V. & Skiena, S., 2004. Integrating microarray data by consensus clustering. *International Journal on Artificial Intelligence Tools,* 13(04), pp. 863-880.

Fred, A. L. & Jain, A. K., 2005. Combining multiple clusterings using evidence accumulation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 27(6), pp. 835-850.

Fujii, S. et al., 2010. Cytoplasmic-nuclear genomic barriers in rice pollen development revealed by comparison of global gene expression profiles among five independent cytoplasmic male sterile lines. *Plant Cell Physiology,* 51(4), pp. 610-620.

Gardner, M. J. et al., 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature,* 419(6906), p. 498–511.

Gasch, A. P. et al., 2001. Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Molecular Biology of the Cell,* Volume 12, p. 2987–3003.

Gasch, A. P. et al., 2001. Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Molecular biology of the cell,* 12(10), pp. 2987-3003.

Gasch, A. P. et al., 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell,* Volume 11, p. 4241–4257.

Gasch, A. P. et al., 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell,* 11(12), pp. 4241-4257.

Ge, H. et al., 2010. Comparative analyses of time-course gene expression profiles of the long-lived sch9Delta mutant. *Nucleic Acids Research,* 38(1), pp. 143-158.

GeneMANIA, 2014. [Online] Available at: http://www.genemania.org/

Getz, G., Levine, E. & Domany, E., 2000. Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences,* 97(22), pp. 12079-12084.

Gilmore, J. M. et al., 2012. Characterization of a highly conserved histone related protein, Ydl156w, and its functional associations using quantitative proteomic analyses. *Molecular & Cellular Proteomics,* April.11(4).

Gionis, A., Mannila, H. & Tsaparas, P., 2007. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data (TKDD),* 1(1), p. 4.

Goffeau, A. et al., 1996. Life with 6000 Genes. *Science,* 274(5287), pp. 546-567.

Golub, T. R. et al., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science,* 286(5439), pp. 531-537.

González-Aguilera, C. et al., 2011. Nab2 functions in the metabolism of RNA driven by polymerases II and III. *Molecular Biology of the Cell,* 22(15), pp. 2729-2740.

Gupta, S., Stamatoyannopolous, J. A., Bailey, T. & Noble, W. S., 2007. Quantifying similarity between motifs. *Genome Biology,* Volume 8.

Harrison, P. M. et al., 2003. Identi®cation of pseudogenes in the Drosophila melanogaster genome. *Nucleic Acids Research,* 31(3), p. 1033±1037.

Hattangadi, S. M., Burke, K. A. & Lodish, H. F., 2010. Homeodomain-interacting protein kinase 2 plays an important role in normal terminal erythroid differentiation. *Blood,* Volume 115, pp. 4853-4861.

Hattangadi, S. M. et al., 2011. From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. *Blood,* Volume 118, pp. 6258-6268.

Hayata, T., Blitz, I. L., Iwata, N. & Cho, K. W., 2009. Identification of embryonic pancreatic genes using Xenopus DNA microarrays. *Developmental Dynamics,* 238(6), p. 1455–1466.

Haykin, S., 1999. *Neural Networks – A Comprehensive Foundation.* 3nd Edition ed. Singapore: Pearson, Prentice Hall.

Hay, S. I. et al., 2009. A World Malaria Map: Plasmodium falciparum Endemicity in 2007. *PLOS Medicine,* 6(10).

Herskowitz, I., 1988. Life cycle of the budding yeast Saccharomyces cerevisiae. *Microbiological Reviews,* 52(4), p. 536–553.

Holm, A. K. et al., 2010. Metabolic and transcriptional response to cofactor perturbations in Escherichia coli. *The Journal of Biological Chemistry,* 285(23), pp. 17498-17506.

Hughes, T. R. et al., 2000. Functional discovery via a compendium of expression profiles. *Cell,* 102(1), pp. 109-126.

Huttenhower, C. et al., 2009. Detailing regulatory networks through large scale data integration. *Bioinformatics,* 25(24), pp. 3267-3274.

Ideker, T. et al., 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science,* 292(5518), pp. 929-934.

Irizarry, R. A. et al., 2003a. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics,* Volume 4, pp. 249-264.

Iyer, V. R. et al., 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature,* Volume 409, pp. 533-538.

Jenner, R. G. et al., 2003. Kaposi's sarcoma-associated herpesvirus-infected primary effusion lymphoma has a plasma cell gene expression profile. *Proceedings of the National Academy of Sciences,* 100(18), pp. 10399-10404.

Kaiser, S. & Leisch, F., 2008. *A toolbox for bicluster analysis in R,* Heidelberg, Germany: Proceedings in ComputationalStatistics.

Kamenšek, S. & Žgur-Bertok, D., 2013. Global transcriptional responses to the bacteriocin colicin M in Escherichia coli. *BMC Microbiology,* 13(42).

Karypis, G. & Kumar, V., 1995. Metis-unstructured graph partitioning and sparse matrix ordering system, version 2.0.

Karypis, G. & Kumar, V., 1998. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing,* 20(1), pp. 359-392.

Keller, M. A. et al., 2006. Transcriptional regulatory network analysis of developing human erythroid progenitors reveals patterns of coregulation and potential transcriptional regulators. *Physiol Genomics,* Volume 28, pp. 114-128.

Kluger, Y., Basri, R., Chang, J. T. & Gerstein, M., 2003. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome research,* 13(4), pp. 703-716.

Kohonen, T. E., 1997. *Self-Organizing Maps.* New York: Springer-Verlag.

Kong, D. et al., 2012. Targeted knockdown of clusterin sensitizes pancreatic cancer MIA-PaCa-2 cell to gmcitabine treatment by inactivation of NF-kB/ Bcl-2. *Biomedical Research,* Volume 23, pp. 91-98.

Kooij, T. W. A., Janse, C. J. & Waters, A. P., 2006. Plasmodium post-genomics: better the bug you know?. *Nature Reviews Microbiology,* Volume 4, pp. 344-357.

Kovacs, L. A. S. et al., 2012. Cyclin-dependent kinases are regulators and effectors of oscillations driven by a transcription factor network. *Molecular Cell,* 45(5), p. 669–679.

Lam, Y. K. & Tsang, P. W., 2012. eXploratory K-Means: A new simple and efficient algorithm for gene clustering. *Applied Soft Computing,* Volume 12, p. 1149–1157.

Lanza, A. M., Blazeck, J. J., Crook, N. C. & Alper, H. S., 2012. Linking yeast Gcn5p catalytic function and gene regulation using a quantitative, graded dominant mutant approach. *PLOS ONE,* 7(4), p. e36193.

Larsson, M. et al., 2013. Functional studies of the yeast med5, med15 and med16 mediator tail subunits. *PLOS ONE,* 8(8), p. e73137.

Laubacher, M. E. & Ades, S. E., 2008. The Rcs phosphorelay is a cell envelope stress response activated by peptidoglycan stress and contributes to intrinsic antibiotic resistance. *Journal of Bacteriology,* 190(6), pp. 2065-2074.

Lazzeroni, L., Owen, A. & others, 2002. Plaid models for gene expression data. *Statistica sinica,* 12(1), pp. 61-86.

Le Roch, K. G. et al., 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science.*

Lee, J. et al., 2008. Indole cell signaling occurs primarily at low temperatures in Escherichia coli. *The ISME Journal,* Volume 2, p. 1007–1023.

Liko, D., Conway, M. K., Grunwald, D. S. & Heideman, W., 2010. Stb3 plays a role in the glucose-induced transition from quiescence to growth in Saccharomyces cerevisiae. *Genetics,* Volume 185, pp. 797-810.

Liko, D., Slattery, M. G. & Heideman, W., 2007. Stb3 binds to ribosomal RNA processing element motifs that control transcriptional responses to growth in Saccharomyces cerevisiae. *Journal of Biological Chemistry,* Volume 282, pp. 26623-26628.

Limb, J.-K.et al., 2009. Regulation of megakaryocytic differentiation of K562 cells by FosB, a member of the Fos family of AP-1 transcription factors. *Cellular and Molecular Life Sciences,* Volume 66, p. 1962 – 1973.

Liu, D. et al., 2004. A random-periods model for expression of cell-cycle genes. *Proc Natl Acad Sci (PNAS),* Volume 11, pp. 7240-7245.

Liu, X. et al., 2008. Genome-wide analysis of gene expression profiles during the kernel development of maize (Zea mays L.). *Genomics,* 91(4), pp. 378-387.

Liu, Z. et al., 2013. Anaerobic α-amylase production and secretion with fumarate as the final electron acceptor in Saccharomyces cerevisiae. *Applied and Environmental Microbiology,* 79(9), p. 2962–2967.

Loi, S. et al., 2007. Definition of clinically distinct molecular subtypes in estrogen receptor--positive breast carcinomas through genomic grade. *Journal of clinical oncology,* 25(10), pp. 1239-1246.

Madeira, S. C. & Oliveira, A. L., 2004. Biclustering algorithms for biological data analysis: a survey. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on,* 1(1), pp. 24-45.

Maeda, T. et al., 2005a. Role of the proto-oncogene Pokemon in cellular transformation and ARF repression. *Nature,* Volume 433, pp. 278-285.

Maeda, T. et al., 2009. LRF is an essential downstream target of GATA1 in erythroid development and regulates BIM-dependent apoptosis. *Developmental Cell,* Volume 17, p. 527–540.

Mardis, E. R., 2010. The $1,000 genome, the $100,000 analysis?. *Genome Medicine,* Volume 2, p. 84.

Martínez-Pastor, M. T. et al., 1996. The Saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE). *EMBO,* 15(9), p. 2227–2235.

Matia-González, A. M. & Rodríguez-Gabriel, M. A., 2011. Slt2 MAPK pathway is essential for cell integrity in the presence of arsenate. *Yeast,* 28(1), pp. 9-17.

MEME, 2014. [Online] Available at: http://meme.nbcr.net/meme/cgi-bin/meme.cgi

Merryweather-Clarke, A. T. et al., 2011. Global gene expression analysis of human erythroid progenitors. *Blood,* 117(13), pp. 4685-4686.

Miller, L. D. et al., 2005. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America,* 102(38), pp. 13550-13555.

Mirkin, B., 1996. *Mathematical classification and clustering.* Dordrecht: Kluwer Academic Press.

Monti, S., Tamayo, P., Mesirov, J. & Golub, T., 2003. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning,* 52(1-2), pp. 91-118.

Morillo-Huesca, M., Clemente-Ruiz, M., Andújar, E. & Prado, F., 2010. The SWR1 histone replacement complex causes genetic instability and genome-wide transcription misregulation in the absence of H2A.Z. *PLOS ONE,* 5(8), p. e12143.

Murali, T. & Kasif, S., 2003. *Extracting conserved gene expression motifs from gene expression data.* Lihue, Hawaii, Pacific Symposium on Biocomputing, pp. 77-88.

Neidigh, J. W., Fesinmeyer, R. M. & Andersen, N. H., 2002. Designing a 20-residue protein. *Nature Structural Biology,* Volume 9, pp. 425-430.

Nilsson, R. et al., 2009. Discovery of Genes Essential for Heme Biosynthesis through Large-Scale Gene Expression Analysis. *Cell Metabolism,* Volume 10, pp. 119-130.

Oghabian, A., Kilpinen, S., Hautaniemi, S. & Czeizler, E., 2014. Biclustering Methods: Biological Relevance and Application in Gene Expression Analysis. *PloS one,* 9(3), p. e90801.

Omelyanchuk, L. V., Trunova, S. A., Lebedeva, L. I. & Fedorova, S. A., 2004. Key events of the cell cycle: regulation and organization. *Russian Journal of Genetics,* 40(3), p. 219–234.

Önskog, J. et al., 2011. Classification of microarrays; synergistic effects between normalization, gene selection and machine learning. *BMC Bioinformatics,* Volume 12, p. e390.

Opitz, C. A. et al., 2003. Damped elastic recoil of the titin spring in myofibrils of human myocardium. *Proc. Natl. Acad. Sci. (PNAS),* 100(22), p. 12688–12693.

Orelle, C. et al., 2013. Tools for characterizing bacterial protein synthesis inhibitors. *Antimicrobial Agents and Chemotherapy,* 57(12), pp. 5994-6004.

Orlando, D. A. et al., 2008. Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature,* Volume 453, pp. 944-947.

Otto, T. D. et al., 2014. A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biology,* Volume 12, p. 86.

PAINT, 2013. *Promoter Analysis and Interaction Network Toolset (V 4.0-pre).* [Online] Available at: http://www.dbi.tju.edu/dbi/tools/paint/ [Accessed March 2013].

Parreiras, L. S., Kohn, L. M. & Anderson, J. B., 2011. Cellular effects and epistasis among three determinants of adaptation in experimental populations of Saccharomyces cerevisiae. *Eukaryot Cell,* 10(10), pp. 1348-1356.

Partridge, J. D. et al., 2008. Characterization of the Escherichia coli K-12 ydhYVWXUT operon: regulation by FNR, NarL and NarP. *Microbiology,* 154(2), pp. 608-618.

Pawitan, Y. et al., 2005. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research,* 7(6), p. R953.

Pena, J. M., Lozano, J. A. & Larranaga, P., 1999. An empirical comparison of four initialization methods for the K-Means algorithm. *Pattern Recognition Letters,* 20(10), pp. 1027-1040.

Peng, J., Chen, J. & Wang, Y., 2013. Identifying cross-category relations in gene ontology and constructing genome-specific term association networks. *BMC Bioinformatics,* 14(Suppl 2).

Pomeroy, S. L. et al., 2002. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature,* 415(6870), pp. 436-442.

Pramila, T. et al., 2006. The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phasegap in the transcriptional circuitryof the cell cycle. *Genes and Development,* Volume 20, p. 2266–2278.

Prelić, A. et al., 2006. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics,* 22(9), pp. 1122-1129.

Quackenbush, J., 2002. Microarray data normalization and transformation. *Nature Genetics,* Volume 32, p. 496–501.

Ramaswamy, S. et al., 2001. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences,* 98(26), pp. 15149-15154.

Reiss, D. J., Baliga, N. S. & Bonneau, R., 2006. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC bioinformatics,* 7(1), p. 280.

Rhouma, M. B. H. & Frigui, H., 2001. Self-organization of pulse-coupled oscillators with application to clustering. *IEEE Trans. Pattern Anal. Mach. Intell.,* 23(2), pp. 1-16.

Roberts, P. C., 2008. Gene expression microarray data analysis demystified. *Biotechnology Annual Review,* Volume 14, pp. 29-61.

Roy, S. et al., 2013. Arboretum: Reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. *Genome Research,* 23(6), pp. 1039-1050.

Salem, S. A., Jack, L. B. & Nandi, A. K., 2008. Investigation of self-organizing oscillator networks for use in clustering microarray data. *IEEE Trans. Nanobioscience,* 7(1), pp. 65-79.

Salem, S., Jack, L. & Nandi, A., 2008. Investigation of self-organizing oscillator networks for use in clustering microarray data. *IEEE Trans. Nanobioscience,* 7(1), pp. 65-79.

Sanz, A. B. et al., 2012. Chromatin remodeling by the SWI/SNF complex is essential for transcription mediated by the yeast cell wall integrity MAPK pathway. *Molecular Biology of the Cell,* 23(14), p. 2805–2817.

Schmitt, A. P. & McEntee, K., 1996. Msn2p, a zinc finger DNA-binding protein, is the transcriptional activator of the multistress response in Saccharomyces cerevisiae. *PNAS,* 93(12), p. 5777–5782.

Segal, E. et al., 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics,* Volume 34, pp. 166-176.

Seiler, M., Huang, C. C., Szalma, S. & Bhanot, G., 2010. ConsensusCluster: a software tool for unsupervised cluster discovery in numerical data. *OMICS A Journal of Integrative Biology,* 14(1), pp. 109-113.

SGD, 2014. *Saccharomyces Genome Database.* [Online] Available at: http://www.yeastgenome.org/

SGD, 2014. *Slim Mapper tool.* [Online] Available at: http://www.yeastgenome.org/cgi-bin/GO/goSlim-Mapper.pl

SGD, 2014. *Term Finder tool.* [Online] Available at: http://www.yeastgenome.org/cgi-bin/GO/goTermFinder.pl

Shabalin, A. A., Weigman, V. J., Perou, C. M. & Nobel, A. B., 2009. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics,* 3(3), pp. 985-1012.

Shalgi, R. et al., 2013. Widespread regulation of translation by elongation pausing in heat shock. *Molecular Cell,* 49(3), p. 439–452.

Siegmund, R. F. & Nasmyth, K. A., 1996. he Saccharomyces cerevisiae Start-specific transcription factor Swi4 interacts through the ankyrin repeats with the mitotic Clb2/Cdc28 kinase and through its conserved carboxy terminus with Swi6. *Molecular and Cellular Biology,* 16(6), p. 2647–2655.

Sievertzon, M., Nilsson, P. & Lundeberg, J., 2006. Improving reliability and performance of DNA microarrays. *Expert Review of Molecular Diagnostics,* Volume 6, pp. 481-492.

Slattery, M. G., Liko, D. & Heideman, W., 2006. The function and properties of the Azf1 transcriptional regulator change with growth conditions in Saccharomyces cerevisiae. *Eukaryotic Cell,* Volume 5, pp. 313-320.

Slonim, N., Atwal, G. S., Tkacik, G. & Bialek, W., 2005. Information-based clustering. *Proc Natl Acad Sci USA,* 102(51), pp. 18297-18302.

Smyth, G. K. & Speed, T., 2003. Normalization of cDNA microarray data. *Methods,* Volume 31, p. 265–273.

Spellman, P. T. et al., 1998. Comprehensive identification of cell cycle--regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell,* 9(12), pp. 3273-3297.

Spellman, P. T. et al., 1998. Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell,* Volume 9, p. 3273–3297.

Stein, T., Kricke, J., Becher, D. & Lisowsky, T., 1998. Azf1p is a nuclear-localized zinc-finger protein that is preferentially expressed under non-fermentative growth conditions in Saccharomyces cerevisiae. *Current Genetics,* Volume 34, p. 287–296.

Strassburg, K. et al., 2010. Dynamic transcriptional and metabolic responses in yeast adapting to temperature stress. *OMICS,* 14(3), p. 249–259.

Strehl, A. & Ghosh, J., 2003. Cluster ensembles---a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research,* Volume 3, pp. 583-617.

Su, A. I. et al., 2002. Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences,* 99(7), pp. 4465-4470.

Suzuki, T. & Iwahashi, Y., 2011. Gene expression profiles of yeast Saccharomyces cerevisiae sod1 caused by patulin toxicity and evaluation of recovery potential of ascorbic acid. *J Agric Food Chem.,* 59(13), pp. 7145-7154.

Suzuki, T. & Iwahashi, Y., 2012. Comprehensive gene expression analysis of type B trichothecenes. *J. Agric. Food Chem.,* 60(37), p. 9519–9527.

Swift, S. et al., 2004. Consensus clustering and functional interpretation of gene-expression data. *Genome biology,* 5(11), p. R94.

Tang, C., Zhang, L., Zhang, A. & Ramanathan, M., 2001. *Interrelated two-way clustering: an unsupervised approach for gene expression data analysis.* Bethesda, Maryland, Proceedings of the IEEE 2nd International Symposium on Bioinformatics and Bioengineering Conference, 2001 , pp. 41-48.

Tchagang, A. B. et al., 2011. Biclustering of DNA Microarray Data: Theory, Evaluation, and Applications. *Handbook of Research on Computational and Systems Biology: Interdisciplinary Applications (1 Vol),* p. 148.

Tchagang, A. B. & Tewfik, A. H., 2006. DNA microarray data analysis: a novel biclustering algorithm approach. *EURASIP Journal on Applied Signal Processing,* Volume 2006, pp. 60-60.

Tchagang, A. B. et al., 2008. Early detection of ovarian cancer using group biomarkers. *Molecular cancer therapeutics,* 7(1), pp. 27-37.

The Gene Ontology Consortium, 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics,* 25(1), pp. 25 - 29.

The Gene Ontology Consortium, 2013. Gene Ontology annotations and resources. *Nucleic Acids Research,* 41(Database), p. D530–D535.

Tkach, J. M. et al., 2012. Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. *Nature Cell Biology,* Volume 14, pp. 966-976.

TOMTOM, 2014. [Online] Available at: http://meme.nbcr.net/meme/cgi-bin/tomtom.cgi

Topchy, A., Jain, A. K. & Punch, W., 2005. Clustering ensembles: Models of consensus and weak partitions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 27(12), pp. 1866-1881.

Tsankov, A. M. et al., 2010. The role of nucleosome positioning in the evolution of gene regulation. *PLOS Biology,* 8(7), p. e1000414.

Vadigepalli, R. et al., 2003. PAINT: A Promoter Analysis and Interaction Network Generation Tool for Gene Regulatory Network Identification. *OMICS,* Volume 7, pp. 235-252.

Vega-Pons, S. & Ruiz-Shulcloper, J., 2011. A survey of clustering ensemble algorithms. *Int J Pattern Recognit Artif Intell,* 25(3), pp. 337-372.

Wade, C. H., Umbarger, M. A. & McAlear, M. A., 2006. The budding yeast rRNA and ribosome biosynthesis (RRB) regulon contains over 200 genes. *Yeast,* Volume 23, p. 293–306.

Wade, S. L., Poorey, K., Bekiranov, S. & Auble, D. T., 2009. The Snf1 kinase and proteasome-associated Rad23 regulate UV-responsive gene expression. *EMBO J.,* 28(19), pp. 2919-2931.

Walasek, M. A. et al., 2012. The combination of valproic acid and lithium delays hematopoietic stem/progenitor cell differentiation. *Blood,* Volume 119, pp. 3050-3059.

Wang, D. et al., 2012. Extensive up-regulation of gene expression in cancer: the normalised use of microarray data. *Molecular BioSystems,* Volume 8, p. 818–827.

Welch, J. J. et al., 2004. Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood,* Volume 104, pp. 3136-3147.

World Health Organization (WHO), 2013. *World Malaria Report,* Geneva: World Health Organization (WHO).

Xiao, X. et al., 2003. *Gene clustering using self-organizing maps and particle swarm optimization.* Indianapolis, s.n., pp. 154-163.

Xie, Y. et al., 2004. A case study on choosing normalization methods and test statistics for two-channel microarray data. *Comparative and Functional Genomics,* Volume 5, p. 432–444.

Xue-Franzén, Y., Henriksson, J., Bürglin, T. R. & Wright, A. P., 2013. Distinct roles of the Gcn5 histone acetyltransferase revealed during transient stress-induced reprogramming of the genome. *BMC Genomics,* Volume 14, p. 479.

Xu, J. et al., 2012. Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Developmental Cell,* Volume 23, p. 796–811.

Yang, J., Wang, H., Wang, W. & Yu, P. S., 2005. An improved biclustering method for analyzing gene expression profiles. *International Journal on Artificial Intelligence Tools,* 14(05), pp. 771-789.

Yang, Y. H. et al., 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research,* Volume 30.

Yeoh, E.-J.et al., 2002. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer cell,* 1(2), pp. 133-143.

Yeung, K. Y., 2001. *Cluster analysis of gene expression data [Ph.D. Thesis].* Seattle: University of Washington.

Yeung, K. Y. et al., 2001. Model-based clustering and data trasformations for gene expression data. *Bioinformatics,,* Volume 17, pp. 977-987.

Yu, M. et al., 2009. Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Molecular Cell,* Volume 36, p. 682–695.

Zhao, L. P., Presntice, R. & Breeden, L., 2001. Statistical modelling of large microarray data sets to identify stimulus-response profiles. *Proc Natl Acad Sci (PNAS),* 98(10), pp. 5631-5636.

Zhao, X., Ning, Q., Sun, X. & Tian, D., 2011. Pokemon reduces Bcl-2 expression through NF-k p65: a possible mechanism of hepatocellular carcinoma. *Asian Pacific Journal of Tropical Medicine,* Volume 4, pp. 492-497.

Zhu, C. et al., 2009. High-resolution DNA binding specificity analysis of yeast transcription factors. *Genome Research,* Volume 19, pp. 556-566.

Zhu, Z., Liu, W., He, S. & Ji, Z., 2012. *Memetic clustering based on particle swarm optimizer and k-means.* Brisbane, Australia, s.n.

Zu, X. et al., 2011. Pro-oncogene Pokemon promotes breast cancer progression by upregulating survivin expression. *Breast Cancer Research,* Volume 13.

Appendix III

Index